Volume 6, Issue 12, December 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Design of an Optimal Method for Disease Prediction Using Data Mining Techniques

Sambasiva Rao Voleti *
Research Scholar, Dept. of Computer Science,
Krishna University, Andhra Pradesh, India

Kiran Kumar Reddi
Assistant Professor, Dept. of Computer Science,
Krishna University, Andhra Pradesh, India

Abstract — The exponential growth of the amount of medical data available raises a problem that is the extraction of valuable and helpful information from these data. It is the foremost challenges in computational biology, which requite the development of efficient computational analysis tool to predict the diseases and is the problem that was presented in this paper. Data mining is a process of investigating data from the different perspectives and summarizing it into useful and valuable information. The main purpose of the data mining process is to identify new patterns from the existing data and to understand the data patterns to present meaningful and helpful information for the users. In this paper, data mining techniques were used to analyze heart disease and diabetes datasets to predict diseases and compared the results of each classification model's performance. And also feature reduction method was used in disease prediction which can reduce the number of tests that are essential to be conducted on a patient. These results may help out physicians and medical scientists in making accurate decisions in heart disease, and diabetes disease treatment. The performances of Naïve Bayes, Back Propagation, SVM, K-NN, C4.5 classification algorithm were evaluated compared. If we consider the accuracy, sensitivity, specificity, precision and mean absolute error, then the best model is SVM.

Keywords — Data Mining, Classification, WEKA, Cleveland Heart Disease Dataset, Pima Indian Diabetes Dataset, Data Preprocessing, Feature Selection.

I. INTRODUCTION

Data Mining is one of the most motivating and essential research areas with an objective of discovering significant information from large amounts of data sets. In present period, Data mining methods are becoming favourites in the medical field because there is a requirement for an effective analytical methodology to find the unknown and precious information hidden in medical data. Data mining offers various benefits in health industry, such as detecting the unfair practices in health insurance industry, availability of various medical treatments for curing the diseases to the patients at lower expenditure, finding the reasons for various diseases and detection of best medical treatment procedures. It can also helps the healthcare researchers for creating efficient medical policies, designing drug recommendation models, preparing the health records of individual patients etc. (Koh & Tan, 2005).

The most common data mining technique used in both academia and industry for data analysis is Classification. It is a supervised learning technique, which is applied to classify a dataset into predefined groups or classes. This paper addressed the accuracy, efficiency and other measures of the classification algorithms.

Heart disease is the primary reasons of death in the entire world over the past ten years. The World Health Organization reported that heart disease is the major leading cause of death in high and low income countries (WHO, 2007). According to the European Public Health Alliance 41% of all deaths were caused by heart attacks and other circulatory diseases (EPHA, 2010-11). According to the Economic and Social Commission of Asia and the Pacific in one fifth of Asian countries, the majority lives were vanished because of non-communicable diseases like cardiovascular, diabetes, and cancer diseases (ESCAP, 2010-11).

Oxygen is crucial to life as it provides fuel for all the body's functions. The heart's role is to supply oxygen-rich blood to the every cell in the human body. Coronary artery disease (CAD), also known as coronary heart disease, is a situation in which plaque builds up within the coronary arteries. These arteries provide the heart muscles with oxygen-rich blood. Plaque is made up of cholesterol, calcium, fat, and other substances which found in the blood. When plaque builds up within the arteries, the condition is known as Atherosclerosis (Wikipedia). Atherosclerosis (sometimes called "hardening" or "blockage" of the arteries) is the formation of cholesterol and fatty deposits (called plaques) on the inner walls of the arteries. These plaques can restrict blood flow into the heart muscle by physically clogging the artery or by causing abnormal artery tone and function. Without an adequate blood supply, the heart becomes starved of oxygen and the vital nutrients it needs to work properly. This can leads to chest pain called angina. If blood supply to a particular portion of the heart muscle is cut off entirely, or if the energy demands of the heart become much higher than its blood supply, a heart attack (injury to the heart muscle) may occur (Wikipedia).

Diabetes is one of the main common non-communicable disease (NCDs) that has extensively contributed to the increased mortality in patients. It is surely one of the most difficult health issues in the present century that evidently is

epidemic in maximum number of developing countries. In 2004, an estimation of 3.4 million people died because of fasting high blood sugar. A same number of deaths had been estimated for the year 2010. According to 2015 statistics 415 million people worldwide had suffering with diabetes and the IDF estimated that by 2016 this would increase to 642 million. More than 80% of diabetes deaths were taking place in middle-and- low-income countries (IDF).

This study focuses to explore the feasible solutions for the patients who are at a threat of developing type II diabetes disease in future. We focused to work on type II diabetes because this type can be prevented by following proactive measures. We propose to develop classifiers and design a prediction model for diabetes prediction based on the currently available data. For predicting the diabetes, we used Pima Indian diabetes dataset. Ultimately, the prediction model would be able to answer the requirement for urgent and important requirement to: (i) stop quick grow in diabetes, (ii) develop health awareness in public, and iii) prevent the beginning of this disease. Insulin is a kind of hormone that controls blood sugar in the human body. Hyperglycemia, also known as raised blood sugar, is a general consequence of uncontrolled diabetes and it leads to severe complications to many parts of the human body system over time, especially for the nerves and blood vessels. Type1 diabetes (previously called as juvenile or insulin-dependent) is the effect of less insulin production and it requires daily management of insulin. The reason for type1 diabetes is not identified and it is not avoidable with currently available knowledge. Type 2 diabetes occurs because of the body's ineffectiveness of utilizing insulin. Type 2 diabetes is mainly the result of physical inactivity and body over weight, and comprises 90% of people with diabetes disease worldwide.

Motivated by the rising mortality of Heart Disease and Diabetes patients every year world-wide and the handiness of enormous amount of patients' data that could be used to mine helpful knowledge; researchers have been utilizing data mining methods to assist medical professionals in the prediction of heart disease and diabetes.

The two medical datasets, Cleveland Heart Disease dataset and Pima Indian Diabetes dataset collected from University of California, Irvine (UCI) machine learning repository was used to evaluate the performance of each classification method. In order to find the optimal classification algorithm the comparative studies of these classification methods were performed using WEKA software. Weka is open source software that can be used to preprocess and analyze the datasets. Weka offers implementations of learning algorithms that can be applied easily to any dataset (Written & Frank, 2005).

II. RELATED WORK

A number of research works have focused on comparison of classification and predictive models within the computational biology and bioinformatics. These studies have been narrowed by comparing small number of (three or less) supervising algorithms in different biological datasets. This study will contribute to the biology and medicine fields by measuring the effectiveness of data mining supervising (classification and predictive) algorithms.

Now a day's research comparing the performance of traditional classifiers and intensive data mining methods has been increasing steadily. Maroco et al., assessed various data mining and traditional classifiers (classification tree, LR, NN, LDA, RF, and SVM) for predicting Alzheimer disease. The data mining methods neural networks, decision tree and SVM were evaluated and compared by Kim et al. in his research with LR for predicting mortality.

The major research study in the field of comparative study of data mining algorithms was completed by Michie, Spiegelhalter and Taylor (1994). This research project analyzed 21 datasets using 23 algorithms. The large number of datasets and algorithms caused difficulties in interpretation of the comparative performance. The authors stated that a hybrid regression-neural network and density estimating algorithms had the best performance. The effectiveness of this study was limited by the absence of a clearly defined research method and the limited scope of the study. This was evidenced in the wide range of sources and characteristics of the datasets used because data mining result or output depends on the quality and type of datasets.

Nidhi Bhatla et al., (2012) performed a work; this work aims at maximizing the effectiveness of proposed system. The observations exemplify that Decision Tree and Naive Bayes using fuzzy logic has better performed then the other data mining techniques.

Carlos Ordonez (2006) performed a work; In this study, search constraints were introduced to discover the association rules that are medically significant only and to make the search process more efficient. In medical terms, association rules relate patient risk factors and heart perfusion measurements to the degree of steno sis in four specific arteries. Association rule significance in medical dataset is assessed with the common measures support and confidence.

Biswadip Ghosh (2012) applied Fuzzy Composite Programming (FCP) to build a classifier for diabetes prediction by using the Pima Indian diabetes dataset. He used Receiver Operating Characteristic (ROC) Curves for evaluation of Fuzzy classifier performance and compared this performance against the performance of a Logistic Regression classifier. The results specified that FCP classifier was better than the Logistic Regression classifier, when the resulting AUC values were compared. The logistic regression technique has achieved an AUC of 64.8%, while the FCP technique has achieved an AUC of 72.2%. He showed that the fuzzy classier was superior than the logistic regression classifier in performance.

Pengyi Yang, et al., (2009) proposed a particle swarm based hybrid system for solving the class imbalance issue in medical data mining. To direct the sampling process, they applied multiple classification techniques. The classification methods applied in this hybrid system composition contains Naive Bayes (NB), Random Forest (RF), Logistic Regression (LOG), K-Nearest Neighbor (KNN), and Decision Tree (J48). A genome wide association study (GWAS) dataset from the genotyping of Single Nucleotide Polymorphisms (SNPs) of Age-related Macular Degeneration (AMD) and four other medical datasets like blood, breast, survival cancer and diabetes datasets from UCI Machine Learning

Repository were used in this study. The different techniques of the study includes Random over Sampling (RO), Random Under Sampling (RU), Particle Swam based Hybrid System (PSO), and Clustering. The metrics used were F-measure, Area under ROC Curve (AUC), and Geometric mean. Particle Swam based Hybrid System (PSO) achieved an average accuracy of 71.6%, cluster with KNN Classifier 64.9%, RU 67.3%, and RO 65.8% for Pima Indian Diabetes. This study showed that PSO better performed than other methods in accuracy.

Zhou, Kasabov and Purvis (2007) proposed a system depending on the statistical analysis of training dataset to choose fuzzy membership functions that can be utilized in association with fuzzy neural networks. This method was first explained and then demonstrated with the help of two different experimental examinations for the medical data.

Leonarda et al., (2006) in their research work applied a multilevel perceptron neural network to automatically detect the symptoms of diabetes in retinal images. The network was trained by using the algorithms for calculating the finest global threshold that can reduce the errors in pixel classification. Performance of the system was measured by using a competent index to give percentage measure in the identification of eye suspect areas depending on the neuro-fuzzy subsystem.

In this study we explored different number of data mining algorithms for developing predictive models for medical patient's datasets with either continuous or binary response variables. For the binary outcome datasets like diabetes and heart disease, we generated a number of models from different methods. This was helpful because it gave us a variety of models and indicates which model is better by evaluating the accuracy and other measures. The entire literature survey specifies that in medical data mining and medical field if we choose and apply these kinds of classification techniques the average accuracy achieved was 75%.

III. METHODOLOGY

In this work, the general approach or steps that were followed for building the predictive model is as follows:

- Collect the data.
- Preprocess the data to reduce the noise.
- Select the important attributes.
- Input the dataset with reduced set of attributes to the classifiers.
- Split the dataset into training and testing datasets.
- Apply data mining techniques to the training dataset.
- Develop the predictive model.
- Evaluate the performance of model using test dataset.
- Repeat this process with other all classification techniques.
- Compare the performance among these data mining techniques.

Data Collection

The researchers compared the performance of the data mining algorithms (Naïve Bayes, Back Propagation, SVM, K-Nearest Neighbour and C4.5 classifiers) for predicting heart disease, and diabetes on respective patient datasets. Each of these algorithms is executed on medical datasets collected from the UCI Machine learning repository. All the files were available for public usage at UCI website.

Data Preprocessing

In present era, the real-world data bases are highly prone to noise, inconsistency and missing because they were collected from several heterogeneous data sources and their huge size. The data with low-quality will produce to low-quality results when mining (Angeline Christobel et al., 2012). During the training phase the knowledge discovery becomes difficult if there is much redundant and irrelevant information. So there is a need for preprocessing the dataset before using it for mining. Data preprocessing is essential to produce quality results in mining. The final dataset used for mining will be the result of data preprocessing. Data preprocessing can be divided into the following categories:

- Data cleaning
- Data reduction
- Data integration
- Data transformations

Data cleaning is used to eliminate the noise and avoid the inconsistencies. Data reduction is used to decrease the size of the data by aggregating, clustering, or removing the redundant features. Data integration combines the data from various sources into a joint data storage area like a data warehouse. Data transformations such as normalization can be used to scale the data values to fall within a range from 0.0 to 1.0. This will enhance the accuracy and efficiency of data mining methods involving distance measures. Depending on the goal for the data mining and the domain of the problem, we can select the appropriate technique.

Data Cleaning

Many research and industrial datasets existing currently have missing values. They are introduced due to reasons, such as incorrect measurements, equipment errors and manual data entry procedures. The most vital step in data cleaning is the processing of missing values. Many different methods have been proposed in the literature to deal with

missing data (Batista et al., 2003). Missing values present in the data should be handled carefully; otherwise the bias might be introduced into the knowledge induced and will affect the performance of classification methods or the entire quality of the classification process (Grace Wabha et al., 1995). Imputation of missing values is a difficult issue in data mining (Briyan, 1996). Because of the complexity with missing values, most of the supervised learning algorithms were not well adapted to some specific application domains (for example Web Applications) because the existing algorithms were developed with the assumption that there were no missing values in the datasets. This tells the need for handling missing values. The reasonable solution to handle missing data is good database design but good analysis can help to minimize the problems (Newman et al., 1998).

In literature, various approaches are proposed to handle missing values in datasets, they are

1. Ignore the tuple which contains missing values
2. Fill the missing values present in the dataset manually
3. Fill the missing values with a global constant
4. Fill the missing values with the attribute mean

The well known imputation method to handle the missing data values was to use a variable's median or mean.

Data transformation

The clinical data are measured on different scales. Normalization scales all values of a given attribute in small specified range so that large numeric attributes do not outweigh smaller range attributes. Also the genetic and clinical data have to be on same scale before feeding them to machine learning algorithms.

Feature selection methods

Feature selection is a data preprocessing step applied to medical datasets. It selects subset of features from whole feature set based on some statistical score and removes redundant features that do not contribute to performance. Feature selection is the main stage in any classification application.

Validity and Reliability

Stratified 10-fold cross-validation is a common validation method used to decrease the bias and variance associated with random sampling of the training and test datasets (Thongkham et. al., 2007). Also, it is a general method for data selection in data mining related to medical research. This stratified 10-fold cross-validation method was applied in the present work in evaluating and validating the predictive model. Confusion matrix is a visualization tool normally used in supervised learning. Each row of the confusion matrix indicates the number of instances in an actual class, while each column of the confusion matrix represents the number of instances in a predicted class as shown in table I.

Table I: Confusion Matrix

		Predicted	
		Positive	Negative
Actual	Positive	*True Positive*	*False Negative*
	Negative	*False Positive*	*True Negative*

In the confusion matrix:

True Positive (**TP**): is the number of actual positive examples classified as positive.

False Positive (**FP**): is the number of actual negative examples classified as positive.

False Negative (**FN**): is the number of actual positive examples classified as negative.

True Negative (**TN**): is the number of actual negative examples classified as negative.

To authenticate the proposed model, various experiments were conducted. Different types of measurements were assessed as indicators of performance, they are

i) Accuracy ii) Sensitivity iii) Specificity iv) Precision v) Area under the ROC curve (AUC)
vi) Mean Absolute Error vii) Efficiency in terms of computing time.

All these above values are calculated using the following formulas

Accuracy: The accuracy of a classification algorithm is the percentage of the testing data set examples that were correctly classified by the classification algorithm.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \ (\%)$$

Sensitivity: Sensitivity is also referred as True positive rate i.e. the proportion of positive examples that were correctly identified. It is the detection of the disease rate that needs to be maximized for accurate prediction.

$$\text{Sensitivity} = TP / (TP + FN) \ (\%)$$

Specificity: Specificity is the True negative rate i.e. the proportion of negative examples that were correctly identified. It is the false alarm rate that is to be minimized for accurate prediction.

Specificity = TN / (FP + TN) (%)

Precision: precision is defined as the proportion of the true positives against all the positive results (both false positives and true positives)

Precision = TP / (TP + FP) (%)

Area Under ROC Curve: It is used to measure the accuracy of a classifier, which has the values in the range 0.0 to 1.0. The closer the AUC is to 1.0, the high accurate the model is.

Mean Absolute Error: It is the absolute value of the mean error calculated from the formula

$$MAE = (\sum_{i=1}^{n} | e_i |) / n$$

where

n is the total amount of data used to predict.

e_i is the difference between the actual data with predicted values.

It is the sum of the difference between the actual data with the predicted values.

Efficiency is how fast the algorithm can execute the prediction task. It is the time of computing of the classification algorithm.

Classification Algorithms

In classification technique data samples are divided into different target classes. The classification process predicts the value of target class for every data point present in the dataset. For example, a patient can be categorized as "heart disease patient" or "non-heart disease patient" on the grounds of their disease pattern by applying the classification technique. It is also called as supervised learning technique having known the class categories. Multilevel and binary are the two kinds of classification. In multilevel approach more target classes were considered like, "high risk", "low risk" and "medium risk" patients. Where as in a binary approach, only two possible classes like, "low" or "high" risk patients may be considered. The original dataset is divided into training dataset and testing datasets. Training dataset can be utilized to train the classifier and the testing dataset could be used to check the correctness of the classifier.

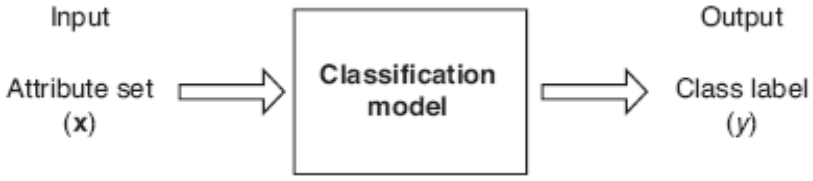

Fig. 1: Classification maps an input attribute x into its class label y

Classification techniques needs that the classes be defined based on the values of attributes in the given data. Pattern recognition is a kind of classification process where an input pattern is classified into one of the different classes based on its likeness to these predefined classes. Classification is considered as a two-step process (Mitchell, 1997).

In step 1, a classifier is developed based on a predetermined set of data classes or concepts. This step is called as the learning step (training phase). Here a classification algorithm develops the classifier by examining or "learning from" a training dataset made up of database tuples and their corresponding class labels. Every tuple is assumed to be belongs to a predefined class known as the class label. Because the class label of each training dataset tuple is provided, this step is also known as supervised learning. The step 1 can also be considered as the creation of a function or mapping, y = f (X), that can generate the corresponding class label y of a given tuple X. Typically, this mapping is described in the form of decision trees, classification rules, or mathematical formula.

In step 2, the developed model is used for classification. First, the classification accuracy is estimated. If we are using the classification accuracy as a measure to select the classifier, this measure would likely be high, because the classifier tends to over fit the data.

The classification methods used in this work are

i) Naive Bayes Classier
ii) Back propagation Algorithm
iii) Support Vector Machine (SVM)
iv) K-Nearest Neighbour Algorithm
v) C4.5 Algorithm

IV. RESULTS

The first dataset used in this research was heart disease dataset, which was collected from Cleveland Clinic Foundation and was supplied by Dr.Robert Detrano, of the V.A. Medical Center, Long Beach, CA. It was part of the databases collected at the University of California, Irvine by David Aha. The purpose of the dataset is to classify the presence of heart disease or absence of heart disease, given the results of various medical tests conducted on a patient. This database consists of 13 features as shown in table II. The database contained 270 examples after removing the instances with missing attribute values. The main criterion that physicians use to predict the heart disease is the narrowing in diameter of any one of the major blood vessel. The prediction was considered to be positive (presence of heart disease) if the diameter of any one of the major vessel was narrowed by more than 50%; and negative otherwise. The dataset contains 120 positive cases and 150 negative cases.

Table II: Cleveland Heart Disease Dataset Attributes

S. No	Name	Type
1	Age	Numeric
2	Sex	Numeric
3	Chest Pain Type	Numeric
4	Trestbps (Blood Pressure)	Numeric
5	Chol (Serum Cholesterol)	Numeric
6	Trestbps (Fasting Blood Sugar)	Numeric
7	Restecg (resting electrographic results)	Numeric
8	Thalach (maximum heart rate achieved)	Numeric
9	Exang (exercise induced angina)	Numeric
10	Oldpeak (ST depression induced by exercise relative to rest)	Numeric
11	Slope (the slope of the peak exercise ST segment)	Numeric
12	Ca (number of major vessels colored by floursopy)	Numeric
13	Thal	Numeric

In this study for selecting features information gain measure was used. Feature selection is made in the heart disease dataset to determine the most important features. It evaluates the worth of an attribute by calculating the information gain with respect to the class.

Info Gain (Class, Attribute) = H (Class) – H (Class | Attribute).

The features were ranked based on the Information Gain value of the attributes in heart disease dataset. Performances of Classifiers for 13 features set combination were presented in the table III.

Table III: Performance of classifiers for 13 feature set combination of CHDD.

	Accuracy	Sensitivity	Specificity	Precision	MAE	ROC	Time
Naïve Bayes	83.70	87.30	79.20	84.00	18.00	0.90	0.02
Back Prop.	79.30	81.30	76.70	81.30	21.00	0.86	1.65
SVM	**83.70**	**87.30**	**79.20**	**84.00**	**16.00**	**0.83**	**0.08**
KNN	75.20	76.70	73.30	78.20	25.02	0.75	0.01
C4.5	76.70	79.30	73.30	78.80	28.00	0.74	0.05

By applying backward elimination method we can remove the lowest information gain feature from the heart disease dataset. The remaining dataset is supplied as input to the classifier.
Performances of Classifiers for the 12 feature set combination were presented in the table IV.

Table IV: Performance of Classifiers for 12 feature set combination of CHDD.

	Accuracy	Sensitivity	Specificity	Precision	MAE	ROC	Time
Naïve Bayes	83.70	87.30	79.20	84.00	18.00	0.90	0.03
Back Prop.	77.80	77.30	78.30	81.70	22.00	0.85	1.48
SVM	**84.00**	**88.00**	**79.20**	**84.10**	**16.00**	**0.84**	**0.09**
KNN	76.30	79.30	72.50	78.30	23.92	0.76	0.01
C4.5	77.40	79.30	75.00	79.90	26.00	0.75	0.05

Performances of Classifier for the 11 feature set combination were presented in the table V.

Table V: Performance of Classifiers for 11 feature set combination of CHDD.

	Accuracy	Sensitivity	Specificity	Precision	MAE	ROC	Time
Naïve Bayes	84.80	88.00	80.80	85.20	18.23	0.90	0.01
Back Prop.	80.70	82.00	79.20	83.10	20.00	0.87	1.42
SVM	**84.00**	**87.30**	**80.00**	**84.50**	**16.00**	**0.84**	**0.08**
KNN	76.30	78.00	74.20	79.10	23.92	0.76	0.01
C4.5	78.10	81.30	74.20	79.70	26.00	0.76	0.03

Performances of Classifier for the 10 feature set combination were presented in the table VI.

Table VI: Performance of Classifiers for 10 feature set combination of CHDD.

	Accuracy	Sensitivity	Specificity	Precision	MAE	ROC	Time
Naïve Bayes	84.10	88.00	79.20	84.10	18.00	0.90	0.03
Back Prop.	81.50	85.30	76.70	82.10	19.00	0.86	1.18
SVM	**84.10**	**88.00**	**79.20**	**84.10**	**16.00**	**0.84**	**0.06**
KNN	75.90	78.00	73.30	78.50	24.29	75.70	0.01
C4.5	78.50	81.30	75.00	80.30	25.00	0.77	0.02

Performances of Classifiers for the 9 feature set combination were presented in the table VII.

Table VII: Performance of Classifiers for 9 feature set combination of CHDD.

	Accuracy	Sensitivity	Specificity	Precision	MAE	ROC	Time
Naïve Bayes	84.80	88.00	80.80	85.20	18.00	0.90	0.02
Back Prop.	81.50	86.70	75.00	81.30	20.00	0.86	1.04
SVM	**84.40**	**88.00**	**80.00**	**84.60**	**16.00**	**0.84**	**0.05**
KNN	81.10	84.70	76.70	81.90	19.14	0.81	0.01
C4.5	78.90	80.70	76.70	81.20	26.00	0.78	0.01

Performances of Classifiers for the 8 feature set combination were presented in the table VIII.

Table VIII: Performance of Classifiers for 8 feature set combination of CHDD.

	Accuracy	Sensitivity	Specificity	Precision	MAE	ROC	Time
Naïve Bayes	85.60	88.00	82.50	86.30	18.00	0.89	0.01
Back Prop.	80.40	84.00	75.80	81.30	21.00	0.88	1.17
SVM	**84.80**	**88.70**	**80.00**	**84.70**	**15.00**	**0.84**	**0.03**
KNN	77.40	80.70	73.30	79.10	22.81	0.77	0.01
C4.5	80.40	85.30	74.20	80.50	25.00	0.82	0.01

Performances of classifiers for the 7 feature set combination were presented in the table IX.

Table IX: Performance of Classifiers for 7 feature set combination of CHDD.

	Accuracy	Sensitivity	Specificity	Precision	MAE	ROC	Time
Naïve Bayes	85.20	90.00	79.20	84.40	19.00	0.89	0.01
Back Prop.	80.70	85.30	75.00	81.00	22.00	0.85	0.69
SVM	**85.90**	**90.70**	**80.00**	**85.00**	**14.00**	**0.85**	**0.03**
KNN	78.90	83.30	73.30	79.60	22.81	0.78	0.01
C4.5	80.40	84.00	75.80	81.30	26.00	0.81	0.01

SVM classifier has given more Accuracy, Sensitivity, Precision and Low Specificity, Mean Absolute Error, low computing time for almost all feature set combinations. So, the SVM classifier was identified as the best algorithm for getting better performances with all combinations of feature sets for the CHDD, which was presented in the figure 2. The SVM classifier outperformed Naïve Bayes, Back Propagation, KNN and C4.5 in terms of Accuracy, Sensitivity, precision and MAE.

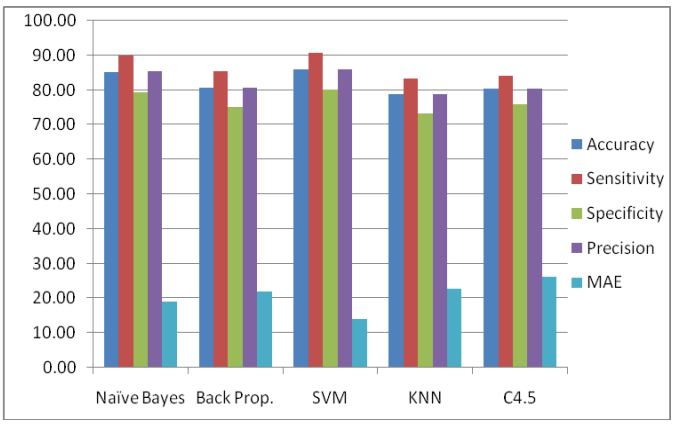

Fig.2. Comparison of Performances of Classification Algorithms on CHDD.

The SVM classifier performances for different feature set combination of Cleveland Heart Disease dataset were shown in the table X.

Table X. Performance of SVM Classifier with different set of features on CHDD.

	Accuracy	**Sensitivity**	**Specificity**	**Precision**	**MAE**	**ROC**	**Time**
13-attr.	83.70	87.30	79.20	83.70	16.00	0.83	0.09
12-attr.	84.00	88.00	79.20	84.10	16.00	0.84	0.08
11-attr.	84.00	87.30	80.00	84.10	16.00	0.84	0.08
10-attr.	84.10	88.00	79.20	84.10	16.00	0.84	0.06
9-attr.	84.40	88.00	80.00	84.40	16.00	0.84	0.05
8-attr.	84.80	88.70	80.00	84.80	15.00	0.84	0.03
7-attr.	**85.90**	**90.70**	**80.00**	**86.00**	**14.00**	**0.85**	**0.03**

From the above table we can observe that by removing the unimportant attributes the accuracy, sensitivity, precision and ROC was increased and the MAE, Time of computing was decreased.

The second dataset used in this study was "The Pima Indians Diabetes Data Set" collected from the UCI Machine Learning Repository. The original possessor of this data set is the National Institute of Diabetes and Digestive and Kidney Diseases. There were several constraints placed on the selection of this diabetes dataset from larger one. In particular, all patients' data present in this dataset were females and at least 21 years age from Pima Indian heritage. The PIDD database at the UCI machine learning repository had became a standard for testing the data mining algorithms to observe their prediction accuracy in diabetes data classification.

The diabetes dataset consists different patients data, as shown in Table XI, The dataset contains 768 patient records with 9 attributes including the class was used to build the predictive model. All attributes are numerical values. Class variable has two values tested_positive for diabetes (268 number of instances) and tested_negetive (500 number of instances).

Table XI: The Pima Indians Diabetes Dataset Attributes

S. No	**Attribute Name**	**Type**
1	Number of times pregnant	Numeric
2	Plasma glucose concentration a 2 hours in an oral	Numeric
3	Diastolic blood pressure (mm Hg)	Numeric
4	Triceps skin fold thickness (mm)	Numeric
5	2-Hour Serum insulin (mu U/ml)	Numeric
6	Body mass index (weight in kg / (height in m)2)	Numeric
7	Diabetes pedigree function Numeric	Numeric
8	Age (years)	Numeric

The features were ranked based on the Information Gain value of the attributes in PIDD. Performances of Classifier for the 8 feature set combination of PIDD were shown in the table XII.

Table XII: Performance of Classifier for 8 feature set combination of PIDD.

	Accuracy	**Sensitivity**	**Specificity**	**Precision**	**MAE**	**ROC**	**Time**
Naïve Bayes	76.30	84.40	61.20	78.90	28.41	0.82	0.31
Back Prop.	75.80	81.80	64.60	75.70	28.72	0.81	3.39
SVM	**77.47**	**90.00**	**54.10**	**78.53**	**22.53**	**0.72**	**0.41**
KNN	70.20	79.40	53.00	75.90	29.88	0.65	0.01
C4.5	73.80	81.40	59.70	73.50	31.57	0.75	0.04

By applying backward elimination method we can remove the lowest information gain feature from the diabetes dataset. And the remaining dataset is supplied as input to the classifier.
Performances of Classifiers for the 7 feature set combination were presented in the table XIII.

Table XIII: Performance of Classifier for 7 feature set combination of PIDD.

	Accuracy	**Sensitivity**	**Specificity**	**Precision**	**MAE**	**ROC**	**Time**
Naïve Bayes	76.43	85.20	60.10	79.92	28.74	0.82	0.03
Back Prop.	74.30	81.60	60.80	79.53	30.47	0.79	2.22

SVM	**76.95**	**89.80**	**53.00**	**78.10**	**23.05**	**0.71**	**0.11**
KNN	67.70	77.00	50.40	74.32	32.34	0.63	0.01
C4.5	74.60	81.60	61.60	74.40	31.37	0.78	0.02

SVM classifier shows more Accuracy, Sensitivity, Precision and Low Specificity, Mean Absolute Error and low computing time for almost all set of features. So, the SVM algorithm is identified as the best algorithm for giving better accuracy with all combinations of feature sets of Pima Indians Diabetes Dataset.

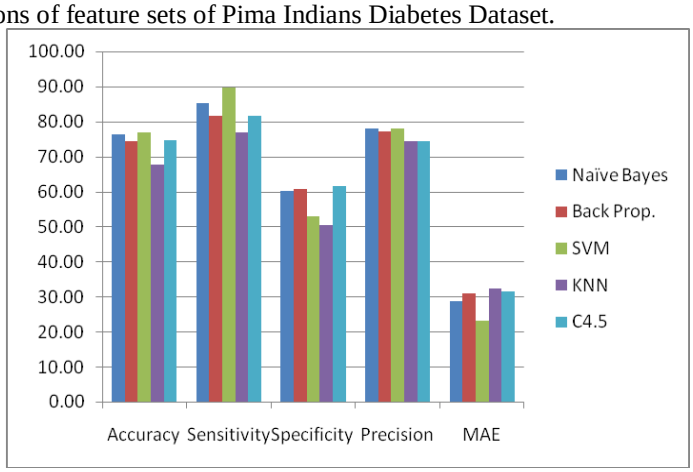

Fig.3: Comparison of performance of classification algorithms on PIDD.

The SVM classifier performances for different feature set combination of PIDD were shown in the table XIV.

Table XIV: Performance of SVM Classifier with Different Feature set Combinations on PIDD.

	Accuracy	Sensitivity	Specificity	Precision	MAE	ROC	Time
8-attr.	77.47	90.00	54.10	77.10	22.66	0.72	0.41
7-attr.	76.95	89.80	53.00	76.50	23.05	0.71	0.11

From the above table we can observe that by removing the unimportant attributes the accuracy, sensitivity, precision, ROC, and MAE was almost same and Time of computing was decreased.

V. CONCLUSION

For the binary outcome dataset, we developed a number of models from diverse data mining techniques. This was helpful because it gave us a variety of models and indicated which model is superior by evaluating the accuracy, sensitivity, precision and other measures. We choose the model with the highest overall accuracy, sensitivity and precision. If we just go by that criterion, then the best model is SVM. The results of the classification algorithms applied to the medical datasets were compared. SVM outperformed Naïve Bayes, Back Propagation, KNN and C4.5algorithms based on accuracy and other performance measures. And the performance of the classifiers increased after removing the unnecessary attributes.

REFERENCES

[1] H. C. Koh and G. Tan, Data Mining Application in Healthcare, Journal of Healthcare Information Management, vol. 19, no. 2, pp.64 -72. (2005)
[2] World Health Organization. 2007. Available from: http://www.who.int/mediacentre/factsheets/ fs310.pdf.
[3] European Public Health Alliance. 2010. Available from: http://www.epha.org/a/2352
[4] ESCAP. 2010 ,Available from: http://www.unescap.org/stat/data/syb2009/9.Healthrisks-causes- of-death.asp.
[5] International Diabetes Federation. IDF Diabetes Atlas: the global burden [Internet]. Available from: http://www.idf.org/diabetesatlas/5e/the-global-burden.
[6] Witten, I. H., & Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. San Francisco, CA: Morgan Kaufman.
[7] Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, de Mendonca A. Data mining methods in the prediction of Dementia: a real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC Res Notes 2011; 4:299.
[8] Kim S, Kim W, Park RW. A comparison of intensive care unit mortality prediction models through the use of data mining techniques. Health Inform Res 2011;17(4):232–243.
[9] Michie.D, Spiegelhalter. D.J and Taylor .C.C. Machine Learning, Neural and Statistical Classification, Chapter 9, page No 157-158.

[10] Nidhi Bhatla, Kiran Jyothi. A Novel Approach for Heart Disease Diagnosis using Data Mining and Fuzzy Logic. International Journal of Computer Applications (0975 – 8887) Volume 54– No.17, 2012.
[11] Carlos Ordonez. Comparing Association Rules and Decision Trees for Disease Prediction. HIKM'06, November 11, 2006, Arlington, Virginia,USA. Copyright 2006 ACM 1-59593-528-2/06/0011 pp.17-24.
[12] Pengyi Yang, LiangXu, Bing B Zhou, Zili Zhang, and Albert Y Zomaya, A particle swarm based hybrid system for imbalanced medical data sampling, BMC Genomics. 2009; 10(Suppl 3): S34.
[13] Biswadip Ghosh, Using Fuzzy Classification for Chronic disease, Indian Journal of Economics & Business, March 2012.pp. 231-240.
[14] Leonarda Carnimeo and A. Giaquinto, An Intelligent System for Improving Detection of Diabetic Symptoms in Retinal Images, IEEE International Conference on Information Technology in Biomedicine, Ioannina, 26-28 October 2006.
[15] Q. Q. Zhou, M. Purvis and N. Kasabov, Membership Function Selection Method for Fuzzy Neural Networks. University of Otago, Dunedin, 2007. http://otago.ourarchive.ac.nz/handle/10523/1027
[16] Angeline Christobel. Y, Dr.P.SivaPrakasam, The Negative Impact of Missing Value Imputation in Classification of Diabetes Dataset and Solution for Improvement, IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278- 0661, ISBN: 2278-8727. Volume 7, Issue 4, 2012, PP 16-23.
[17] Newman, D.J. & Hettich, S. & Blake, C.L. & Merz, C.J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.
[18] Brian D. Ripley (1996), Pattern Recognition and Neural Networks, Cambridge University Press, Cambridge.
[19] Grace Whaba, Chong Gu, Yuedong Wang, and Richard Chappell (1995), Soft Classification a.k.a. Risk Estimation via Penalized Log Likelihood and Smoothing Spline Analysis of Variance, in D. H. Wolpert (1995), The Mathematics of Generalization, 331-359, Addison-Wesley, Reading, MA.
[20] G.E.A.P.A. Batista and M.C. Monard, An analysis of four Missing Data treatment methods for supervised learning, Applied Artificial Intelligence 17 (2003)
[21] Thongkam, J., Xu, G., Zhang, Y., & Huang, F. (2007). Breast cancer survivability via AdaBoost algorithms. Australian workshop on health data and knowledge management, Wollongong, NSW, Australia. Retrieved October 9, 2010, from the Association for Computing Machinery (ACM) Digital Library.
[22] Mitchell, T. (1997). Machine Learning. San Francisco, CA: McGraw-Hill.
[23] Han, J., & Kamber, M. (2006). Data Mining: Concepts and Techniques. San Francisco, CA: Morgan Kaufman.
[24] Sambasiva Rao Voleti, Kiran Kumar Reddi, Classifiers Performance Improvement through Integration of Clustering Technique, International Journal of Advanced Research in Computer Science and Software Engineering, 2015, Vol.5, Issue 5, pp.808-813.

ABOUT AUTHOR

Sambasiva Rao Voleti obtained his M.Tech degree in Computer Science and Engineering from JNTU, Hyderabad. He is pursuing Ph.D in KrishnaUniversity, Andhra Pradesh. He has more than 15 years of teaching experience. He has around 10 publications in various national and international Journals. He is member of CSI.

Dr. Kiran Kumar Reddi obtained his Ph.D from Acharya Nagarjuna University, Guntur, Andhra Pradesh, and M.Tech from JNTU, Kakinada, Andhra Pradesh. At Present he is working as HOD, Dept. of CS, Krishna University, Machilipatnam, Andhra Pradesh. He has more than 18 years of teaching experience. He has more than 50 publications in various national and international journals. He is member of ISTE, IETE and CSI.

