



Data Mining Approach to Detect Heart Diseases

Authors

Vikas Chaurasia

Research Scholar, Sai Nath University,

chaurasia.vikas@gmail.com
Ranchi, Jharkhand, India

Saurabh Pal

Dept. of MCA, VBS Purvanchal University,

drsaaurabhpal@yahoo.co.in
Jaunpur, Uttar Pradesh, India

Abstract

Globally, heart diseases are the number one cause of death. About 80% of deaths occurred in low- and middle income countries. If current trends are allowed to continue, by 2030 an estimated 23.6 million people will die from cardiovascular disease (mainly from heart attacks and strokes).

The healthcare industry gathers enormous amounts of heart disease data which, unfortunately, are not "mined" to discover hidden information for effective decision making. The reduction of blood and oxygen supply to the heart leads to heart disease. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. This research paper intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques which will be useful for medical practitioners to take effective decision. The objective of this research work is to predict more accurately the presence of heart disease with reduced number of attributes. Originally, thirteen attributes were involved in predicting the heart disease. Thirteen attributes are reduced to 11 attributes. Three classifiers like Naive Bayes, J48 Decision Tree and Bagging algorithm are used to predict the diagnosis of patients with the same accuracy as obtained before the reduction of number of attributes. In our studies 10-fold cross validation method was used to measure the unbiased estimate of these prediction models.

Key Words

Bagging algorithm, Data Mining, Heart disease Diagnosis, J48 Decision Tree, Naïve Bayes.

I. INTRODUCTION

According to the World Health Organization heart disease is the first leading cause of death in high and low income countries and occur almost equally in men and women [1]. By the year 2030, about 76% of the deaths in the world will be due to non-communicable diseases (NCDs) [2].

Cardiovascular diseases (CVDs), also on the rise, comprise a major portion of non communicable diseases. In 2010, of all projected worldwide deaths, 23 million are expected to be because of cardiovascular diseases. In fact, CVDs would be the single largest cause of death in the world accounting for more than a third of all deaths [3]. For CVDs specifically, in 2005, the age standardized mortality rate for developing nations like India, China, and Brazil was between 300-450 per 100,000, whereas it was around 100-200 per 100,000 for developed countries like USA and Japan [4]. According to a recent study by the Registrar General of India (RGI) and the Indian Council of Medical Research (ICMR), about 25 percent of deaths in the age group of 25- 69 years occur because of heart diseases.

The core functionalities of data mining are applying various methods and algorithms in order to discover and extract patterns of stored data [5]. From the last two decades data mining and knowledge discovery applications have got a rich focus due to its significance in decision making and it has become an essential component in various organizations. The field of data mining have been prospered and posed into new areas of human life with various integrations and advancements in the fields of Statistics, Databases, Machine Learning, Pattern Reorganization and healthcare.

Medical Data mining in healthcare is regarded as an important yet complicated task that needs to be executed accurately and efficiently. Healthcare data mining attempts to solve real world health problems in diagnosis and treatment of diseases [6]. This research paper aims to analyze the several data mining techniques proposed in recent years for the diagnosis of heart disease. Many researchers used data mining techniques in the diagnosis of diseases such as tuberculosis, diabetes, cancer and heart disease in which several data mining techniques are used in the diagnosis of heart disease such as KNN, Neural Networks, Bayesian classification, Classification based on clustering, Decision Tree, Genetic Algorithm, Naive Bayes, Decision tree, WAC which are showing accuracy at different levels.

Each data mining technique serves a different purpose depending on the modeling objective. Naïve Bayes is one of the successful data mining techniques used in the diagnosis of heart disease patients [7-8]. Naive Bayes classifiers have works well in many complex real-world situations. Naive Bayes or Bayes Rule is the basis for many machine-learning and data mining methods. The rule is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. It learns from the “evidence” by calculating the correlation between the target (i.e., dependent) and other (i.e., independent) variables. By theory, this classifier has minimum error rate but it may not be case always. However, inaccuracies are caused by assumptions due to class conditional independence and the lack of available probability data. Observations show that Naïve Bayes performs consistently before and after reduction of number of attributes.

Bagging plays an important role in the field of medical diagnosis. Many research works in this aspect is depicted in related work. Bagging algorithms used to improve model stability and accuracy. Bagging works well for unstable base models and can reduce variance in predictions.

Boosting can be used with any type of model and can reduce variance and bias in predictions. Bagging means Bootstrap aggregation [9] an ensemble method to classify the data with good accuracy.

J48 Decision Tree is a popular classifier which is simple and easy to implement. J48 Decision Tree with reduced error. It requires no domain knowledge or parameter setting and can handle high dimensional data. Hence it is more appropriate for exploratory knowledge discovery. It still suffers from repetition and replication. Therefore necessary steps need to be taken to handle repetition and replication. The performance of decision trees can be enhanced with suitable attribute selection. Correct selection of attributes partition the data set into distinct classes. Observations show that Decision trees outperform the other two classifiers but take more time to build the model.

II. BACKGROUND

In the diagnosis of heart disease large number of work is carried out, researchers have been investigating the use of data mining techniques to help professionals. Many risk factors associated with heart disease like age, sex, chest pain, blood pressure, cholesterol, blood sugar, family history of heart disease, obesity, and physical inactivity. Knowledge of these risk factors medical professionals can diagnosis the heart disease in patients easily. Naive Bayes is an important data mining technique.

My Chau Tu's [10] compare the bagging with C4.5 algorithm, bagging with Naïve bayes algorithm to diagnose the heart disease the patient. Rajkumar and Reena investigated comparing naïve bayes, k-nearest neighbour, and decision list in the diagnosis of heart disease patients [11]. Cheung applied naive bayes classifier on the heart disease dataset [12]. Ratanamahatana and Gunopulos applied comparative study on bayesian classifier performs reliably better than naïve bayes [13]. Ramana, Babu et al. applied classification technique with bagging and boosting in the diagnosis of Liver disease [14].

Bagging algorithms used in many research work to improve model stability and accuracy of medical data set. My Chau Tu's [15] used bagging algorithm to identify the heart disease. Tsirogiannis's [16] applied bagging algorithm on medical databases. Kaewchinporn C's [17] presented a new classification decision tree with bagging and clustering. Pan wen [18] conducted experiments on ECG using decision tree algorithm C4.5 with bagging. Jinyan LiHuiqing Liu's [19] experimented on ovarian tumor data using C4.5 with and without bagging. Dong-Sheng Cao's [20] proposed a new decision tree in the area of chemometrics related to pharmaceutical industry. Liu Ya-Qin's [21] experimented on breast cancer data using C5 algorithm with bagging. Tan AC's [22] used C4.5 decision tree, bagged decision tree on cancerous micro array data, and compared the prediction.

Sitair-Taut et al. used the weka tool to investigate applying J48 Decision Trees for the detection of coronary heart disease. Tu et al. used the weka tool in the diagnosis of heart disease and applying J48 Decision Tree.

III. METHODOLOGY

In this paper we use the following data mining techniques:

A. Naïve Bayes

Naïve bayes is the data mining techniques that show success in classification in diagnosing heart disease patients [23, 24]. Naïve bayes is based on probability theory to find the most likely possible classifications [25,26, 27]. This algorithm uses the Bayes formula, which calculates the probability of a data record Y having the class label c_j :

$$P(\text{label} = c_j|Y) = \frac{P(Y|\text{label} = c_j) * P(c_j)}{P(Y)}$$

Dominator, $P(Y)$, can be safely eliminated as it does not depend on the label. The class label c_j , with the largest conditional probability value, determines the category of the data record. Let the actual values of features a_1, a_2, \dots, a_n for the data record Y be equal to $\bar{a}_1, \bar{a}_2, \dots, \bar{a}_n$. Assuming that the features are independent with respect to the class label, the above probability can be rewritten as follows:

$$P(\text{label} = c_j|Y) = P(c_j) * \prod_{i=1}^n P(a_i = \bar{a}_i|c_j)$$

where $P(a_i = \bar{a}_i)$ is the ratio of the samples that have value a_i for the i th feature, among all the samples with class label c_j and $P(c_j)$ is the ratio of the samples with class label c_j to all the available samples.

B. J48 Decision Tree

It is also based on Hunt's algorithm. J48 handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes, J48 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values. J48 uses Gain Ratio as an attribute selection measure to build a decision tree. It removes the biasness of information gain when there are many outcome values of an attribute. At first, calculate the gain ratio of each attribute. The root node will be the attribute whose gain ratio is maximum. J48 uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification.

C. Bagging

Bagging means Bootstrap aggregation [28] an ensemble method to classify the data with good accuracy. In this method first the decision trees are derived by building the base classifiers c_1, c_2, \dots, c_n on the bootstrap samples D_1, D_2, \dots, D_n respectively with replacement from the data set D. Later the final model or decision tree is derived as a combination of all base classifiers c_1, c_2, \dots, c_n with the majority votes.

Bagging can be applied on neural networks, Bayesian algorithms, Rule based algorithms, neural networks, Support vector machines, Associative classification, and Distance based methods and Genetic Algorithms. Applying bagging on classifiers especially on decision trees, Neural networks increases accuracy of classification. Bagging plays an important role in the field of heart disease diagnosis.

IV. HEART DISEASE DATA

The data used in this study is the Hungarian Institute of Cardiology. Heart disease data set available at <http://archive.ics.uci.edu/ml/datasets/Heart+Disease>. The data set has 76 raw attributes. However, all of the published experiments only refer to 11 of them. Consequently, to allow comparison with the literature, all the predictor and response variables which were derived from the database are given in Table I for reference. The data set contains 294 rows.

TABLE I: SELECTED HUNGARIAN INSTITUTE OF CARDIOLOGY

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	1 = male 0 = female
Cp	Discrete	Chest pain type: 1 = typical angina 2 = atypical angina 3 = non-angina pa 4 =asymptomatic
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar > 120 mg/dl: 1 = true 0 = false
Restecg	Discrete	Resting electrocardiographic results: 0 = normal 1 = having ST-T wave abnormality 2 =showing probable or define left ventricular hypertrophy by Estes'criteria
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise induced angina: 1 = yes 0 = no
Slope	Discrete	The slope of the peak exercise segment : 1 = up sloping 2 = flat 3= down sloping
Diagnosis	Discrete	Diagnosis classes: 0 = healthy 1= possible heart disease

V. DATA MINING MODEL

Experiments are conducted using Weka tool and the results are compared with bagging and without bagging using 10-fold cross validation. Weka is a collection of machine learning algorithms for data mining tasks. The classify panel enables the user to apply classification and

regression algorithms to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, or the model itself. The three classifiers like Naive Bayes, J48 Decision Tree and Bagging algorithm were implemented in WEKA. Since there is no separate evaluation data set, this is necessary to get a reasonable idea of accuracy of the generated model. These predictive models provide ways to predict whether a patient having heart disease or not.

VI. EXPERIMENTAL RESULT AND DISCUSSION

After applying the pre-processing and preparation methods, we try to analyze the data visually and figure out the distribution of values. Figure 1 depicts the distribution of values of Heart disease patients.

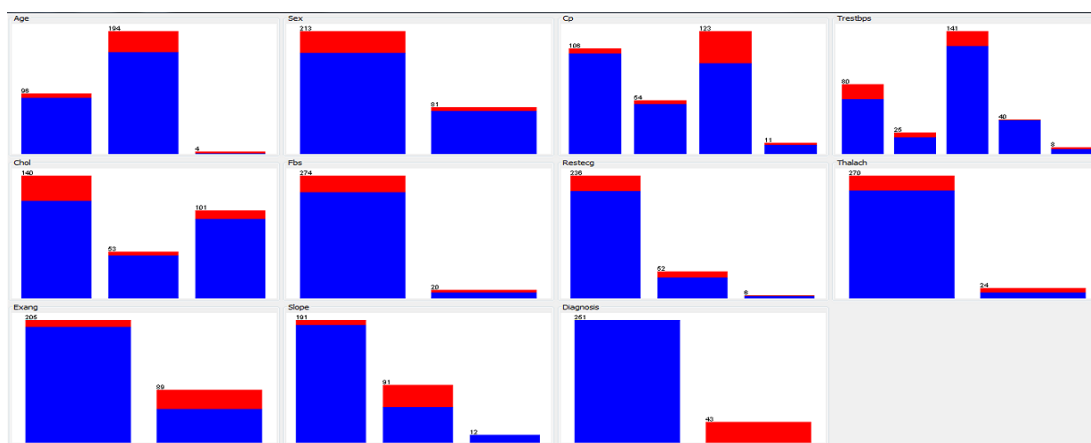


FIGURE 1: VISUALIZATION OF THE HEART PATIENTS

The results of the experiments are shown in table II. We have carried out some experiments in order to evaluate the performance and usefulness of different classification algorithms for predicting Heart Patients.

TABLE II: PERFORMANCE OF THE CLASSIFIERS

Evaluation Criteria	Classifiers		
	Naive Bayes	J48	Bagging
Timing to build model (in Sec)	0.02	0.05	0.05
Correctly classified instances	242	248	250
Incorrectly classified instances	52	46	44
Accuracy (%)	82.31%	84.35%	85.03%

As accuracy is very important in the field of medical domain, the performance measure accuracy of classification is considered in this study. So bagging classifier has more accuracy than other classifiers.

Kappa statistic, mean absolute error and root mean squared error will be in numeric value only. We also show the relative absolute error and root relative squared error in percentage for references and evaluation. The results are shown in Tables III.

TABLE III: TRAINING AND SIMULATION ERROR

Evaluation Criteria	Classifiers		
	Naive Bayes	J48	Bagging
Kappa statistic(KS)	0.4062	0.0942	0.0824
Mean absolute error(MAE)	0.1839	0.2161	0.21
Root mean squared error (RMSE)	0.3577	0.3455	0.334
Relative absolute error (RAE)	73.10%	85.89%	83.46%
Root relative squared error (RRSE)	101.20%	97.74%	94.50%

Here we check how accurate our predictive model is, it is necessary to check, the Accuracy of the predictive model is calculated based on the precision, recall values of classification matrix.

We have trained the classifiers to classify the medical data set as either “healthy” or “possible heart disease”. For the given two classes, we consider in terms of positive tuples (diagnosis =healthy) versus negative tuples (diagnosis = possible heart disease). True positives refer to the positive tuples that were correctly labeled by the classifier, while true negatives are the negative tuples that were correctly labeled by the classifier. False positives are the negative tuples that were incorrectly labeled by the classifier, while false negatives are the positive tuples that were incorrectly labeled by the classifier. The precision is used for the percentage of samples labeled as “healthy”. These measures are defined as

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Here true negatives (i.e sick samples that were correctly classified) and false positives (“possible heart disease samples that were incorrectly labeled as healthy).

Recall is fraction of relevant instances that are retrieved. It is usually expressed as a percentage. It is calculated as total number of true positives divided by total number of true positives + total number of false negatives.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

Comparison of detailed accuracy by class is shown in table IV.

TABLE IV: COMPARISON OF ACCURACY MEASURES

Classifier	TP	FP	Precision	Recall	Class
Naive Bayes	0.857	0.372	0.931	0.857	Healthy
	0.628	0.143	0.429	0.628	Possible Heart Disease
J48	0.972	0.907	0.862	0.972	Healthy
	0.093	0.028	0.364	0.093	Possible Heart Disease
Bagging	0.984	0.93	0.861	0.984	Healthy
	0.07	0.016	0.429	0.07	Possible Heart Disease

Execution of the learning techniques is highly dependent on the nature of the training data. Confusion matrices are very useful for evaluating classifiers. To evaluate the robustness of classifier, the usual methodology is to perform cross validation on the classifier. The columns represent the predictions, and the rows represent the actual class.

TABLE V: CONFUSION MATRIX

Classifier	Healthy	Possible Heart Disease
Naive Bayes	215	36
	16	27
J48	244	7
	39	4
Bagging	247	4
	40	3

Figures 2 are the graphical representations of the Classifiers.

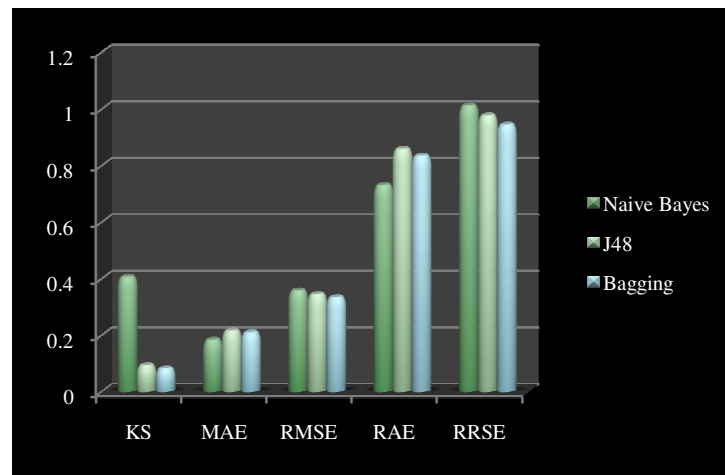


FIGURE 2: COMPARISON BETWEEN PARAMETERS

From the above Figures 2, and Table II, we can see that the highest accuracy is 85.03% and the lowest is 82.31%. The other algorithm yields an average accuracy of 84.35%. An average of 246 instances out of total 294 instances is found to be correctly classified with highest score of 250 instances compared to 242 instances, which is the lowest score. The total time required to build the model is also a crucial parameter in comparing the classification algorithm. In this simple experiment, from Table II, we can say that a Bagging, J48 requires around 0.05 seconds

compared to Naive Bayes which requires around 0.02 seconds.

Kappa statistic is used to assess the accuracy of any particular measuring cases, it is usual to distinguish between the reliability of the data collected and their validity [29].

The average Kappa score from the selected algorithm is around 0.08-0.40. Based on the Kappa Statistic criteria, the accuracy of this classification purposes is substantial [29]. From Figure 2, we can observe the differences of errors resultant from the training of the three selected algorithms. This experiment implies a very commonly used indicator which is mean of absolute errors and root mean squared errors. Alternatively, the relative errors are also used. Since, we have two readings on the errors, taking the average value will be wise.

VII. CONCLUSION

In medical diagnosis various data mining techniques are available. In this study, for classification of medical data we employed bagging algorithm because it produce human readable classification rules which are easy to interpret. Researchers have been investigating applying different data mining techniques to help health care professionals in the diagnosis of heart disease. Bagging algorithm is one of the successful data mining techniques used in the diagnosis of heart disease patients. This paper investigates experiments are conducted to find the best classifier for predicting the diagnosis of heart disease patients. This paper systematically investigates applying different methods of classifier technique in the diagnosis of heart disease patients. The results show that bagging algorithm accuracy of 85.03% and the total time taken to build the model is at .05 seconds in the diagnosis of heart disease patients. Finally, some limitations on this work are noted as pointers for future research.

The empirical results show that we can produce short but accurate prediction list for the heart patients by applying the predictive models to the records of incoming new patients. This study will also work to identify those patients which needed special attention.

REFERENCES

- [1] World Health Organization. 2007 7-February 2011]; Available from: <http://www.who.int/mediacentre/factsheets/fs310.pdf>.
- [2] Preventing Chronic Disease: A Vital Investment. World Health Organization Global Report. 2005
- [3] Global Burden of Disease. 2004 update (2008). World Health Organization.
- [4] Coronary Heart Diseases in India. Mark D Huffman. Center for Chronic Disease Control. http://sancd.org/uploads/pdf/factsheet_CHD.pdf
- [5] Fayadd, U., Piatetsky -Shapiro, G., and Smyth, P. 1996. From Data Mining To Knowledge Discovery in Databases, AAAI Press / The MIT Press, Massachusetts Institute Of Technology. ISBN 0-26256097-6 Fayap.
- [6] Liao, S.-C. and I.-N. Lee, Appropriate medical data categorization for data mining classification techniques. MED. INFORM., 2002. Vol. 27, no. 1, 59-67, .

- [7] Sitar-Taut, V.A., et al., Using machine learning algorithms in cardiovascular disease risk evaluation. *Journal of Applied Computer Science & Mathematics*, 2009.
- [8] Srinivas, K., B.K. Rani, and A. Govrdhan, Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. *International Journal on Computer Science and Engineering (IJCSE)*, 2010. Vol. 02, No. 02: p. 250-255.
- [9] L. Breiman, "Bagging predictors", *Machine Learning*, 26, 1996, 123-140.
- [10] My Chau Tu, Dongil Shin, Dongkyoo Shin, "Effective Diagnosis of Heart Disease through Bagging Approach", 2nd International Conference on Biomedical Engineering and Informatics, 2009.
- [11] Rajkumar, A. and G.S. Reena, Diagnosis Of Heart Disease Using Datamining Algorithm. *Global Journal of Computer Science and Technology*, 2010. Vol. 10 (Issue 10).
- [12] Cheung, N., Machine learning techniques for medical analysis. *School of Information Technology and Electrical Engineering, B.Sc. Thesis, University of Queensland.*, 2001.
- [13] Ratanamahatana, C.A. and D. Gunopulos, Scaling up the Naive Bayesian Classifier: Using Decision Trees for Feature Selection. *Proc. Workshop Data Cleaning and Preprocessing (DCAP '02)*, at IEEE Int'l Conf. Data Mining (ICDM '02), 2002.
- [14] Ramana, B.V., M.S.P. Babu, and N.B. Venkateswarlu, A critical evaluation of bayesian classifier for liver diagnosis using bagging and boosting methods. *International Journal of Engineering Science and Technology*, 2011. Vol. 3 No. 4.
- [15] My Chau Tu, Dongil Shin, Dongkyoo Shin, "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms" Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2009.
- [16] Tsirogiannis, G.L, Frossyniotis, D, Stoitsis, J, Golemati, S, Stafylopatis, A Nikita, K.S, "Classification of Medical Data with a Robust Multi-Level Combination scheme", IEEE international joint Conference on Neural Networks.
- [17] Kaewchinporn, C, Vongsuchoto, N, Srisawat, A "A Combination of Decision Tree Learning and Clustering for Data Classification", 2011 Eighth International Joint Conference on Computer Science and Software Engineering (JCSSE).
- [18] Pan Wen, "Application of decision tree to identify a abnormal high frequency electrocardiograph", *China National Knowledge Infrastructure Journal*, 2000.
- [19] Jinyan Li, Huiqing Liu, See-Kiong Ng and Limsoon Wong, "Discovery of significant rules for classifying cancer diagnosis data", *Bioinformatics* 19(Suppl. 2) Oxford University Press 2003.
- [20] Dong-Sheng Cao, Qing-Song Xu, Yi-Zeng Liang, Xian Chen, "Automatic feature subset selection for decision tree-based ensemble methods in the prediction of bioactivity", *Chemometrics and Intelligent Laboratory Systems*.
- [21] Liu Ya-Qin, Wang Cheng, Zhang Lu, "Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data", 3rd International Conference on Bioinformatics and Biomedical Engineering, 2009.
- [22] Tan AC, Gilbert D. "Ensemble machine learning on gene expression data for cancer classification", *Appl Bioinformatics*. 2003;2(3 Suppl):S75-83.

- [23] Andreeva, P., Data Modelling and Specific Rule Generation via Data Mining Techniques. International Conference on Computer Systems and Technologies - CompSysTech, 2006.
- [24] Sitar-Taut, V.A., et al., Using machine learning algorithms in cardiovascular disease risk evaluation. Journal of Applied Computer Science & Mathematics, 2009.
- [25] Wu, X., et al., Top 10 algorithms in data mining analysis. Knowl. Inf. Syst., 2007.
- [26] S. K. Yadav & Pal, S. 2012. Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification, World of Computer Science and Information Technology (WCSIT), 2(2), 51-56.
- [27] S. K. Yadav, B. K. Bharadwaj & Pal, S. 2011. Data Mining Applications: A comparative study for predicting students' performance, International journal of Innovative Technology and Creative Engineering (IJITCE), 1(12).
- [28] L. Breiman, "Bagging predictors", Machine Learning, 26, 1996, 123-140.
- [29] Kappa at <http://www.dmi.columbia.edu/homepages/chuangj/kappa>.

AUTHORS' BIOGRAPHY



Vikas Chaurasia is M.Sc. (Math) and MCA from UNSIET VBS Purvanchal University, U.P., India. Since 2010 he has been working as lecturer in the department of Pharmacy. He is presently working as Lecturer in Department of Pharmacy, KHBS College of Pharmacy, Jaunpur, U.P, and India. His area of research includes Data Mining, Cloud Computing, Network Security, Web Technologies, and Artificial Intelligence.



Saurabh Pal received his M.Sc. (Computer Science) from Allahabad University, UP, India (1996) and obtained his Ph.D. degree from the Dr. R. M. L. Awadh University, Faizabad (2002). He then joined the Dept. of Computer Applications, VBS Purvanchal University, Jaunpur as Lecturer. At present, he is working as Head and Sr. Lecturer at Department of Computer Applications.

Saurabh Pal has authored more than 40 numbers of research papers in international/national Conference/journals and also guides research scholars in Computer Science/Applications. He is an active member of CSI, Society of Statistics and Computer Applications and working as reviewer and member of editorial board for more than 15 international journals. His research interests include Image Processing, Data Mining, Grid Computing and Artificial Intelligence.