ISSN No. 0976-5697

Volume 8, No. 3, March – April 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Performance Analysis of Classification Algorithms in Predicting Diabetes

Meraj Nabi
Deparment of CS & IT
Maulana Azad National Urdu University
Hyderabad, India

Abdul Wahid
Deparment of CS & IT
Maulana Azad National Urdu University
Hyderabad, India,

Pradeep Kumar
Deparment of CS & IT
Maulana Azad National Urdu University
Hyderabad, India

Abstract: In Data mining classification is one the most important technique. Today, we have data in abundant from numerous sources but in order to get meaningful information from it is very tedious task. Machine learning algorithms to train classifiers to decode the meaningful information from the data, this analysis approach has gained much popularity in recent years. This paper explores evaluation performance of Naïve Bayes, Logistic Regression and Decision tree, Random forest using datasets (Pima Indian Diabetes data from UCI Repository). Naïve Bayes algorithm is depending upon likelihood and probability; it is fast and stable to data changes. Logistic Regression, calculate the relationship of each feature and weights them based on their impact on result. Random forest algorithm is an ensemble algorithm, fits multiple trees with subset of data and averages tree result to improve performance and control over-fitting. Decision tree can be nicely visualized uses binary tree structure with each node making a decision depending upon the value of the feature. This paper concludes with a comparative evaluation of Naïve Bayes, Logistic Regression, Decision tree and Random Forest in the context of Pima Indian Diabetes Dataset(take from UCI repository) in order to predict the diabetic patients.

Keywords: Naïve Bayes, Logistic Regression, Random Forest, Classification, Decision tree

I. INTRODUCTION

Machine learning is a powerful artificial intelligence tool that enables us to crunch petabytes of data and make sense of a complicated world. And it's transforming a wide variety of industries. It's becoming increasingly ubiquitous with more and more and more application that we can't even think of them. Most people probably already know that email provider uses a machine learning algorithm to identify spam. From past few years Google, Tesla and other are building self-driving system that will soon augment or replace the human drivers. And E-commerce giant like Amazon and technology companies like Braintree are using it in conjunction with other tools to stop credit card fraud. Mining is one of the most significant applications of Machine learning. More often during analyses or, possibly, when trying to establish relationships between multiple features people become prone to making mistakes. This makes them tiresome for them to find a solution to certain problems. Here comes Machine learning which can be utilized to apply successfully to these problems, improving the Data efficiency of the system and designs of machines. The same set of features represents every instance of any dataset that is used by machine learning algorithms. The features can be Continuous, Discreet or Binary. Whenthe instances of the dataset are provided withknown labels (corresponding correct output) that learning is known as Supervised Machine learning. While in case of Unsupervised learning instances are without known labels. By implementing these unsupervised (clustering) learning algorithm researchers anticipate to discover unknown, but useful classes of items.

In supervised learning, the gathered data after preprocessing feed to the algorithm which analyze the data and build a model which then predicts the result on new data, example problems are Classification and Regression. In contrast with Unsupervised learning the gathered data is un-label therefore algorithm analyze the data and create the model by deducing the structure present in the input data, example problems are clustering, reduction in dimensionality and learning rule for association. [1][2]

Reinforcement machine learning is another different kind of machine learning technique where learning is taking place by interacting with the environment. Here learner is not provided what actions to take, as in most forms of machine learning, but instead it discovers by it-self which actions produces the most reward by trying them. A reinforcement learning agent learns from the outcomes of its actions, rather than from being explicitly taught and selection of its actions depends upon its past experiences (exploitation) and also by new choices (exploration), which is essentially trial and error learning.[3][4]

In Robotics Reinforcement learning is common, where the collection of sensors readings at one point in time is a data point, and the algorithm decide the robot's next action. In the Internet of Things (IoT) applications also a natural fit into it.

Here in this paper, we are predicting if a person will develop diabetes and analyze the model created by using simple but powerful algorithms like Naïve Bayes, Logistic Regression and Decision Tree and Random Forest. The data we have chosen is from a Pima Indian Diabetes study.[5]

II. PROCESS WORK FLOW

The orchestrated and repeatable pattern which systematically transforms and processes information to create prediction solutions [4]. The format of this process is shown in Fig 1.

The first step is to define the problem by asking the right question to solve the problem i.e. identifying and defining it. Next step is collecting the Dataset. In order to identify the most

informative fields (attributes, features) we can take a suggestion from a requisite expert if available otherwise, we can apply Brute-Force means measuring everything available and expect right (informative, relevant) features can be isolated. However, a dataset collected from "brute-force" method is directly unsuitable for induction. It contains noise and missing features in most cases and therefore requires significant pre-processing.

The Next step is the data preparation and data preprocessing. Depending on the circumstances, Researchers have a number of methods to choose from to handle missing data. Instance selection is not only used to handle noise but to cope with the infeasibility from the very large dataset. Selection of instances in these dataset comes under optimization problem that strives to maintain the mining quality while minimizing the sample size.

We have used here Pima Indian diabetes data which is taken from UCI repository. The Dataset used in the study consist of 768 instances out of which 500 are negative tested for diabetes and 268 are positively tested. All the attribute and their types are shown in Table 1.

Byremoving the irrelevant and redundant features we make feature subset and this process termed as feature subset selection. This helps in reducing the dimensionality of the data and help to operate effectively and faster. In this paper, we have imputed missing data with mean and deleted the externous correlated features which contribute to the better comprehensibility of the produced classifier and the better understanding of the learned concept. [4][6].

The Prepared data after cleaning, used for training and testing the model. The data is split 70% for training and 30% for testing. Then we train the algorithm on 70% of the data set and keeping the test data aside. This training process will produce the training model based on the logic and the algorithm and the values of the features in the training data.

Then test the model on the unseen data to evaluate the model. If we trained the model on the entire set of data, then it produces a good result on the test data as it has seen the biases and when we use this model to the real world data, then it will perform poorly as it is unaware of the biases present in the real world data. Therefore we keep the testing data separated from training data so that it produces better results.

Table I: Diabetes Dataset Attributes.

S. No	Attributes	Type
1	Number of Times pregnant	Continuous
2	Plasma glucose concentration 2 hours in an oral glucose tolerance test	Continuous
3	Diastolic blood pressure (mm Hg)	Continuous
4	Triceps skin fold thickness (mm)	Continuous
5	2-Hour serum insulin (mu U/ml)	Continuous
6	Body mass index (weight in kg/(height in m)^2)	Continuous
7	Diabetes pedigree function	Continuous
8	Age (years)	Numeric
9	Class variable (0 or 1)	Discreet

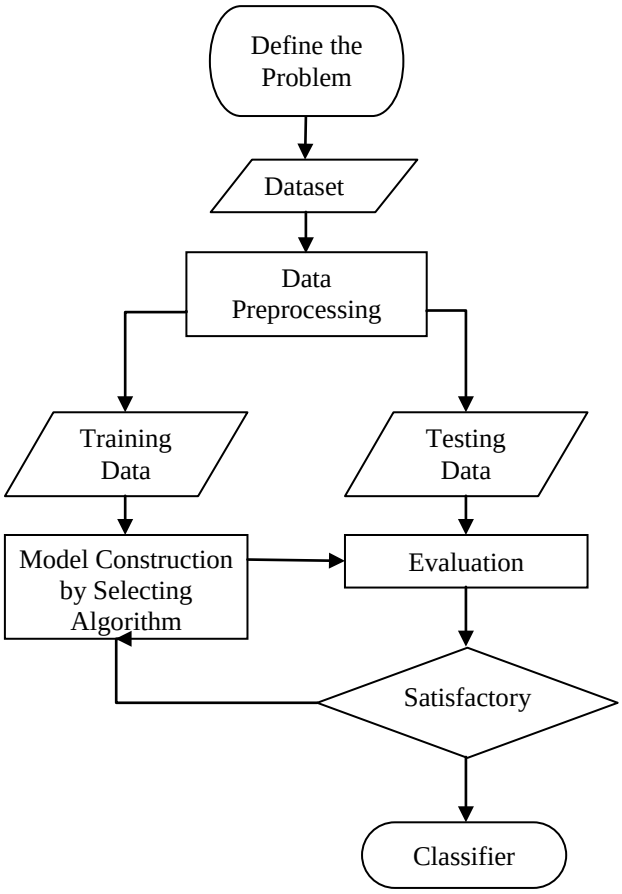

Fig 1: The Process of Conceptual Framework.

III. MODEL CONSTRUCTION

In the Next Step, Model Construction will take place using, Naïve Bayes, Logistic Regression Decision tree and Random Forest and their performance will be evaluated.

A. Naïve Bayes

The naïve Bayes algorithm is a simple probabilistic classifier that calculates a collection of probabilities by investigating frequency and combination of values in a given data set. The algorithm is based on applying Bayes theorem with the "naïve" assumption of independence between every pair of features.

Due to simple structure of Naive Bayes, construction of it is very simple and also has several advantages. Moreover, the inference (classification) is achieved in a linear time (while the inference in Bayes networks with a general structure is known to be NP-complete). Also, it does not require much larger data set smaller data set can also be used. Finally, the construction of naive Bayes is incremental, in the sense that it can be easily updated (namely, it is always easy to consider and take into account new cases in hand). [7][8]

Suppose Ci be diabetes risk group i and N be input variables that are used in a model and under the assumption of all variables are independent. To predict a class of diabetes risk, a model of Naive Bayes can be defined by

$$P(C_i|\,N) \quad = \quad \frac{P(N\,|C_i) \;\times\; P(C_i)}{P(N)} \qquad (1)$$

Where $P(C_i|\,N)$ is a posterior probability of a training data set with variable N that will beC_i.

B. Logistic regression

Despite its name, Logistic regressionis basically a linear model for classification rather than regression. It is also known as the logit regression, maximum-entropy classification (MaxEnt) or the log-linear classifier. In this model,we use logistic regression to model probabilistically described outcomes of a single trial. It is a basic model which describes dichotomous output variables and can be extended for disease classification prediction. [9][10]

Suppose there are N input variables where their values are indicated by m_1, m_2, m_3, …,m_N. Let us assume that the P probability of that an event will occur and 1- P be a probability that event will not occur. Logistic regression model is given by

$$\log\left(\frac{P}{1-p}\right) = \log it\,(P) = \beta_0 + \beta_1 m_1 + \cdots + \beta_N m_N \qquad (2)$$

Where β_0 is the intercept and $\beta_1, \beta_2, \ldots, \beta_N$ are regression coefficients.

C. Decision Tree

It creates a binary tree. The decision tree approach is most useful in the classification problem. With this technique, a tree is constructed to model the classification process. It consists of three types of nodes root node, child node, and leaf node. The algorithm starts with defining a root node from the most relationship between every input and output variables. Next, the child node is selected by calculating Information Gain (IG).

$\mathrm{IG(parent, child)} = \mathrm{Entropy(parent)} - [\mathrm{P}(x_1) \times \mathrm{Entropyx1} + \mathrm{Px1} \times \mathrm{Entropyx1} + \ldots 3$

$\mathrm{Entropy}(C_i) = -\mathrm{P}(x_i)\log \mathrm{P}(x_i)$ and $\mathrm{P}(x_i)$ is the probability of child node i. Node having the highest IG will become the parent for next generation. This process is repeated until it gets a leaf node and completed decision tree. The stopping criteria for decision tree is that all the sample for a given node belong to the same class, there aren't remaining attributes for any further partitioning and there aren't any leftover sample.

It requires little data preparation. While different techniques typically require data normalization, creation of dummy variables and blank values to be removed. Note however that this module does not support missing values.[11][12]

Decision trees tend to over-fit on data having a vastnumber of features. Obtaining the right ratio of samples to a number of features is important since a tree with few samples in high dimensional space is very likely to over-fit.

D. Random Forest

Random Forest is an ensemble algorithm which was modeled from trees algorithm and Bagging algorithm. It is developed by Breiman20, he found that the algorithm can potentially improve classification accuracy. It also works well with a data set with a vast number of input variables. The algorithm begins by creating a combinationof trees which each will vote for a class as shown in Fig. 2. The figure presents how to model the Random Forest. [8][13]-[16]

Suppose that there are X data and Y input variables in a data set where the real data used in this paper compose of 768 data and 9 input variables. Let z be the number of sampling groups, $x_i$and y_i be a number of data and variablesin group i where i is equal to 1, 2, ... and z. Each sampling group is as followed:

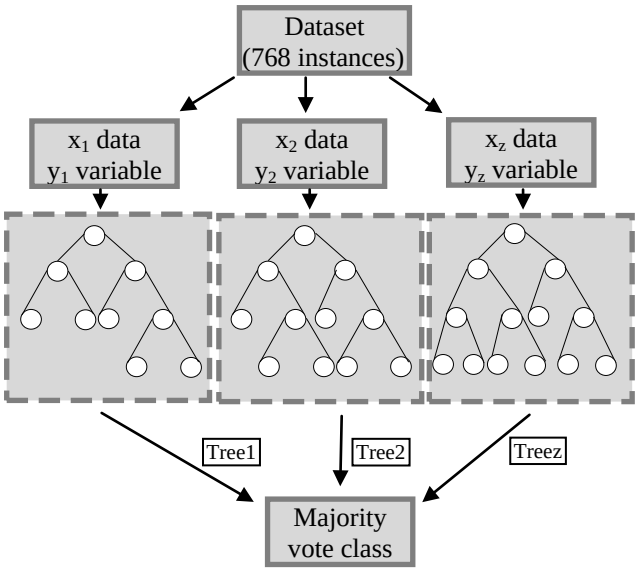

Fig 2: Random Forest.

$x_i$variables where $x_i$is not greater than X are selected randomly from X.

$y_i$variables where y_i is not greater than Y are selected randomly from Y.

A tree is grown and gives a prediction class.

After Step one to three was recurrent for z times, these trees become a forest. Then the classification will be elected by a majority vote of all trees within the forest.Note that all data have to be returned to the data set before selecting a newsampling group.

IV. PERFORMANCE EVALUATION CRITERIA FOR MODEL

To analyze and compare the performance of the data mining methods presented in our study, we apply various statistics such as MAE, RMSE, NRMSE and Confusion Matrixcomputed as follows.[17][18]

1) Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\,Predicted - Actual\,| \qquad (4)$$

MAE is the quantity used to measure how close predictionsare to the actual outcomes.

2) Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(Predicted - Actual)^2}{n}} \qquad (5)$$

Where n is the number of observations for correspondingpredicted and observed values of the model.

3) Normalized Root Mean Square Error (NRMSE)

$$NRMSE = \sqrt{\frac{\sum_{i=1}^{n}(Predicted\ -Actua\ )^2}{\sum_{i=1}^{n} Predicted\ ^2}} \quad (6)$$

NRMSE is a non-dimensional form of the RMSE, which is a normalized form of RMSE to the range of the observed data.

1. Confusion matrix

The information about actual and predicted classification system is hold by the Confusion matrix. It demonstrates the accuracy of the solution to a classification problem.

The table no. 2 shows the confusion matrix for a two class classifier. The entries in the confusion matrix have the following meaning in the context of our study.

t_p is the number of correct predictions that an instance is positive.

f_n is the number of incorrect predictions that an instance is negative.

f_p is the number of incorrect predictions that an instance is positive and

t_n is the number of correct predictions that an instance is negative.

Table II:Confusion Matrix

		Predicted	
		Positive	Negative
Actual	Positive	t_p	f_n
	Negative	f_p	t_n

4) Recall /True Positive Rate /Sensitivity

True positive rate (TP) is the proportion of positive cases that were correctly identified, as calculated using the equation:

$$R = \frac{t_p}{t_p + f_n} \quad (7)$$

5) Precision

Precision (P) is the proportion of the predicted positive cases that were correct,as calculated using the equation:

$$P = \frac{t_p}{t_p + f_p} \quad (8)$$

6) Accuracy

The proportion of the total number of predictions that were correct is known to be as Accuracy(AC). It shows overall effectiveness of classifier. It is determined using the equation:

$$AC = \frac{t_p + t_n}{t_p + f_n + f_p + t_n} \quad (9)$$

7) ROC

A receiver operating characteristics (ROC) graph is a method for conceptualize, organizing and selecting classifiers on the basis of their performance. ROC graphs are bi-dimensional graphs where on the Y axis t_p rate is plotted and on the X axisf_p rate is plotted. A ROC graph describe relative tradeoffs between benefits (true positives) and costs (false positives).

V. RESULTS

We are using the machine learning workflow to process and transform Pima Indian diabetes data to create prediction model.This model must predict which people are likely to develop diabetes, using Naïve Bayes, Logistic Regression, Decision tree (J48) and Random forest. The Performance of the algorithms is summarized in the table no. III.

Table III*: Summary of Prediction for different algorithms.*

Algoritms	CC	IC	MAE	RMSE	RAE	RRSE
Naïve Bayes	76.95	23.04	.2677	.3863	59.56	82.78
Logistic Regression	80.43	19.56	.2987	.3748	66.44	80.31
J48	76.52	23.47	.3206	.4239	71.33	90.85
Random Forest	76.52	23.47	.3095	.3884	68.86	83.24

VI. RESULT ANALYSIS

Table 2 shows Naïve Bayes classifier being the simplest classifier have performed well with an accuracy of 76.52%, while having relative absolute error 59.56%. The performance of Logistic Regression in classifying instances correctly is higher than Naïve Bayes, J48 and Random forest. Among the applied algorithms Logistic Regression has higher accuracy of about 80.43% which is quitwell and having the lowest RMSE value 37.48.Fig 4 shows comparative analysis of algorithm in terms of Mean Absolute Error,Root Mean Square Error.Fig 5 shows comparative analysis of algorithm in terms of Correctly Classified Instances, Incorrectly Classified Instances, Relative Absolute Error and Root Relative squared Error. Fig 6 compares Recall, Precision, Accuracy calculated using confusion matrix. Also, the area under the ROC is also compared for all the four applied algorithms namely Naïve Bayes, Logistic Regression, J48 and Random Forest. The ROC area of Logistic Regression is highest among Naïve Bayes, J48 and Random Forest. More the area covered better is the classifier. These measurements are taken by using Weka tool on Pima Indian Diabetes Data set taken from UCI repository. The results shown in Table 2 appears to be marginally better quantitatively in terms of accuracy and prediction capability. Further, the results may be improved by applying large size updated data sets of realistic context. However we need to apply other machine learning algorithms using updated data sets before generalized the results.

VII. CONCLUSION

In this paper, we have inspected the execution of four machine learning algorithms namely Naïve Bayes, Logistic Regression, J48 and Random forest to predict the populationwho are most likely to develop diabetes on Pima Indian diabetes data.The performance measurement is compared in terms of MAE and NRMSE obtained from the test setthe results are marginally better quantitatively in terms of accuracy and prediction capability. Further, we plan to recreate our study of Classification models by introducing the intelligent machine learning algorithms applied to a largecollection of real life data set.

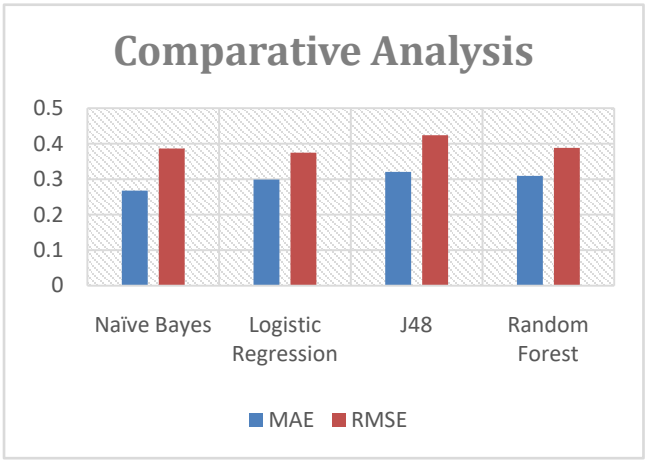

Fig 3: Comparative analysis of algorithms in terms of MAE, RMSE.

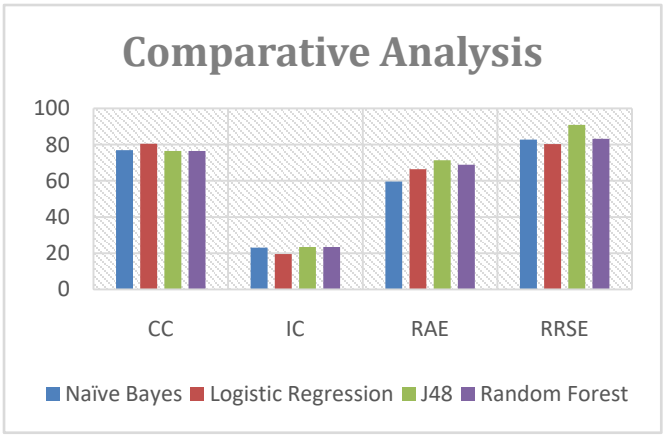

Fig 4: Comparative analysis of algorithms in terms of CC, IC, RAE, RRSE.

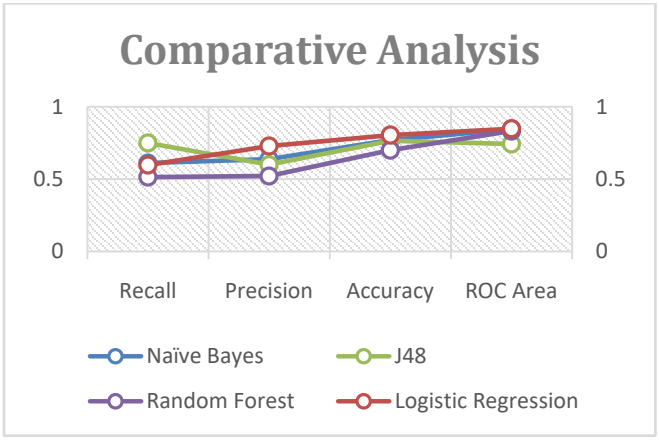

Fig 5:Comparative Analysis of algorithms in terms of Recall, Precision, Accuracy and ROC Area.

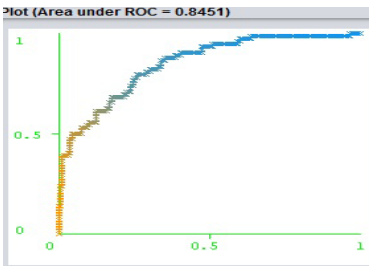

Fig 6: ROC plot for Naïve Bayes.

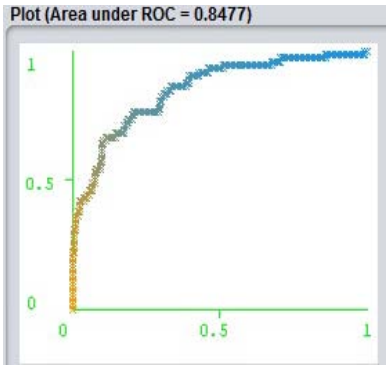

Fig 7: ROC plot for Logistic Regression.

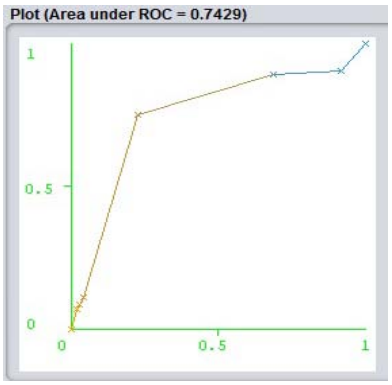

Fig 8: ROC plot for J48.

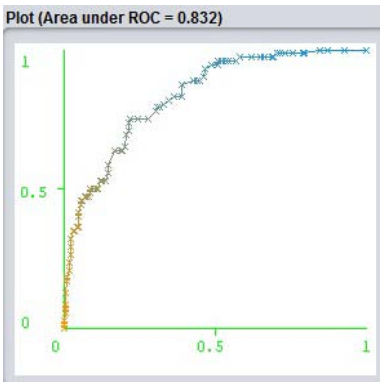

Fig 9: ROC plot for Random Forest.

REFERENCES

[1] Margaret H. Danham,S. Sridhar, "Data mining, Introductory and Advanced Topics", Person education , 1st ed., pp. 75-84,2006.

[2] Aman Kumar Sharma, SuruchiSahni, "A Comparative Study of Classification Algorithms for Spam Email Data Analysis", IJCSE, Vol. 3, No. 5, pp. 1890-1895,2011.

[3] Barto, A. G. & Sutton, R., "Introduction to Reinforcement Learning", MIT Press.M. Young, The Technical Writer's Handbook Mill Valley, CA: University Science, pp. 45-60,1997.

[4] S. B. Kotsiantis, I. D. Zaharakis, P. E. Pintelas, "Machine learning: a review of classification and combining techniques", Springer Science+Business Media B.V., ArtifIntell Rev, Vol. 26, pp. 159–190,2007.

[5] Leslie Pack Kaelbling, Michael L. Littman,"Reinforcement Learning:A Survey", Journal of Atificial Intelligence Research, Vol. 4, pp. 237-285,1996.

[6] Lei Yu, Huan Liu (). "Efficient Feature Selection via Analysis of Relevance and Redundancy", Journal of Machine Learning Research, Vol. 5,pp 1205–1224,2004.

[7] AnshulGoyal, Rajni Mehta, "Performance Comparison of Naive Bayes and J48 Classification Algorithms", IJAER, Vol. 7, No 11, pp. 281-297,2012.

[8] NongyaoNai-arun, RungruttikarnMoungmai, "Comparison of Classifiers for the Risk of Diabetes Prediction", Original Research Article Procedia Computer Science, Vol. 69, pp. 132-142,2015.

[9] Agresti A, "An Introduction to Categorical Data Analysis". 2nd ed. New York: Wiley; 1996.

[10] Tabaei B, Herman W(). "A Multivariate logistic regression equation to screen for diabetes",Diabetes Care, Vol.25, pp. 1999–2003,2002.

[11] Quinlan JR, "Induction of decision tree". Machine Learning 1, Khuwer Academic Publisher, pp. 81-106,1986.

[12] Han J, KanberM. Pei J, "Data Mining: Concepts and Techniques", 3rd ed. USA: Morgan Kaufman; 2012.

[13] Ali J, Khan R, Ahmad N, MaqsoodI(). "Random forests and decision tree", Journal of Computer Science, 9(5), pp. 272-278,2012.

[14] Sittidech P, Nai-arun N, "Random Forest Analysis on Diabetes Complication Data", Proceeding of the IASTED International Conference, pp. 315-320,2014.

[15] Kellie J, Archer, Ryan VK, "Empirical characterization of random forest variable importance measures", Journal of Computational Statistics & Data Analysis, Vol. 52, pp. 2249-2260,2008.

[16] Verikas A, Gelzinis A, Bacauskiene M. "Mining data with random forests: A survey and Results of new tests", Journal of Pattern Recognition, Vol. 44,pp 330-349.2011.

[17] Pradeep Kumar, Abdul Wahid . "Performance Evaluation of Data Mining Techniques for Predicting Software Reliability", World Academy of Science, Engineering and Technology,2015.