PREDICTION AND ANALYSIS OF DIABETIC DATA USING MACHINE LEARNING TECHNIQUES

[1]Monika , [2]Pooja Sharma
[1]Student in Dept of CSE at IKG Punjab Technical University main campus Kapurthala, Punjab,India
[2]Assistant Professor in Dept of CSE at IKG Punjab Technical University main campus Kapurthala, Punjab,India

Abstract **–** Data mining is a procedure of getting the data from a dataset and changes it into unambiguous structure. Medicinal information mining has been a incredible capacity for finding concealed examples in the informational indexes of the medicinal area. Diabetes is one of the major worldwide wellbeing issues. There are numerous Data digging methods for the forecast of maladies like heart diseases, cancer, kidney stones, etc. Prediction of Diabetes is a developing and quickest developing innovation. Utilizing different Data mining strategies we can foresee Diabetes from the informational index of a patient.. This paper proposed a technique for prediction of diabetes patients by combining Gin i index-based feature selection with balanced random forest. Experimental results demonstrate that proposed technique outperforms the existing techniques**.**

Keywords **–Clinical Databases, Diabetes, Prediction, Gini Index, Feature Selection, Balanced Random Forest.**

I. INTRODUCTION

Data Mining is utilized to create learning out of information and displaying it in a condition that is effortlessly reasonable to people. It is a procedure to investigate a lot of information gathered. Data innovation assumes a fundamental part to implement the Data mining procedures in different areas like managing an account, instruction, and so forth. In the field of medicinal area information mining can be successfully utilized for the forecast of maladies by utilizing different information mining methods. There are two overwhelming objectives of information mining have a tendency to be expectation and portrayal. Expectation includes a few factors or fields in the informational collection to anticipate obscure or future estimations of different factors of intrigue. Portrayal centers around finding the examples specifying the information that can be translated by people. Fundamental origination of development and qualities influencing diabetes from outer sources is particularly basic before building prescient models. The thought is to foresee the diabetes and to discover the components in charge of diabetes utilizing information mining strategies. Information mining methods can be utilized for early expectation of the infection with more prominent quality keeping in mind the end goal to spare the human life and it will likewise lessen the treatment cost. As indicated by International Diabetes Federation expressed that 382 million individuals are influenced with diabetes around the world. By 2035, this will be served as 592 million [1].

Diabetes is one of the primary points for medicinal research because of the life span of the diabetes and the tremendous cost on the social insurance suppliers. Early identifying of diabetes eventually diminishes cost on social insurance suppliers for treating diabetic patients, yet it is a testing undertaking. For early distinguishing of diabetes, specialists can exploit the patient's social insurance information to change over crude information into important data and concentrate shrouded learning by applying information mining, for example [2], choice tree or SOM to build a wise prescient model.

Rest of the paper is organized as follows. The Section 2 discuss the related work, In Section 3 proposed methodology is presented followed by the evaluation measures, Section 4 show the experimental result, Section 5 paper concluded with future work.

II. RELATED WORK

Shivakumar et al. gives a study of information mining techniques that have been ordinarily connected to Diabetes information examination and forecast of the malady. Different information mining strategies enable diabetes to explore and at last enhance the nature of human services for diabetes patients [4]. Yuvaraj et al. proposes the novel usage of machine learning calculations in hadoop based groups for diabetes expectation. Human services frameworks are simply intended to address the issues of expanding populace all around. Individuals around the world are influenced with various kinds of deadliest maladies. Among the diverse kinds of ordinarily existing illnesses, diabetes is a noteworthy reason for visual deficiency, kidney disappointment, heart assaults, and so on. Medicinal services observing frameworks for various illnesses and indications are accessible all around the globe [5].

MacDougall et al. brings the use of research based evidence into practice so as to develop clinical guidelines into practice. This paper provides a review of current research on the integration of Health Information Technology (HIT) into clinical guidelines so as to achieve more accurate results [6]. Srinivas et al. uses data mining application techniques in the Healthcare and the prediction of Heart Attacks. The author has deeply examined the use of data mining techniques in classification such as Rule based

technique, Decision Tree technique, Naïve Bayes Classification technique and Artificial Neural Network technique for the extraction of huge amount of patterns from the abundant data which is not mined so as to discover the hidden information from the data [7]. Sundar et al. finds out the accuracy of the result by using the K-means Clustering Algorithmic Technique for the prediction and diagnosis of Heart disease. It uses two datasets – real and artificial datasets. The real dataset is the dataset taken from the real life patients of hospitals and patients of laboratory tests whereas the artificial dataset is the dataset taken from the UCI machine learning Databases, 2004 [8].

Chaudhari et al. explores applying KNN to help human services experts in the determination of illness uniquely coronary illness. It additionally explores if coordinating voting with KNN can improve its exactness in the analysis of coronary illness patients. The outcomes demonstrate that applying KNN could accomplish higher exactness than neural system troupe in the analysis of coronary illness patients. The outcomes additionally demonstrate that applying voting couldn't improve the KNN exactness in the finding of coronary illness [9]. Mythili et al. proposes a technique that means to facilitate the patients experiencing different therapeutic tests, which the vast majority of them consider as a dull assignment and tedious. The parameters distinguished for diagnosing diabetes have been planned such that, the client can foresee on the off chance that he is influenced with diabetes himself. Back Propagation calculation is utilized for determination [10].

Ahmed et al. proposed to discover the heart illnesses through information mining, Support Vector Machine (SVM), Genetic Algorithm, harsh set hypothesis, affiliation rules and Neural Networks [11]. Thangaraju et al. considers the examination of diabetes guaging approaches utilizing grouping methods. Here creators are utilizing three various types of grouping systems named as Hierarchical bunching; Density based bunching, and Simple K-Means bunching. Weka is utilized as an instrument [12]. Durairaj et al. directed a point by point study on the utilization of various delicate figuring systems for the expectation of diabetes. This review is intended to distinguish and propose a successful method for prior forecast of the illness [13].

III. PROPOSED METHODOLOGY

The dataset used, type of feature selection technique used, and Balanced Random Forest algorithm follow by the evaluation measures is discussed in the following sections.

1. **Collection of data**
   Dataset were mainly collected from UCI repository and from various hospitals.

2. **Preprocessing and Filtering**
   In preprocessing step, it selects an attribute for selecting a subset of attribute so that it can provide good predicted capability. It also contains the conversion of data types like numeric to nominal or vice versa. It handles all the missing values and remove them. If an attribute contains more than 5% missing values, then the records should not be deleted and it is advised to put the values where the data is missing using some suitable methods and helps in feature selection and class label .

3. **Feature Selection**
   The proposed Gini index feature selection addresses the issues of uneven distribution of prior class probability and global goodness of a feature in two stages. First, it transforms the samples space into a feature specific normalized samples space without compromising the intra-class feature distribution. In the second stage of the framework, it identifies the features that discriminates the classes most by applying Gini coefficient of inequality.
   The specific algorithm: Suppose the collection of data samples is S of s having m different values of class label attribute which defines different classes of C_i, (i = 1; 2...; m). According to the class label attribute
   value, S can be divided into m subsets (S_i, i = 1; 2...; m). If Si is the subset of samples which belongs to class C_i, and s_i is the number of the samples in the subset Si, then the Gini Index of set S is

$$GiniIndex(S) = 1 - \sum_{i=1}^{m} {P_i}^2$$

   Where Pi is the probability of any sample of C_i, which is estimated by s_i/s. When the minimum of GiniIndex (S) is 0, i.e. all records belong to the same category at this collection, it indicates that the maximum useful information can be obtained. When all the samples in this collection have uniform distribution for a certain category, GiniIndex(S) reaches maximum, indicating the minimum useful information obtained. The initial form of the Gini Index is used to count the "impurity" of attribute for classification. The smaller its value, i.e. the lesser the "impurity", the better the attribute. On the other hand,

$$GiniIndex(S) = \sum_{i=1}^{m} {P_i}^2$$

   measuring the "purity" of attribute for categorization, the bigger its value, i.e. the better the "purity", the better the attribute.

4. **Classification**:
   Classification is a technique for machine learning by which it is used to predict the grouping membership of different data instances. It will perform the task by which it will generalize the well-known structure so as to apply it on new data. Here Random forest classifier has been used for quality measurement of dataset will be consider on the basis of

percentage of correctly classified instances. For validation phase we use 10-fold cross validation method. Random forest classifier helps in identifying the characteristics of patient with Diabetes diseases.

In this work prediction of diabetes is done by combining Gini index-based feature selection with balanced random forest. Results of the proposed technique is evaluated on basis of various parameters and are compared with decision tree, naive bayes, and random forest algorithm.

Decision tree:

The Decision tree is one of the classification techniques wherein class is completed by way of the splitting standards. The choice tree is a drift chart like a tree structure that classifies times by means of sorting them based on the attribute (feature) values. Each and every node in a decision tree represents an attribute in an instance to be categorized. All branches represent an outcome of the test, each leaf node holds the magnificence label. The example is classified based totally on their characteristic cost [3].

Naïve Bayes:

Naive Bayes classifier is a honest and effective set of rules for the classification mission. It works on conditional opportunity. Conditional possibility is the possibility that something will manifest, for the reason that something else has already took place. Using the conditional possibility, we can calculate the opportunity of an event the use of its previous expertise. Naive Bayes is a kind of classifier which makes use of the Bayes Theorem. It predicts membership chances for each elegance along with the chance that given report or statistics factor belongs to a specific elegance. The elegance with the very best chance is considered as the most in all likelihood magnificence.

$$P(H|\boldsymbol{X}) = \frac{P(H|\boldsymbol{X})P(H)}{P(X)}$$

Random Forest:

The random forest is a group approach that can likewise be thought of as a type of closest neighbor predictor. Random forest is only a change over the highest point of the choice tree calculation. The center thought behind Random Forest is to produce different little decision trees from irregular subsets of the information (subsequently the name "Random Forest"). Every one of the decision tree gives a one-sided classifier (as it just thinks about a subset of the information). They each catch diverse patterns in the information. This group of trees resembles a group of specialists each with a little information over the general subject however thourough in their specialized topic.

Balanced Random Forest

1. For each iteration in random forest, draw a bootstrap sample from the minority class. Randomly draw the same number of cases, with replacement, from the majority class.
2. Induce a classification tree from the data to maximum size, without pruning. The tree is induced with the C4.5 Tree algorithm, with the following modification: At each node, instead of searching through all variables for the optimal split, only search through a set of m- try randomly selected variables.
3. Repeat the two steps above for the number of times desired. Aggregate the predictions of the ensemble and make the final Prediction.

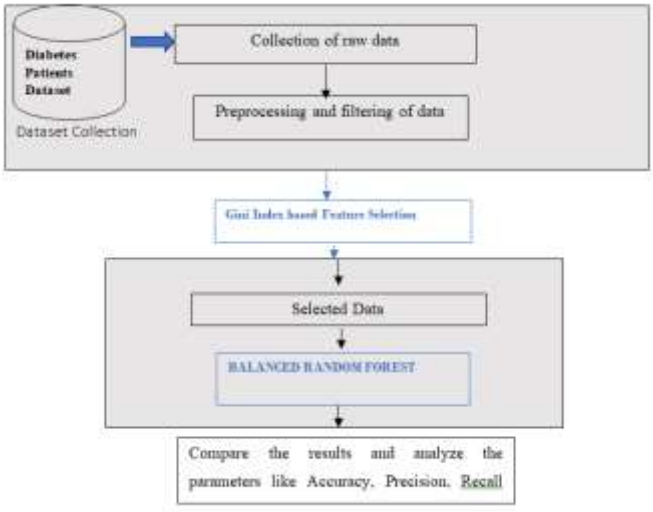

Figure 1: Block Diagram of Proposed Technique

Table 1: Contingency table for evaluation

	Actual label (expectation)	
Predicted label (observation)	a (true positive) correct result	b (false positive) unexpected result
	c (false negative) missing result	d (true negative) correct result

Evaluation Measures

Proposed technique (Gini index with balanced random forest) is evaluated on the basis of following parameters:-

- Correctly Classified Instances
- Incorrectly Classified Instances
- Kappa Statistics
- Mean Absolute Error
- Root Mean Squared Error
- Accuracy
- Precision
- Recall
- F-Measure
-

Precision and recall

Precision and recall are the two measurements that are generally to evaluate execution in content mining, and in content examination field like data recovery. These parameters are utilized for estimating exactness and completeness respectively.

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

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

F-measure

F-Measure is the harmonic mean of precision and recall. The esteem computed utilizing F-measure is a balance among precision and recall.

$$\text{F-measure} = \frac{2*\text{recall}*\text{precision}}{\text{precision+recall}}$$

Accuracy

Accuracy is the basic measure for arrangement execution. Exactness can be estimated as accurately ordered occurrences to the aggregate number of occasions, while mistake rate utilizes inaccurately arranged cases rather than effectively grouped examples.

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

Mean Absolute Error (MAE)

The MAE calculates the standard size of the mistake in an display of gauges, lacking thinking about their bearing.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

Root Mean Squared Error (RMSE)

The RMSE is a quadratic scoring guideline that calculates the normal size of the mistake. The condition for the RMSE is provided in both of the references.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

IV. EXPERIMENTAL RESULTS

This section presents experimental results of the proposed technique. Proposed technique is implemented using weka tool.

Weka: Weka (Waikato Environment for Knowledge Analysis) is a famous suite of machine learning programming written in Java, created at the University of Waikato, New Zealand. Weka is free programming accessible under the GNU General Public License. The Weka workbench contains a gathering of perception devices and calculations for information examination and prescient displaying, together with graphical UIs for simple access to this usefulness. Weka is a gathering of machine learning calculations for taking care of certifiable data mining issues. It is composed in Java and keeps running on any stage. The calculations can either be connected specifically to a dataset or called from your own Java code.

Table 2: Comparison of proposed technique with existing technique based on correctly classified instances and incorrectly classified instances

Parameter/ Algorithm	Decision Tree	Naive Bayes	Random Forest	Proposed Algorithm
Correctly Classified Instances	570	588	662	720
Incorrectly Classified Instances	198	180	106	48

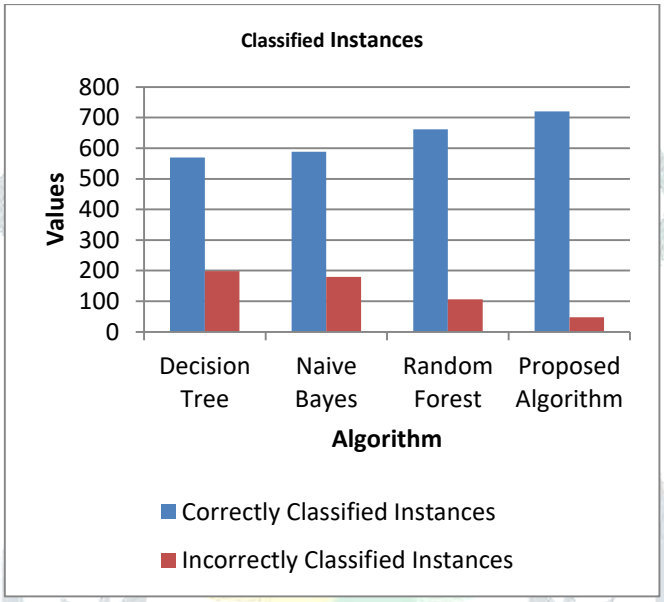

Figure 2: Showing comparison of proposed technique with existing technique based on correctly classified instances and incorrectly classified instances

Table 3: Comparison of proposed technique with existing technique based on kappa statistics

Decision Tree	Naive Bayes	Random Forest	Proposed Algorithm
0.4216	0.4723	0.6962	0.8663

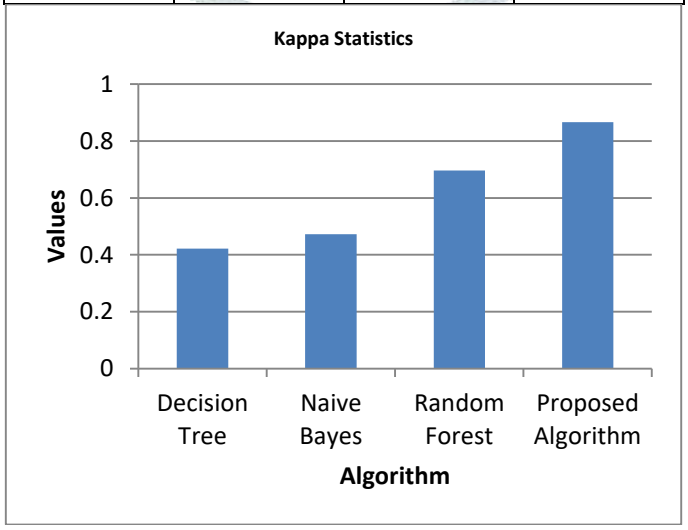

Figure 3: Showing comparison of proposed technique with existing technique based on kappa statistics

Table 4: Comparison of proposed technique with existing technique based on mean absolute error and root mean squared error

Error/ Algorithm	Decision Tree	Naive Bayes	Random Forest	Proposed Algorithm
Mean Absolute Error	0.3152	0.2853	0.2139	0.1108
Root Mean Squared Error	0.441	0.418	0.321	0.2234

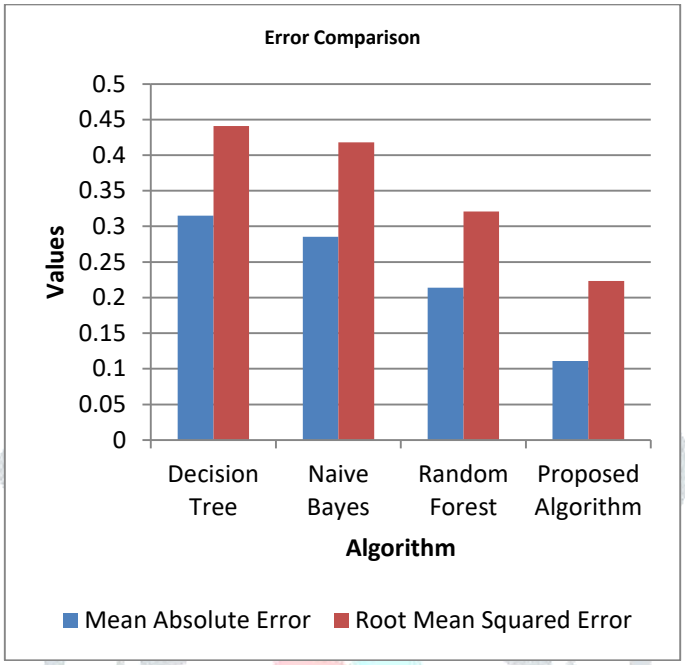

Figure 4: Showing comparison of proposed technique with existing technique based on mean absolute error and root mean squared error

Table 5: Comparison of proposed technique with existing technique based on accuracy

Decision Tree	Naive Bayes	Random Forest	Proposed Algorithm
74	77	86.20	93.75

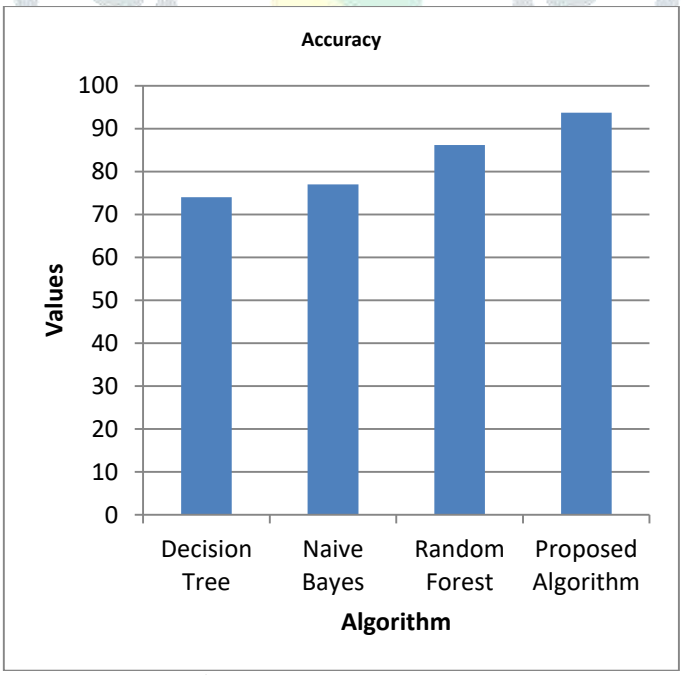

Figure 5: Showing comparison of proposed technique with existing technique based on accuracy

Table 6: Comparison of proposed technique with existing technique based on class parameters

Parameter/ Algorithm	Decision Tree	Naive Bayes	Random Forest	Proposed Algorithm
Precision	0.642	0.682	0.846	0.903
Recall	0.59	0.616	0.762	0.929
F-Measure	0.615	0.647	0.801	0.916

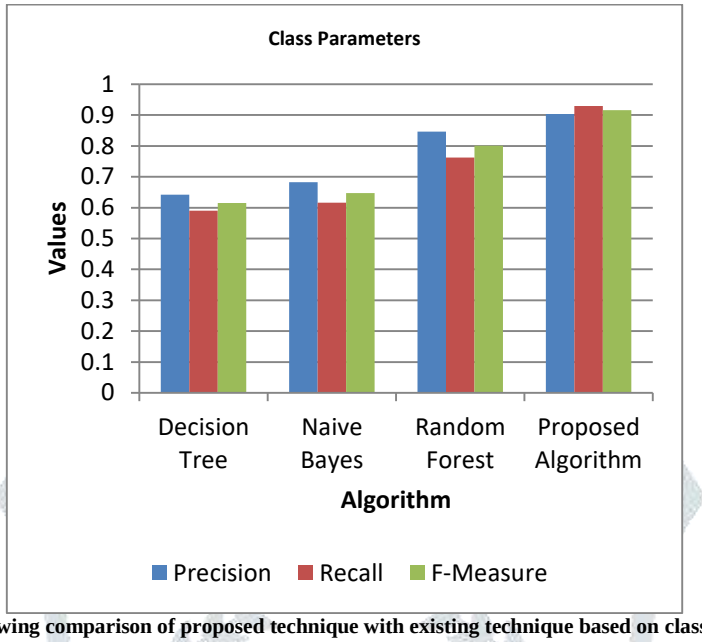

Figure 6: Showing comparison of proposed technique with existing technique based on class parameters

V. CONCLUSION AND FUTURE WORK

This paper proposed a technique for prediction of diabetes patients by combining Gini index-based feature selection with balanced random forest. Dataset were mainly collected from UCI repository and from various hospitals. Various parameters are used to evaluate the performance of the proposed technique, like accuracy, precision, recall, f-measure, kappa statistics, correctly classified instances, incorrectly classified instances, mean absolute error, and root mean squared error. Proposed algorithm is compared with existing algorithms such as decision tree, naive bayes and random forest. Results show that proposed technique has more accuracy, precision, recall, f-measure, kappa statistics and correctly classified instances and less error and incorrectly classified instances as compared to existing techniques. This work focuses on prediction of diabetes patients, in future, we may also analyse medical data using this technique.

REFERENCES

[1] K.Priyadarshini, Dr.I.Lakshmi (2017) "A Survey on Prediction of Diabetes Using Data Mining Technique", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 6, Issue. 11, pp. 369-373.
[2] Tahani Daghistani,Riyad Alshammari (2016) "Diagnosis of Diabetes by Applying Data Mining Classification Techniques", International Journal of Advanced Computer Science and Applications, Vol. 7, No. 7, pp. 329-332.
[3] Uppin ShravanKumar and M A Anusuya (2014) "Expert System design to predict Heart and Diabetes Diseases", International Journal of Scientific Engineering and Technology Vol: 03.
[4] B.L.Shivakumar, S. Alby (2014) "A Survey on Data-Mining Technologies for Prediction and Diagnosis of Diabetes", IEEE, International Conference on Intelligent Computing Applications.
[5] N. Yuvaraj, K. R. SriPreethaa (2017) "Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster", Springer, Cluster Computing,.
[6] Mac Dougall Candice, Percival Jennifer and Mc Gregor Carolyu (2009) "Integrating Health Information Technology into Clinical Guidelines", Annual International Conference of the IEEE, EMBS Minneapolis, Minnesota, USA, September 2-6.
[7] Srinivas K, Kavihta Rani B. and Dr. Govrdhan A (2010) "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", International Journal on Computer Science and engineering Vol. 02, No. 02, pp. 250-255.
[8] Sundar V Bata and Tevi T, Saravanan N (2012) "Development of a Data Clustering Algorithm for Predicting Heart", International Journal of Computer Applications(0975-888) Volume 48, No. 7, June, Coimbatore, India.
[9] Anand A. Chaudhari, Prof.S.P.Akarte (2014) " Fuzzy and Data Mining based Disease Predection using K-NN Algorithm", International Journal of Innovations in Engineering and Technology, Vol. 3, Issue No. 4, April.
[10] Prof.Sumathy, Prof.Mythili, Dr.Praveen Kumar, Jishnujit T M, K Ranjith Kumar(2010) "Diagnosis of Diabetes Mellitus based on Risk Factors", International Journal of Computer Applications, Vol.10, Issue No.4, November.
[11] Aqueel Ahmed, Shaikh Abdul Hannan (2012) " Data Mining Techniques to Find Out Heart Diseases: An Overview", International Journal of Innovative Technology and Exploring Engineering, Vol. 1, Issue No. 4, September.
[12] P. Thangaraju, B.Deepa, T.Karthikeyan (2014) "Comparison of Data mining Techniques for Forecasting Diabetes Mellitus", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue No. 8, August.
[13] M. Durairaj, G. Kalaiselvi (2015) " Prediction Of Diabetes Using Soft Computing Techniques- A Survey", International Journal of Scientific & Technology Research, Vol. 4, Issue No.3, March.