

Comparing Naive Bayes and Decision Tree Techniques for Predicting the Risk of Diabetic Retinopathy

G. Parthiban and Dr. S.K.Srivatsa

*Research Scholar, Dr. MGR Educational Research and Institute,
MGR University, Chennai, India.*

*Sr. Professor, Dept. of E & I, Prathyusha Institute of Technology and Management,
Chennai, India.*

trgparthi@gmail.com, profsks@rediffmail.com

Abstract

Classifying data is a common task in Machine learning. Data mining in health care is an emerging field of high importance for providing prognosis and a deeper understanding of medical data. Most data mining methods depend on a set of features that define the behaviour of the learning algorithm and directly or indirectly influence the complexity of resulting models. Diabetic retinopathy the most common diabetic eye disease, is caused by complications that occurs when blood vessels in the retina weakens or distracted. We have applied machine learning methods to predict the early detection of eye disease diabetic retinopathy and found that Decision Tree method to be 90% accurate. The performance was also measured by sensitivity and specificity.

Keywords - Data Mining, Naïve Bayes Method, Decision Tree, Diabetes, Diabetic Retinopathy.

I. INTRODUCTION

To extract hidden patterns and relationships from large data bases, Data mining merges statistical analysis, machine learning and database technology.[1] Diabetes is a chronic disease which causes serious health complications including heart disease, kidney failure and blindness. [2]. The commonest cause of blindness among working class is Diabetic Retinopathy which often leads to the complete loss of vision. [3]. The World Health Organization (WHO) has estimated that Diabetic Retinopathy is responsible for 4.8% of the 37 million cases of blindness throughout the world. Therefore a prediction technique is conceived so that early precautions or controls can

be implemented. People with diabetes are susceptible to impairment of other vital organs such as heart, kidney and eyes. [4].

In the first stage which is called non-proliferative diabetic retinopathy there are no symptoms, it is not visible to the naked eye. On the second stage, as abnormal new blood vessels form at the back of the eye as a part of proliferative diabetic retinopathy, they can burst and bleed. Image analysis tools can be used for automated detection of these various features and stages of Diabetes Retinopathy and can be referred to the specialist accordingly for intervention. Thus such tools will be useful for effective screening of Diabetic Retinopathy patients [5]. Therefore, a prediction technique has been conceived so that early precautions or controls can be implemented.

‘Diabetes Data Analysis and Prediction Model Discovery Using Rapid Miner’[6] analyse a Pima Indians diabetes data set containing information about patients with and without diabetes. This work focuses on data pre-processing, including attribute identification and selection, outlier removal, data normalization and numerical discretization, visual data analysis, hidden relationships discovery, and a diabetes prediction model construction.

IHDPS [7] prototype predicts the possibility of patients getting a heart disease from the Cleveland heart disease database using data mining techniques decision trees, naïve Bayes and neural network with 9 medical attributes. The results show that the most effective model to predict patients with heart diseases is naïve Bayes (86.12%) followed by neural network and decision trees. Furthermore, it can incorporate other data mining techniques such as time series, clustering and association rules.

‘Empirical Study on the Performance of Integrated Hybrid Prediction Model on the Medical Datasets’[8] system has been proposed to improve the diagnostic accuracy of diabetic disease by selecting informative features of Pima Indians Diabetes dataset. The hybrid prediction model proposed combines two different functionalities of data mining clustering and classification with F-score selection approach to identify the optimal feature subset of the Pima Indians Diabetes dataset. The proposed model was validated using four parameters, namely the accuracy of the classifier, area under ROC curve, sensitivity and specificity. The two traditional classification methods (logistic regression and Fisher linear discriminant analysis) and four machine-learning classifiers (neural networks, support vector machines, fuzzy c-mean, and random forests) were compared [9] to classify persons with and without diabetes.

During the recent years there have been many studies on automatic diagnosis of diabetes, diabetic retinopathy, heart disease etc., In [10] a method has been proposed for automated detection and classification of vascular abnormalities using several techniques such as scale and orientation, selective Gabor filter banks, In [11] Kaplan-Meier method to generate univariate survival curves to identify patients who were at a higher risk for retinopathy, and results showed duration of diabetes, systolic blood pressure, glycosylated haemoglobin, albuminuria, gender and diabetes therapy were significantly associated with the occurrence of retinopathy. As in depth study [12] was made to evaluate the efficiency of three plant components viz, cinnamaldehyde, cinnamic acid and cinnamyl alcohol in inhibiting Aldose reductase

(AR), an enzyme associated with retinopathy of both type 1 and type 2 diabetic patients. A product [13] made from whole leaf concentrate of Stevia, found to reduce hyper glycaemia in type 2 diabetic women. In (14), it was suggested that increased awareness and treatment of diabetes should begin with prevention. According to (15) data mining applications can be developed to evaluate the effectiveness of medical treatments.

II. EXPERIMENTAL METHODOLOGY

Data mining technology is useful for extracting non trivial information from medical databases. It is the intelligent computational analysis of large sets of data by using a combination of machine learning, statistical analysis and database technology, with the objective to discover patterns and rules useful for guiding decisions about future activities [16], [17]. Data mining technique was used to predict the chances of diabetic retinopathy. Under the data exploration mode, almost all attribute selection modules applicable for the data to collect optimal subset of attributes were explored.

There are various data mining techniques available with their suitability dependent on the domain application. Data mining application in health can have tremendous potential and usefulness. It automates the process of finding predictive information in large databases. Data mining classification technology consists of two models such as classification model and evaluation model. The classification model makes use of training data set in order to build classification predictive model. Testing data set is used for testing the classification efficiency.

Data mining tools predict future trends and behaviours help organizations to make proactive knowledge-driven decisions. Rapid Miner was chosen as the data mining tool due to its learning operators and operator framework, which allows forming nearly arbitrary processes. The application of Rapid miner software is being used for business and industries besides research, education, training, rapid prototyping. It helps in coordinated activities in machine learning, data mining, text mining, predictive analytics and business analytics. Thus it supports all stages of the data mining process to get valid and optimized results with clear visualization. [18], [19]. The proposed architecture is shown in Figure 1.

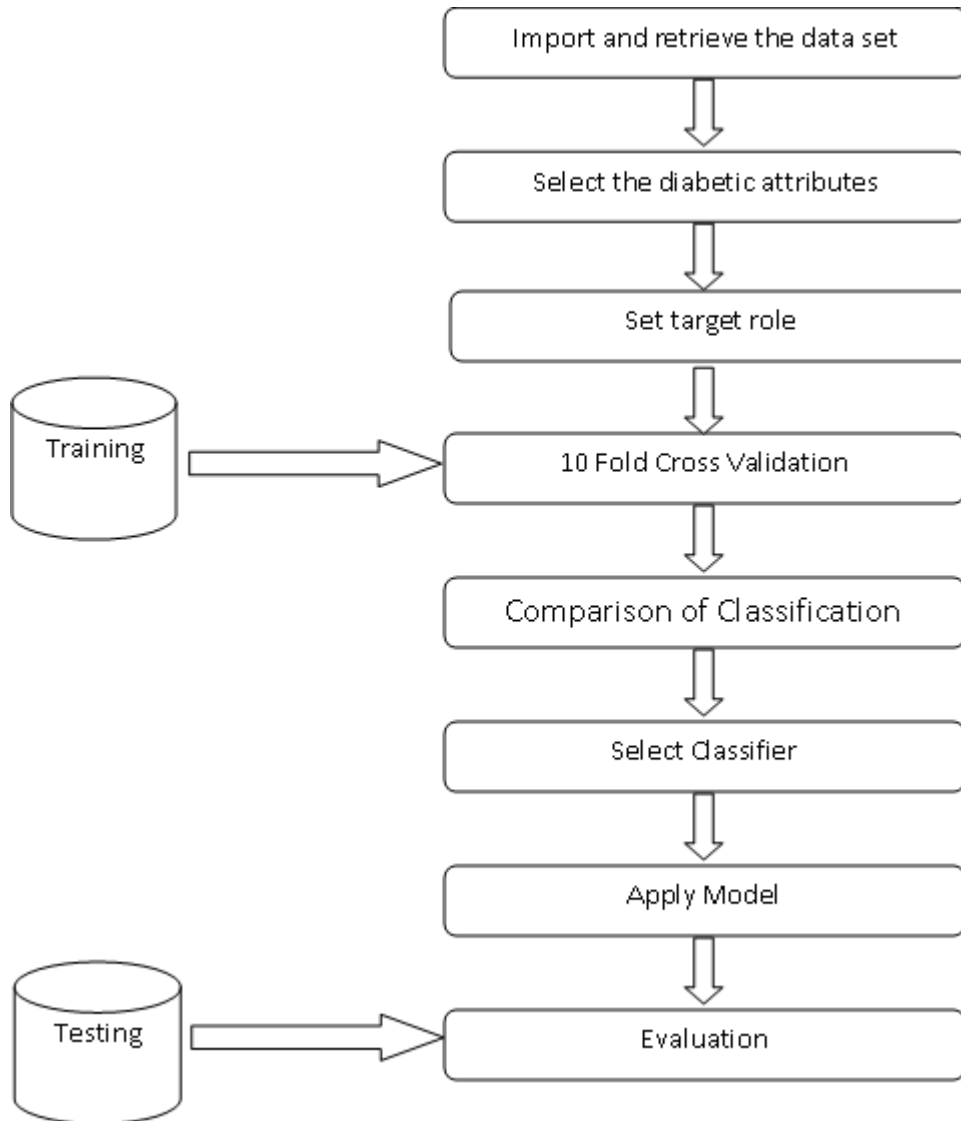


Figure 1 Proposed Architecture

Though there is availability of Cleveland Clinic Foundation Heart Disease dataset, for the sake of determining the accuracy rate in Indian region, we have collected 500 clinical records from Dr. Seshiah Diabetes Centre, Chennai, Tamil Nadu. The clinical data set specification provides concise, unambiguous definition for items related to diabetes.

Typically, cross-validation is used to generate a set of training, validation folds, and we compared the expected error on the validation folds after training on the training folds. Cross validation works were carried out by using part of the data to train the model, and the rest of the dataset to test the accuracy of the trained model. In this case, we have divided the dataset into 10 parts with training and testing data for each part.

The attributes data view of each record is shown in Table 1.

TABLE 1 Diabetic Attributes used in our Experimentation

Attribute Role	Attribute Name	Attribute Type	Description
Regular	Sex	Binomial	Sex of the patient. Values: Male, Female
Regular	Age	Integer	Age of the patient
Regular	Family Heredity /	Polynomial	Indicates whether the patient's parents were affected by diabetes. Values: Father, Mother, Both
Regular	Weight	Numeric	Weight of the patient
Regular	BP	Polynomial	Blood pressure of the patient
Regular	Fasting	Integer	Fasting Blood Sugar
Regular	PP	Integer	Post prondial Blood Glucose
Regular	A1C	Numeric	Glycosylated Hemoglobin Test
Regular	LDL	Integer	Low Density Lipoprotein
Regular	VLDL Cholesterol HDL	Integer Numeric	Very Low Density Lipoprotein High Density Lipoprotein
Label	Vulnerability	Polynomial	Indicates the Vulnerability of the patients to Retinopathy. Values : High, Medium, Low

III. MACHINE LEARNING METHODS

Most of these systems have successfully employed Machine learning methods such as Naïve Bayes and Decision tree for the classification purpose.

A. *Naives Bayes Method*

Naïve Bayes Classifier is a term dealing with simple probabilistic classifier based on applying Bayes Theorem with strong independence assumptions. It assumes that the presence or absence of particular feature of a class is unrelated to the presence or absence of any other feature [20].

Naives Bayes method is based on probabilities which are conditional and given the probability of another event that has already occurred, the probability of an event occurring is found using Bayes theorem. [21] If 'A' is referred as prior event and 'B' as dependent event, Bayes' theorem can be given as

$$\text{Prob}(B|\text{given}A)=\text{Prob}(A\text{and}B)/\text{Prob}(A)$$

The Naïve bayes performance screen is shown in Figure 2.

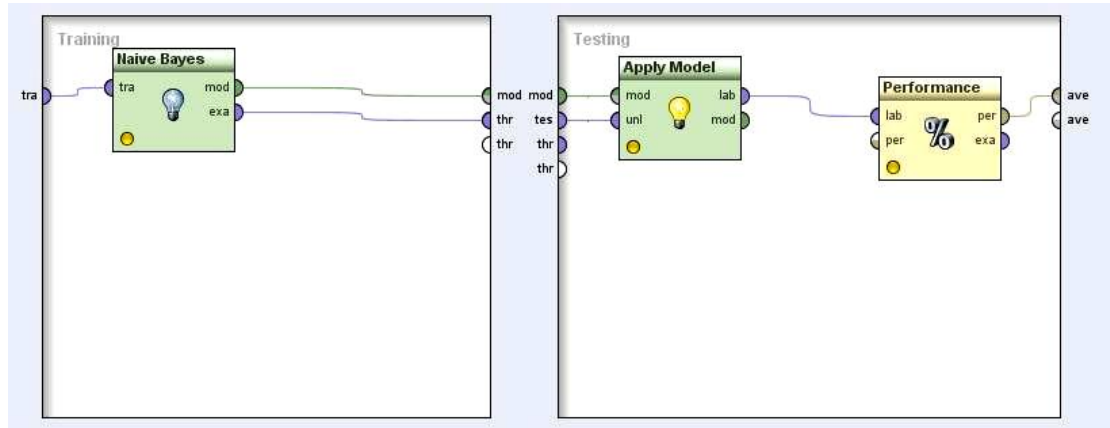


Figure 2 Naïve Bayes Performance Screen

B. Decision Tree

Decision tree is a popular classifier and prediction method for handling high dimensional data and it looks like a tree structure [22]. It builds classification or regression models in the form of tree structure. It breaks down smaller and smaller subsets while at the same time an association decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

The Decision tree performance screen is shown in Figure 3.

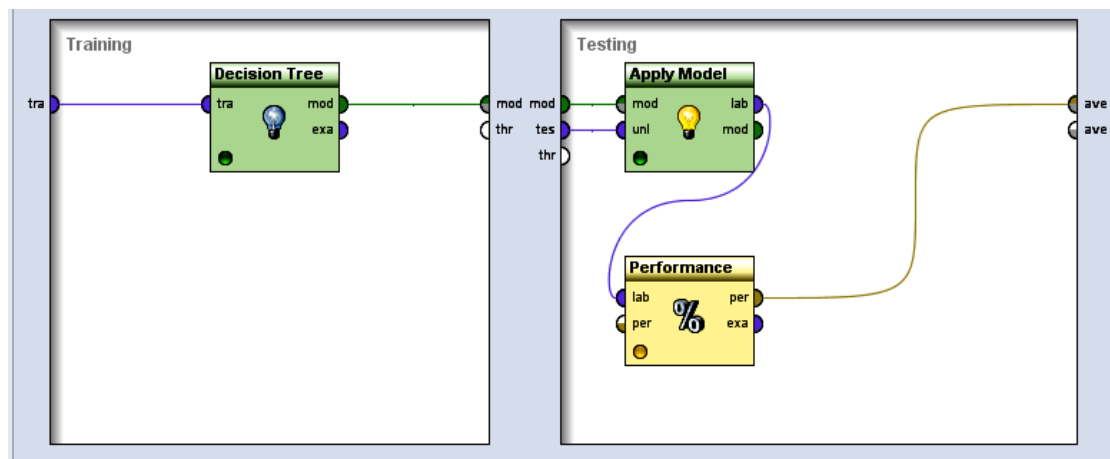


Figure 3 Decision Tree Performance Screen

IV. RESULTS AND DISCUSSION

The basic approach taken with this Rapid miner tool is to prepare a process model which uses 10-Cross validation along with the machine learning algorithm to increase the accuracy of the model. Further out of 10 subsets, which are classifier trained, one

subset is tested. By this the whole training set is provided once in each instance and cross validation arrived. It shows the correct classification data in percentage.

In Naïve Bayes classification technique, we are assumed that the probability distribution for an attribute follows a normal or Gaussian distribution.

We have tried to take age, sex, smoking, alcohol, cholesterol HDL as the prime attribute to evaluate Naive Bayes with the plot.

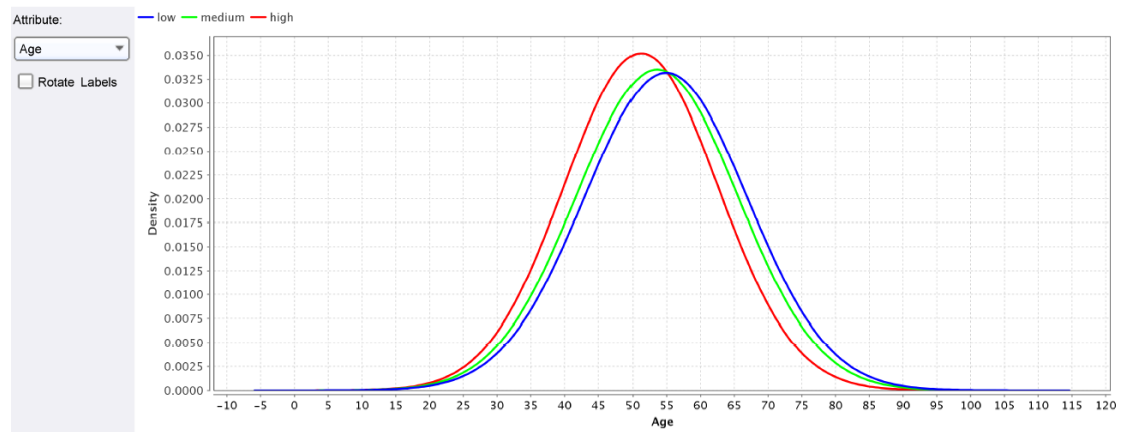


Figure 4 Bayes distribution plot by age attribute

In Figure 4, X axis denotes age and Y axis denotes the density. At the age of 52, the risk is high, at the age of 54, the risk is medium and at the age of 55 the risk is low.

Similarly, its distribution table for above five attributes are also calculated and table for age attribute is shown in Table 2.

TABLE 2 Naïve Bayes distribution table for age attribute

Attribute	Parameter	Low	Medium	High
Age	Mean	54.92	53.62	51.24
Age	Standard deviation	12.03	11.90	11.33

The performance of the proposed model is evaluated by accurately calculating the correctly predicted True Positive and True Negative classifications, arrived from out of proportion of instances.

The Sensitivity, Specificity and Accuracy values are calculated using the formulas.

$$\text{Sensitivity} = \text{True Positive} / (\text{True Positive} + \text{False Negative}).$$

$$\text{Specificity} = \text{True Negative} / (\text{True Negative} + \text{False Positive}).$$

Accuracy = (True Positive + True Negative) / True Positive + False Positive + True Negative + False Negative.

The matrix indicating the accuracy of the Naïve bayes classifier for the given data sets is shown in Table 3.

TABLE 3 Bayes Distribution accuracy: 81.37%

	True low	True medium	True high	Class precision
pred. low	620	52	20	88.38%
pred. medium	37	95	22	60.12%
pred. high	10	23	84	70.49%
class recall	92.66%	52.97%	64.66%	

The decision tree using various split methods such as Gain ratio, Information gain and Gini index has been tried as shown in Table 4 which gives different levels of accuracy.

TABLE 4 Accuracy by split methods using decision tree

Split method criteria	Accuracy in percentage	Classification error in percentage
Gain ratio	88.49	11.71
Information gain	89.99	9.21
Gini Index	77.69	26.31

For classification problems, it is natural to measure a classifier’s performance in terms of the error rate. The classifier predicts the class of each instance. If it is correct, that is counted as success; if not, it is an error. The error rate is just the proportion of errors made over a whole set of instances, and it measures the overall performance of the classifier. The part of decision tree diagram is shown in Figure 5 and the text view is shown in Figure 6.

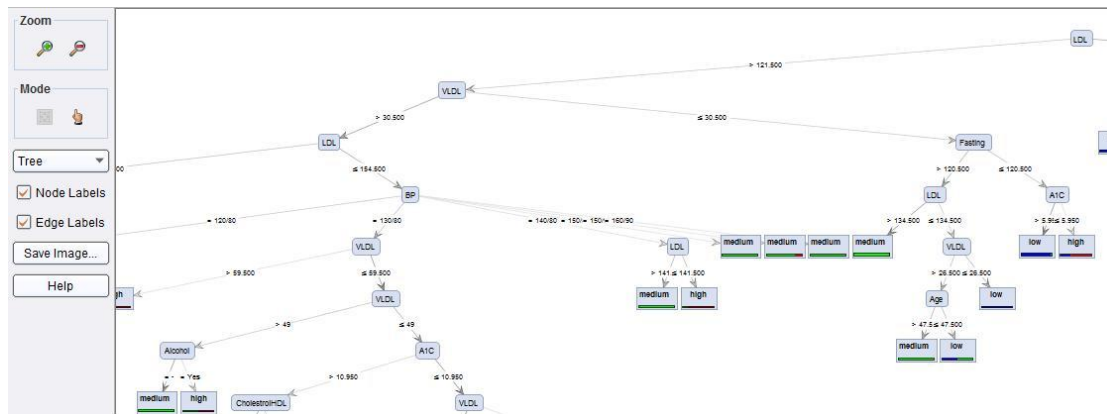


Figure 5 Decision Tree diagram

Tree

```

LDL > 121.500
| VLDL > 30.500
| | LDL > 154.500
| | | VLDL > 43.500: high {low=1, medium=0, high=62}
| | | VLDL 43.500
| | | | Age > 43: high {low=0, medium=1, high=23}
| | | | Age 43
| | | | | Age > 39.500: medium {low=0, medium=4, high=0}
| | | | | Age 39.500: high {low=0, medium=0, high=2}
| | LDL 154.500
| | | BP = 120/80
| | | | VLDL > 49.500
| | | | | Fasting > 133.500: high {low=0, medium=1,
high=6}
| | | | | Fasting 133.500: medium {low=0, medium=8,
high=0}
| | | | VLDL 49.500: medium {low=0, medium=26, high=0}
| | | BP = 130/80
| | | | VLDL > 59.500: high {low=0, medium=0, high=6}
| | | | VLDL 59.500
    
```

Figure 6 Decision Tree text view

With the use of the information gain as split parameter in decision trees, the results are exhibited by average precision, recall and accuracy of this technique was found to be 89.99 % shown in Table 5.

TABLE 5 Performance of Decision tree with an accuracy of 89.99% using information gain as split parameter

	True low	True medium	True high	Class precision
pred. low	652	21	10	94.81%
pred. medium	14	140	21	79.57%
pred. high	4	10	101	85.00%
class recall	96.48%	80.00%	76.69%	

As in table 5, a total of 500 records with medical attributes, having the results of two models, Decision tree appear to be most effective as it has the highest percentage of correct predictions (89.99%) for patients with retinopathy, followed by naive Bayes.

TABLE 6 Accuracy of various Classification techniques

Technique	Accuracy
Naïve Bayes	81.37%
Decision Tree	89.99%

As shown in Table 6, for the results of two models, decision tree appears to be most effective as it has the highest percentage of correct predictions (89.99%) for patients with diabetic retinopathy, followed by naïve Bayes.

V. CONCLUSIONS

Application of Data mining in analysing the medical data is a good method for considering the existing relationships between variables. We have applied naïve Bayes and decision tree classification algorithms on Diabetes dataset and the performance of those algorithms have been analysed and found that decision tree is to be more accurate. Thus this work presents a successful Diabetic Retinopathy predicting method which helps to predict the disease in early stage that can eventually reduce the manual work. Performances of our method were also measured by Specificity 95% and Sensitivity 96.65%. The proposed approach has shown that mining helps to retrieve useful correlation even from attributes which are not direct indicators of the class we are trying to predict. Our future work for this paper is to implement other algorithms like neural network and clustering with use of medical datasets in Weka tool. Moreover these data analysis results can be used for further research in enhancing the accuracy of the prediction system in future.

ACKNOWLEDGMENT

We are grateful to Dr.V.Shesiah, Chairman and Managing director of Dr.V.Shesiah Diabetes Centre, Chennai for providing an access to medical diabetic data and for his involvement in this domain.

REFERENCES

- [1] Thuraisingham, B.(2000) "A Primer for Understanding and Applying Data Mining", IT Professional, pp: 28-31.
- [2] World Health Organization. Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: Available: <http://www.who.int/diabetes/en>
- [3] Akara Sopharak, Bunyarit Uyyanonvara, Sarah barman (2011) 'Automatic Microaneurysm Detection from Non-dilated Diabetic Retinopathy Retinal Images Using Mathematical Morphology Methods', IAENG International Journal of Computer Science, IJCS_38_3_15

- [4] Lily Tapak & Hossein Mahjub, 'Real Data Comparison of Data Mining Methods in Prediction of Diabetes in Iran', Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3810525/>
- [5] Neera Singh, Ramesh Chandra Tripathi (2010), 'Automated Early Detection of Diabetic Retinopathy Using Image Analysis Techniques', *International Journal of Computer Applications* (0975-8887), volume 8, No.2.
- [6] Jianchao Han, Juan Rodriguze & Mohsen Beheshti (2008), 'Diabetes Data Analysis and Prediction Model Discovery Using Rapid Miner', *In Proceedings of the 2nd International Conference on Future Generation Communication and Networking*, vol.3, pp. 96-99
- [7] Sellappan Palaniappan & Rafiah Awang (2008), 'Intelligent Heart disease Prediction System using Data Mining Techniques', *International Journal of Computer Science and Network Security*, vol.8, no. 8, pp. 343-350
- [8] Sarojini Balakrishnan, Ramaraj Narayanaswamy. "An Empirical Study on the Performance of Integrated Hybrid Prediction Model on the Medical Datasets". *International Journal of Computer Applications* 29(5):1-6, September 2011. Published by Foundation of Computer Science, New York, USA. ISBN: 978-93-80864-76-4.
- [9] Lily Tapak, Hossein Mahjub, Omid Hamidi, & Jalal Poorolajal (2013), 'Real-Data Comparison of Data Mining Methods in Prediction of Diabetes in Iran', *Health Inform Res.* Vol 19, pp.177-185
- [10] Vallabha, D., Dorairaj, R., Namuduri K. and Thompson., H, "Automated Detection and Classification of Vascular Abnormalities in Diabetic Retinopathy", *IEEE*, 2004.
- [11] Cho, H. Y., Lee, D. H., Chung, S. E., & Kang, S. W. (2010). Diabetic Retinopathy and Peripapillary Retinal Thickness. *Korean J Ophthalmol*, 24(1), 16-22
- [12] Sivakumari K, Flora Mary Cyril Rathinabai A, Kaleena P.K, Jayaprakash P & Srikanth R (2010), 'Molecular docking study of bark-derived components of *Cinnamomum cassia* on aldose reductase', *Indian Journal of Science and Technology*, Vol.3, .No.8, pp. 1081-1088.
- [13] Parimalavalli R & Radhaisri S (2011), 'Glycaemic index of stevia product and its efficacy on blood glucose level in type 2 diabetes', , *Indian Journal of Science and Technology*, Vol.4, .No.3, pp. 318-321.
- [14] Akkarapol Sa-ngasoongsong & Jongsawas Chongwatpol (2012), 'An Analysis of Diabetes Risk Factors Using Data Mining Approach', pp. 1-11
- [15] Salim Diwani, —Overview Applications of Data Mining In Health Care: The Case Study of Arusha Region|| , *International Journal of Computational Engineering Research*, Vol 03, Issue, 8, August 2013
- [16] Han KamberJ, M. 2006. *Data Mining: Concepts and Techniques*, 2nd ed. San Francisco: Morgan Kaufman.
- [17] Fayyad U, .Piatetsky-Shapiro, and.Smyth P (1996), "From Data Mining to Knowledge Discovery in Databases", *AI Magazine*, Vol.17, pp.37-54.
- [18] Statistica tool, <<http://en.wikipedia.org/wiki/STATISTICA>>

- [19] Rapid Miner, '*Machine learning software getting started*' <<http://rapidminer.com/learning/getting-started/>>
- [20] Naïve bayes classifier based on applying bayes theorem: [http://en.wikipedia.org/wiki/Naive bayes classifier](http://en.wikipedia.org/wiki/Naive_bayes_classifier)
- [21] Naïve Bayes Classifier, '*Bayes theorem*' <[http://en.wikipedia.org/wiki/Naive bayes classifier](http://en.wikipedia.org/wiki/Naive_bayes_classifier)>
- [22] Sudha A, Gayathri P, N.Jaisankar, "Effective Analysis and Predictive Model of Stroke Disease using Classification Methods", International Journal of Computer Applications (IJCA) Volume 43-No.14, April 2012, 0975-8887.