A Novel Technique for Prediction of Diabetic Patients Data Using Naives Bayes Classification in Orange Tool

S.Nivetha
Assistant Professor, Department of Computer Science,
Kamban Arts and Science College,
Pollachi, Tamil Nadu, India.
Email: nive0501@gmail.com
Dr.A.Geetha
Assistant Professor, PG & Research Department of Computer Science,
Chikkanna Govt. Arts College,
Tirupur, Tamil Nadu, India.
Email: gee_sam@yahoo.com

Abstract - Data mining is a process of extracting information from a dataset and transform it into understandable structure to discover patterns in large data sets. Data mining for healthcare is useful in evaluating the effectiveness of clinical treatments to its roots in databases records system getting to know and facts visualization. Diabetic ailment refers back to the heart disorder that develops in persons with diabetes. The term diabetes is a continual ailment that occurs both when the pancreas does now not produce sufficient insulin. The blood vessels despite the fact that many data mining type techniques exist for the prediction of heart disorder there is inadequate records for the prediction of heart illnesses in a diabetic character. A number of experiments had been conducted the use of orange tools for contrast of the performance of predictive facts mining techniques on the diabetic dataset with attributes. The naive bayes classifier method has been carried out in orange tool prediction model using minimal training set to diagnose vulnerability of diabetic sufferers. All the above experiments find the probabilities of risk in diabetic patients for coronary heart sickness. In this test a comparative examine has been performed at the classifiers which result in the chance of diabetic patients getting heart disease from a system. The performances additionally had been in comparison the use of accuracy and additionally in terms of precision and exhibited a great overall performance.

Keywords- Data Mining, Diabetic Data, RF, NB, ORANGE.

1. INTRODUCTION

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge.

Flat files: Flat files are actually the most common data source for data mining algorithms, especially at the research level. Flat files are simple data files in text or binary format with a structure known by the data mining algorithm to be applied.

Relational Databases: It consists of a set of tables containing either values of entity attributes or values of attributes from entity relationships. Tables have columns and rows, where columns represent attributes and rows represent tuples.

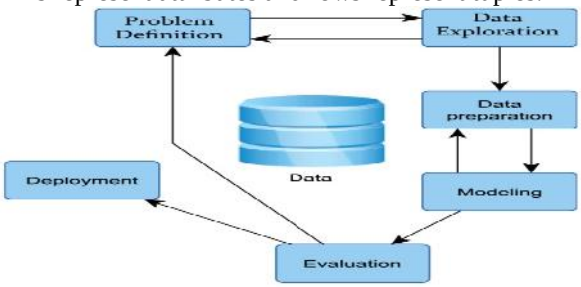

Fig 1.1 Steps in Data mining

Data Warehouses: A data warehouse as a store house is a repository of data collected from multiple data sources and is intended to be used as a whole under the same unified schema.

Transaction Databases: A transaction database is a set of records representing transactions, each with a time stamp, an identifier and a set of items.

Multimedia Databases: Multimedia databases include video, images, audio and text media. They can be stored on extended object-relational. Multimedia is characterized by its high dimensionality, which makes data mining even more challenging.

Spatial Databases: Spatial databases are databases that, in addition to usual data, store geographical information like maps and global or regional positioning.

World Wide Web: The World Wide Web is the most heterogeneous and dynamic repository available. A very large number of authors and publishers are continuously contributing to its growth and metamorphosis, and a massive number of users are accessing its resources daily.

Time-Series Databases: Time-series databases contain time related data such stock market data or logged activities. These databases usually have a continuous flow of new data coming in, which sometimes causes the need for a challenging real time analysis.

2. RELATED WORKS

Anuja Kumari et al. [1] described the Support vector machine, a supervised machine learning method as the classifier for diagnosis of diabetes using Pima Indian diabetic database in Classification of Diabetes Disease Using Support

Vector Machine. They have used the basic concepts of SVM and kernel function selection and experiments have been conducted on Matlab.

Asha Gowda Karegowda et al. [2] describes diabetes can occur in anyone. However, people who have close relatives with the disease are somewhat more likely to develop it. Other risk factors include obesity, high cholesterol, high blood pressure and physical inactivity. The risk of developing diabetes also increases, as people grow older. People who are over 40 and overweight are more likely to develop diabetes, although the incidence of type-2 diabetes in adolescents is growing.

Jayshri Sonawane et al. [3] presented the heart is the organ that pumps blood, with its life giving oxygen and nutrients, to all tissues of the body. If the pumping action of the heart becomes inefficient, vital organs like the brain and kidneys suffer and if the heart stops working altogether, death occurs within minutes. The term heart disease applies to a number of illnesses that affect the circulatory system, which consists of heart and blood vessels.

Jianchao Han et al. [4] analyzed 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.

Karthikeyani et al. [5] presented the classification of supervised data mining algorithms based on diabetes disease dataset in Comparative of Data mining classification algorithm in Diabetes disease Prediction. Different classification algorithms like C4.5 decision tree, Classification and Regression Trees, Support Vector Machine, K-Nearest Neighbour and Prototype Neural Network classification have been used to analyze the Pima Indian Diabetes dataset with 9 attributes and 768 instances.

Sarojini Balakrishnan et al. [6] proposed a system to improve the diagnostic accuracy of diabetic disease by selecting informative features of Pima Indians Diabetes dataset in Empirical Study on the Performance of Integrated Hybrid Prediction Model on the Medical Datasets. They propose a hybrid prediction model that 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.

Selvakuberan et al. [7] presented the diabetes is one of the major causes of premature illness and death worldwide. In developing countries, less than half of people with diabetes are diagnosed. There is no time for diagnoses and adequate treatment, complications and morbidity from diabetes rise exponentially.

Vahid Rafe et al. [8] developed the medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis. Data mining technology provides a user oriented approach to novel and hidden patterns in the data. Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently.

Vijayarani et al. [9] discussed the heart disease plays an important role in data mining due to occurrence of death in heart diseases. Medical diagnosis plays a vital role and it is yet a complicated task that needs to be executed efficiently and accurately. To reduce cost for achieving clinical tests an appropriate computer based information and decision support should be provided.

Vijaya Lakshmi et al. [10] discussed Gestational Diabetes Mellitus is defined as any abnormal carbohydrate that begins or is first recognized during pregnancy. It does not exclude the possibility that unidentified glucose intolerance have preceded the pregnant state.

.

3. DATA MINING TOOL

3.1 ORANGE

Orange is a component-based data mining and machine learning software suite, featuring a visual programming front-end for explorative data analysis and visualization and Python bindings and libraries for scripting. It includes a set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation and exploration techniques. Its graphical user interface builds upon the cross-platform framework

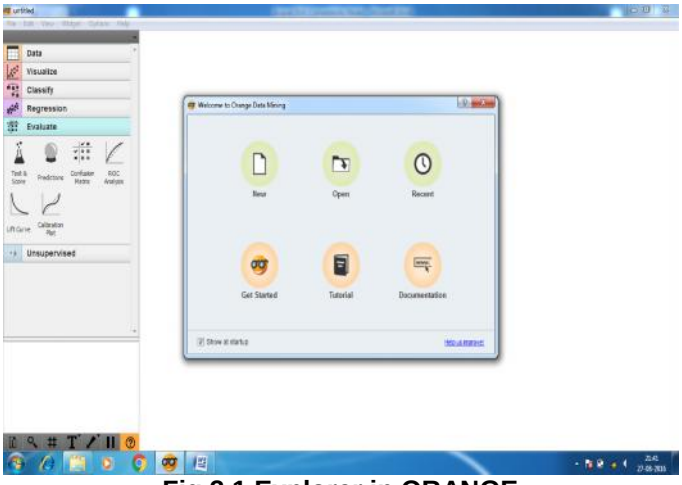

Fig 3.1 Explorer in ORANGE

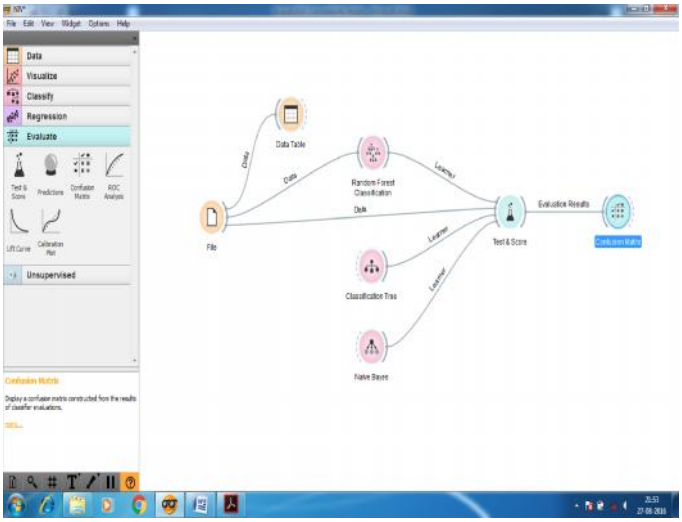

Fig 3.2 Diabetic Attributes Selected in ORANGE

4.1 EXISTING METHODOLOGY

4.1.1 J48 Pruned Tree

J48 is a module for generating a pruned or unpruned C4.5 decision tree. When we applied J48 onto refreshed data, got the results shown as below on Figure.

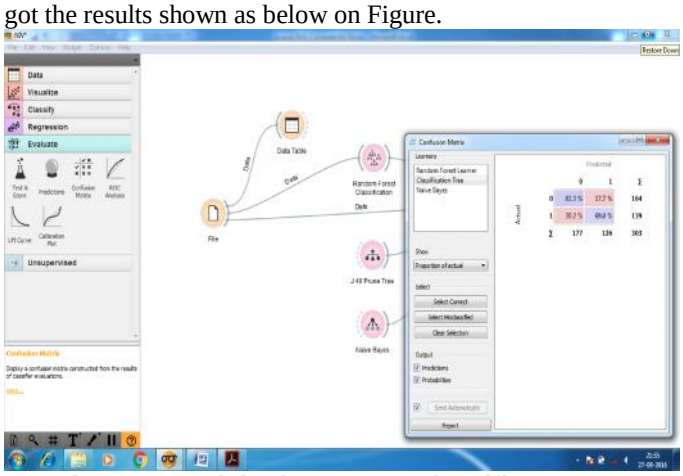

Fig 4.1 J48 Diabetic Dataset Classifier Output Predicted

4.1.2 Random Forest

Random forest is an algorithm that consists of many decision trees. The model uses a bagging approach and the random selection of features to build a collection of decision trees with controlled variance. The instances class is to the class with the highest number of votes, the class that occurs the most within the leaf in which the instance is placed. By using trees that classify the instances with low error the error rate of the forest decreases. The correlation and strength of the forest increases with the number m of variables selected. A smaller m returns a smaller correlation and strength. To improve the prediction's accuracy, a bootstrap method is used to create different trees.

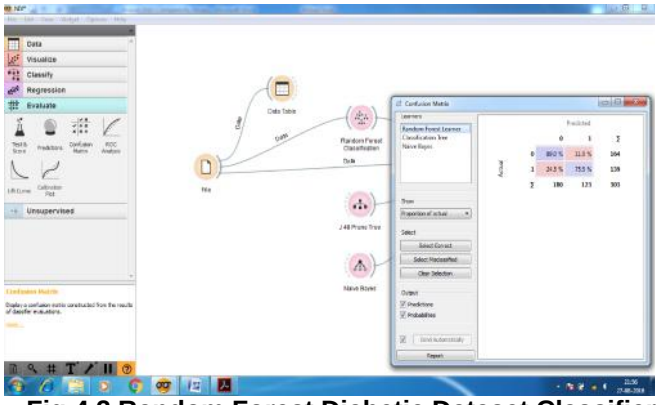

Fig 4.2 Random Forest Diabetic Dataset Classifier Predicted Output

4.2 PROPOSED METHODOLOGY

Classification is the process of finding a model that describes and distinguishes data classes or concepts, for the able purpose of being to use the model to predict the class of objects whose class label is unknown. It is a technique which is used to predict group membership for data instances. Classification is a two step process, first, it builds classification model using training data. Every object of the dataset must be pre-classified and the second the model generated in the preceding step is tested by assigning class labels to data objects in a test dataset. Here using diabetes dataset now a days the percentage of diabetes patient is growing very fast. India accounts for the largest number of people suffering from diabetes in the world. The diabetes in country's population is likely to be affected from the disease. It is estimated that every five person with diabetes will be an Indian. It means that India has highest number of diabetes in any one of the country in the world. The attributes predict whether a person having diabetes or not.

4.2.1 Naive Bayes Approach

Naive Bayes classifier as a term dealing with a simple probabilistic classifier based on application of Bayes theorem with strong independence assumptions. Since independent variables are assumed, only the variances of the variables for each class need to be determined. It can be used for both binary and multi class classification problems. Naive Bayes data mining classifier technique has been applied which produces an optimal prediction model using minimum training set to predict the chances of diabetic patient getting heart disease. The diagnosis of diseases plays vital role in medical field. Using diabetic's diagnosis, the proposed system predicts attributes such as age, sex, blood pressure and blood sugar and the chances of a diabetic patient getting a heart disease. It should be noted that the attributes used in our proposed method are those used for diagnosis of diabetes and are not direct indicators of heart disease. Each algorithm requires submission of data in a specified format.

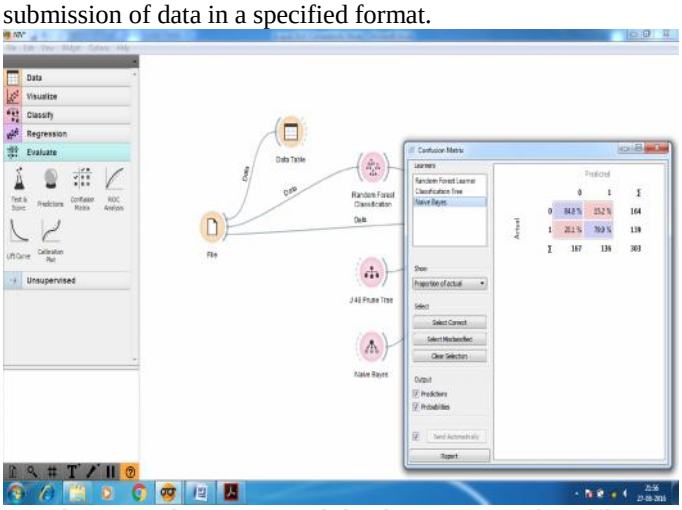

Fig 4.6 Naive Bayes Diabetic Dataset Classifier Predicted Output

5. EXPERIMENTATION & RESULTS

5.1 Performance evaluation

To measure the performance sensitivity, accuracy and specificity are used. TP is true positive, FP is false positive, TN is true negative and FN is false negative. TPR is true positive rate, which is equivalent to Recall.

$$\text{Accuracy} = \frac{TP+TN}{(TP+TN+FP+FN)} \quad \ldots\ldots 1$$

Methods / Parameters	J48 tree	Random Forest	Naive Bayes Classifier
Number of Instances	768	768	768
Accuracy	82.3%	83.0%	84.8%

Table.5.1 Comparison Results

From the above table 5.1 shows the performance of naive bayes classifier. The fig.5.1 shows comparison graphical representation of methods. The method can over perform the traditional method with classify recall rate of 0.842.

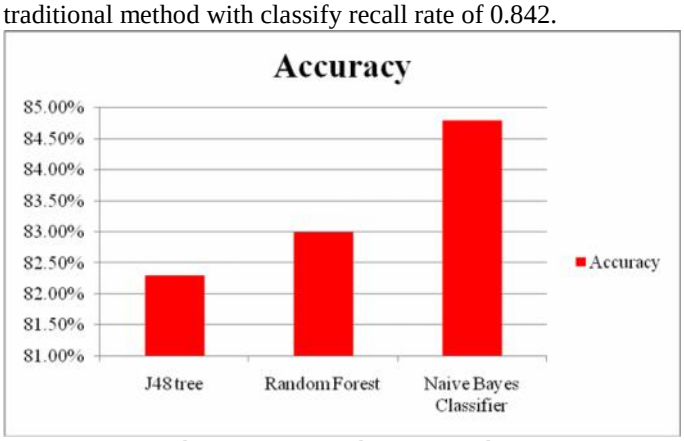

Fig.5.1 Comparison Results

6. CONCLUSION

Data mining for healthcare is useful in evaluating the effectiveness of medical treatments and ensures detection of fraud and abuse. The data mining techniques give the necessary standard in prediction. The performance in prediction depends on the various attributes which are helpful in predicting disease efficiently and patients receive better and more affordable healthcare services. The naive Bayes data mining classifier technique has been applied which produces an optimal prediction model using minimum training set to predict the chances of diabetic patient. Orange tool is considered being a successful tool for classification purpose and evidence is the proposed system is quite good, since it has proved and shown good accuracy on the prediction of diabetic. To determine the most accurate technique to predict the risk in diabetic patients. The diabetic patients based on their predictive accuracy. In overall accuracy, in terms of precision and recall exhibited a very consistent performance.

REFERENCES

1. Anuja Kumari, V & Chitra, R 2013, 'Classification of Diabetes Disease Using Support Vector Machine', International Journal of Engineering Research and Applications, vol. 3, no. 2, pp. 1797-1801.
2. Asha Gowda Karegowda, Manjunath AS and Jayaram MA 2011, 'Application of Genetic Algorithm Optimized Neural Network Connection Weights for Medical Diagnosis of Pima Indians Diabetes', International Journal on Soft Computing, vol. 2, no. 2, pp.15-23.
3. Jayshri Sonawane, S, Dharmaraj Patil, R & Vishal Thakare, S 2013, 'Survey on Decision Support System For Heart Disease, International Journal of Advancements in Technology, vol.4, no.1, pp. 89-96.
4. 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.
5. Karthikeyani, V & Parvin Begum 2012, 'Comparative of Data mining classification algorithm in Diabetes disease Prediction', International Journal of Computer Applications, vol. 60, no. 12, pp. 26-31.
6. Sarojini Balakrishnan, Ramaraj Narayanaswamy & Ilango Paramasivam 2011,'An Empirical Study on the Performance of Integrated Hybrid Prediction Model on the Medical Datasets', International Journal of Computer Applications, vol.29, no.5, pp. 1-6.
7. Selvakuberan, K, Kayathiri, D, Harini, B & Indra Devi, M 2011, 'An Efficient Feature Selection Method for Classification in Health care Systems using Machine Learning Techniques', In Proceedings of the 3rd International Conference on Electronics Computer Technology, Kanyakumari, India vol. 4, pp. 223-226.
8. Vahid Rafe & Roghayeh Hashemi Farhoud 2013, 'A Survey on Data Mining Approaches in Medicine', International Research Journal of Applied and Basic Sciences, vol.4, no.1, pp. 196-202.
9. Vijayarani, S & Sudha, S 2013, ' An Effective Classification Rule Technique for Heart Disease Prediction',International Journal of Engineering Associates, vol. 1, no.4, pp. 81-85.
10. A.Geetha,G.M.Nasira, "Rainfall Prediction using Logistic Regression Technique",CiiT International Journal of Artficial Intelligence Systems and Machine Learning,Vol.6,No.7,pp.246-250,2014.
11. Sripriya, A.Geetha, "Cyclone Storm Prediction using KNN Algorithm", Indian Journal of Engineering, Discovery Publication,12(30), pp.350-354,2015.