An Empirical Comparison by Data Mining Classification Techniques for Diabetes Data Set

Nilesh Jagdish Vispute
M.Tech Scholar, Department of CSE
Sri Satya Sai College of Engineering
Bhopal M.P., India

Dinesh Kumar Sahu
Ph.D Scholar, Department of Computer Science
Barkatullah University
Bhopal M.P., India

Anil Rajput
Professor, Department of Mathematics & Comp. Sc.
CSA, Govt. P.G. College
Sehore M.P., India

ABSTRACT
Data mining is a process of extracting information from a dataset and transform it into understandable structure for further use, also it discovers patterns in large data sets . Data mining has number of important techniques such as preprocessing, classification. Classification is one such technique which is based on supervised learning.. diabetic is a life threatening disease prevalent in several developed as well as developing countries like India. the data classification is diabetic patients data set is developed by collecting data from hospital repository consists of 1865 instances with different attributes. The instances in the dataset are two categories of blood tests, urine tests. In this paper we discuss various algorithm approaches of data mining that have been utilized for diabetic disease prediction. Data mining is a well known technique used by health organizations for classification of diseases such as diabetes and cancer in bioinformatics research. In the proposed approach we have used WEKA with 10 cross validation to evaluate data and compare results. Weka has an extensive collection of different machine learning and data mining algorithms.

In this paper we have firstly classified the diabetic data set and then compared the different data mining techniques in weka through Explorer, knowledge flow and Experimenter interfaces. Furthermore in order to validate our approach we have used a diabetic dataset with 108 instances but weka used 99 rows and 18 attributes to determine the prediction of disease and their accuracy using classifications of different algorithms to find out the best performance. The main objective of this paper is to classify data and assist the users in extracting useful information from data and easily identify a suitable algorithm for accurate predictive model from it. From the findings of this paper it can be concluded that Naïve Bayes the best performance algorithms for classified accuracy because they achieved maximum accuracy= 76.3021% correctly classified instances, maximum ROC = 0.819 , had least mean absolute error and it took minimum time for building this model through Explorer and Knowledge flow results.

Keywords
Weka, Data mining, Classification, Diabetic Disease Prediction.

1. INTRODUCTION
The main focus of this paper is the classification of different types of datasets that can be performed to determine if a person is diabetic. The solution for this problem will also include the cost of the different types of datasets. For this reason, the goal of this paper is classifier in order to correctly classify the datasets, so that a doctor can safely and cost effectively select the best datasets for the diagnosis of the disease. The major motivation for this work is that diabetes affects a large number of the world population and it's a hard disease to diagnose. A diagnosis is a continuous process in which a doctor gathers information from a patient and other sources, like family and friends, and from physical datasets of the patient. The process of making a diagnosis begins with the identification of the patient's symptoms. The symptoms will be the basis of the hypothesis from which the doctor will start analyzing the patient. This is our main concern, to optimize the task of correctly selecting the set of medical tests that a patient must perform to have the best, the less expensive and time consuming diagnosis possible. A solution like this one, will not only assist doctors in making decisions, and make all this process more agile, it will also reduce health care costs and waiting times for the patients.

The major contributions of this paper are:

(1) To extract useful classified accuracy for prediction of Diabetes diseases.

(2) Comparison of different data mining algorithms on Diabetes dataset.

(3) Identify the best performance algorithm for prediction of diseases.

This paper will focus on the analysis of data from a data set called diabetes data set

2. RELATED WORK
The few medical data mining applications as compared to other domains. Reported their experience in trying to automatically acquire medical knowledge from clinical databases. They did some experiments on three medical databases and the rules induced are used to compare against a set of predefined clinical rules. Past research in dealing with this problem can be described with the following approaches:

(a) Discover all rules first and then allow the user to query and retrieve those he/she is interested in. The representative approach is that of templates . This approach lets the user to specify what rules he/she is interested as templates. The system then uses the templates to retrieve the rules that match the templates from the set of discovered rules.

(b) Use constraints to constrain the mining process to generate only relevant rules. Proposes an algorithm that can take item constraints specified by the user in the association rule mining processor that only those rules that satisfy the user specified

item constraints are generated. The study helps in predicting the state of diabetes i.e.,whether it is in an initial stage or in an advanced stage based on the characteristic results and also helps in estimating the maximum number of women suffering

from diabetes with specific characteristics. Thus patients can be given effective treatment by effectively diagnosing the characteristics. Our research work based on the concept from Data Mining is the knowledge of finding out of data andproducing it in a form that is easily understandable and comprehensible to humans in general. These further extended in this to make an easier use of the data's available with us in the field of Medicine.

The main use of this technique is the have a robust working model of this technology. The process of designing a model helps to identify the different blood groups with available Hospital Classification techniques for analysis of Blood group data sets. The ability to identify regular diabetic patients will enable to plan systematically for organizing in an effective manner. Development of data mining technologies to predict treatment errors in populations of patients represents a major advance in patient safety research.

3. OVERVIEW OF PROPOSED APPROACH

3.1 WEKA

In order to carry out experimentations and implementations Weka was used as the data mining tool. Weka (Waikato Environment for Knowledge Analysis) is a data mining tool written in java developed at Waikato. WEKA is a very good data mining tool for the users to classify the accuracy on the basis of datasets by applying different algorithmic approaches and compared in the field of bioinformatics. Explorer, Experimenter and Knowledge flow are the interface available in WEKA that has been used by us. In this paper we have used these data mining techniques to predict the survivability of Diabetes disease through classification of different algorithms accuracy .

Figure 1 visualizes the interface of WEKA Data mining tool. It has four applications:

(1)Explorer: The explorer interface has several panels like preprocess, classify, cluster, associate, select attribute and visualize. But in this interface our main focus is on the Classification Panel .

(2) Experimenter: This interface provides facility for systematic comparison of different algorithms on basis of given datasets. Each algorithm runs 10 times and then the accuracy reported.

(3)Knowledge Flow: It is an alternative to the explorer interface. The only difference between this and others is that here user selects Weka component from toolbar and connects them to make a layout for running the algorithms .

(4) Simple CLI: Simple CLI means command line interface. User performs operations through a command line interface by giving instructions to the operating system. This interface is less popular as compared to other three.

Figure 1 visualizes the interface of WEKA Data mining tool

3.2 Classification

In data mining tools classification deals with identifying the problem by observing characteristics of diseases amongst patients and diagnose or predict which algorithm shows best performance on the basis of WEKA's statistical output .

Three techniques have been adopted in this paper, the first technique uses explorer interface and depends on algorithms like Naïve Bayes, SMO, J48, REP Tree and RANDOM Tree, used in areas to represent, utilize and learn the statistical knowledge and significant results have been achieved.

The second technique uses Experimenter interface. This study allows one to design experiments for running algorithms such as Naïve Bayes, J48, REP Tree and RANDOM Tree on datasets. These algorithms can be run on experimenter and analyze the results. It configures the test option to use cross validation 10 folds. This interface provides provision for running all the algorithms together and thus a comparative result was obtained.

The third technique uses Knowledge Flow. In this study we classified the accuracy of different algorithms Naïve Bayes, SMO, J48, REP Tree and random Tree on different data sets and compared the results to know which algorithm shows best performance. In order to predict Diabetes Disease for survivability by user one can select this weka component from toolbar, place them in a layout like manner and connect its different components together in order to form a knowledge flow web for preprocessing and analyzing data.

All the algorithms used by us were applied to a Diabetes data set explained in detail in section 4. In order to obtain better accuracy 10 fold cross validation was performed. For each classification we selected training and testing sample randomly from the base set to train the model and then test it in order to estimate the classification and accuracy measure for each classifier. The thrust classifications and accuracy used by us are:

3.2.1 Correctly Classified Accuracy

It shows the accuracy percentage of test that is correctly classified.

3.2.2 Incorrectly Classified Accuracy

It shows the accuracy percentage of test that is incorrectly classified.

3.2.3 Mean Absolute Error
It shows the number of errors to analyze algorithm classification accuracy.

3.2.4 3.3.4. Time
It shows how much time is required to build model in order to predict disease.

3.2.5 ROC Area
Receiver Operating Characteristic19 represent test performance guide for classifications accuracy of diagnostic test

3.3 Data Mining Techniques
The data mining technique have been used by us to predict diabetes disease. Predictions have been done by us using weka data mining tool for classification and accuracy by applying different algorithms approaches. The interfaces of weka used in this paper are the following:

3.3.1 Explorer Interface
It first preprocesses the data and then filters the data. Users can then load the data file in CSV (Comma Separated Value) format and then analyze the classification accuracy result by selecting the following algorithms using 10 cross validation: Naïve Bayes, J48, SMO, REP Tree, and Random Tree.

Figure 2 shows the interface of explorer when The output obtained by scoring of NaïveBayes, J48, SMO, REPTree , Random Tree algorithm accuracy of is given on the basis of time, accuracy, error and ROC.

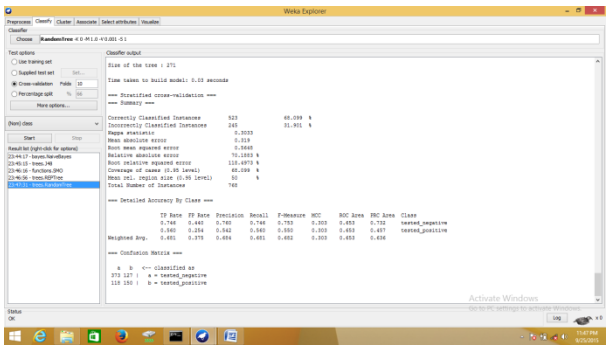

Fig. 2. Screenshot view of Explorer Interface Accuracy

3.3.2 Experimenter Interface
Experimenter Interface has been used in this paper to analyze data by experimenting through algorithms such as Naïve Bayes, J48, REP Tree and Random Tree to classify the data using train and test sets.In Figure 3 we run four different algorithms on Diabetes datasets and analyze algorithms accuracy.

(a) Naïve Bayes
It is one of the fastest algorithm works on probability of all attribute contained in dat sample individually and then classifies them accurately.

(b) J48 Tree
We used J48 tree to decide the target value based on various attribute of dataset to predict algorithms accuracy.

(c) REP Tree
We used Weka classifier tree algorithm analyze accuracy applied on Diabetes dataset.

(d) Random Tree
We used Random classifier tree algorithm to analyze classification based on our dataset. Figure 3 analyzes experiment test of all four algorithms, each algorithm is run 10 times and ccuracy is reported. "v" stand for best accuracy prediction and "*" stand for worse accuracy prediction. This means it predicts best and worse scoring accuracy amongst the four different algorithms listed below respectively:

- Naïve Bayes
- J48 Tree
- SMO
- REP Tree
- Random Tree

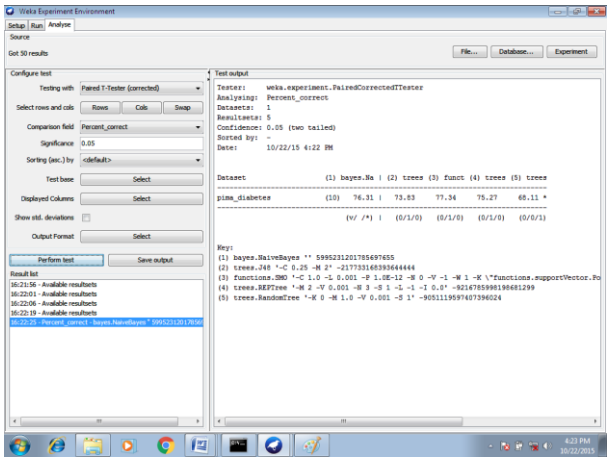

Fig. 3. Screenshot view of Experimenter Algorithm Accuracy

3.3.3 Knowledge Flow Interface
Knowledge Flow is an alternative to the explorer.[18]the user lays out the data by connecting them together in order to form a knowledge flow by selecting weka component from a tool bar as shown in Figure 4 .For the purpose of our experimentation we have connected together CSV loader, class assigner, Cross validation, and then an algorithm such as SMO, REP tree etc followed by Classifier Performance evaluator and finally we view the output using text viewer.

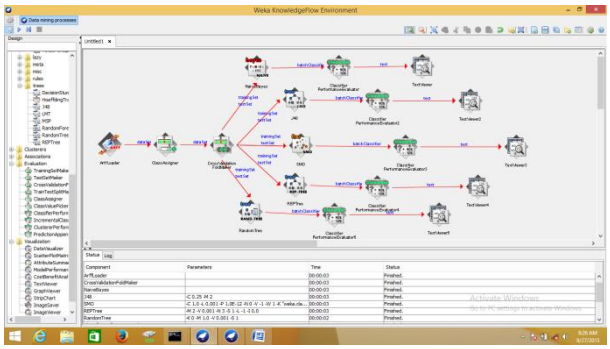

Fig. 4. Screenshot view of Knowledge Flow Interface

4. RESULTS/ DISCUSSION

Explorer, Experimenter and Knowledge flow are the data mining techniques that have been used by us using different algorithms Naïve Bayes, J48, SMO, RANDOM tree and REP tree. Through these techniques we trained out results on the basis of time taken to build model, correctly classified instances, error and ROC area. Algorithm scoring accuracy is shown in Table 1.Naïve Bayes 76.3021 % correctly instances accuracy with minimum Naïve Bayes Mean Absolute Error = 0.2841 having maximum Naïve Bayes ROC =0.819 time taken to build model=0.05 seconds. So from Explorer Interface data mining technique we can deduce that Naïve Bayes have maximum accuracy , least error and it takes less time to build model it and has maximum ROC.

In Table 2 Naïve Bayes classified 76.3021% correctly instances accuracy with minimum Naïve Bayes Mean Absolute Error = 0.2841, having maximum Naïve Bayes ROC =0.819 and time taken to build model=0 seconds. So from Knowledge flow Interface data mining technique result Naïve Bayes have maximum accuracy, least error, less time taken to build model and maximum ROC. Explorer and Knowledge flow achieved same scoring to classify accuracy but there is approx. change in ROC Value of Naïve Bayes as compared to other because Knowledge flow is an alternative method of Explorer.

In Table 3 Naïve Bayes and SMO scoring accuracy is high that is best prediction (V) as compared to REP Tree and Random Tree having low algorithm accuracy called worse prediction (*).

Table 1. .Explorer result

Algorithm	Time Taken to Build Model (seconds)	Correctly Classified Instances %Accuracy	Incorrectly Classified Instances %Accuracy	Mean Absolute Error	ROC Area
Naïve Bayes	0.05	76.3021% (586)	23.6979% (182)	0.2841	0.819
J48	0.09	73.8281% (567)	26.1719% (201)	0.3158	0.751
SMO	0.5	77.3438% (594)	22.6563% (174)	0.2266	0.720
REPTree	0.06	75.2604% (578)	24.7396% (190)	0.3272	0.766
Random Tree	0	68.099% (523)	31.901% (245)	0.319	0.653

Table 2. Knowledge Flow result

Algorithm	Time Taken to Build Model (seconds)	Correctly Classified Instances %Accuracy	Incorrectly Classified Instances %Accuracy	Mean Absolute Error	ROC Area
Naïve Bayes	0	76.3021% (586)	23.6979% (182)	0.2841	0.819
J48	0	73.8281% (567)	26.1719% (201)	0.3158	0.751
SMO	0	77.3438% (594)	22.6563% (174)	0.2266	0.720
REPTree	0	75.2604% (578)	24.7396% (190)	0.3272	0.766
Random Tree	0	68.099% (523)	31.901% (245)	0.319	0.653

Table 3 : Experimenter result

Algorithm	Best Accuracy Prediction(v)	Worse Accuracy Prediction(*)
Naïve Bayes	76.31	-
J48 Tree	-	73.83
SMO	77.34	-
REP Tree	-	75.27
Random Tree	-	68.11

Finally from these three data mining technique it is observed that Naïve Bayes the best classifier performance to predict

the survivability of diabetes disease prediction among patient using WEKA because it classifies more accurately, has maximum ROC Area, least mean absolute error and takes minimum time to build model . The Accuracy of test depends on dataset with and without disease. Accuracy measured by ROC area =0.819 shows a perfect and excellent test as Patient will get effective diagnosis timely and in an accurate manner.

5. CONCLUSION / FUTURE WORK

The discovery of knowledge from medical databases is important in order to make effective medical diagnosis. The aim of data mining is to extract knowledge from information stored in database and generate clear and understandable description of patterns. The main aim of this paper is to predict diabetes disease using WEKA data mining tool. It has four interfaces. Out of these four we have used three interfaces: Explorer, Experimenter and knowledge flow. Each interface has its own classifier algorithms. We have used five algorithms i.e. Naïve Bayes, J48, SMO, REP Tree and Random tree for our experimentation. Then these algorithms were implemented using WEKA data mining technique to analyze algorithm accuracy which was obtained after running these algorithms in the output window. After running these algorithms the outputs were compared on the basis of accuracy achieved. In Explorer and Knowledge flow there are several scoring algorithms for accuracy but for our experimentation we have used only five algorithms. The outputs obtained from both Explorer and Knowledge flow is approximately same because knowledge flow is an alternative method of Explorer. It is just a different way of carrying out experimentations. These algorithms compare classifier accuracy to each other on the basis of correctly classified instances, time taken to build model, mean absolute error and ROC Area.Through Explorer and Knowledge Flow and Experimenter technique compare these three technique it was inferred that Nave Bayes the best performance classifier algorithms as they achieved an accuracy of 76.3021 %, takes less time taken to build and shows maximum ROC area = 0.819, and had least absolute error. Maximum ROC Area means excellent predictions performance as compared to other algorithms.

The proposed approach is used with diabetes data set but The future work can be applied to blood groups to identify the relationship that exits between diabetic, diagnosing cancer patients based on blood cells orpredicting the cancer types on the blood groups, blood pressure, personality traits and medical diseases.

6. REFERENCES

[1] S , Liver Disease Prediction Using Bayesian Classification , Special Issues , 4th National Conference on Advance Computing , Application Technologies, May 2014

[2] SolankiA.V., Data Mining Techniques using WEKA Classification for Sickle Cell Disease, International Journal of Computer Science and Information Technology,5(4): 5857-5860,2014.

[3] Joshi J, Rinal D, Patel J, Diagnosis And Prognosis of Breast Cancer Using Classification Rules, International Journal of Engineering Research and General Science,2(6):315-323, October 2014.

[4] David S. K., Saeb A. T., Al Rubeaan K., Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics, Computer Engineering and Intelligent Systems, 4(13):28-38,2013.

[5] Vijayarani, S., Sudha, S., Comparative Analysis of Classification Function Techniques for Heart Disease Prediction, International Journal of Innovative Research in Computer and Communication Engineering, 1(3): 735-741, 2013.

[6] Kumar M. N., Alternating Decision trees for early diagnosis of dengue fever .arXiv preprint arXiv:1305.7331,2013.

[7] Durairaj M, Ranjani V, Data mining applications in healthcare sector a study. Int. J. Sci. Technol. Res. IJSTR, 2(10), 2013.

[8] Sugandhi C , Ysodha P , Kannan M , Analysis of a Population of Cataract Patient Database in WEKA Tool , International Journal of Scientific and Engineering Research ,2(10) ,October ,2011.

[9] Yasodha P, Kannan M, Analysis of Population of Diabetic Patient Database in WEKA Tool, International Journal of Science and Engineering Research, 2 (5), May 2011.

[10] Bin Othman M. F , Yau, T. M. S., Comparison of different classification techniques using WEKA for breast cancer, In 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006, Springer Berlin Heidelberg, 520-523,January 2007.

[11] Wikipedia, http://en.m.wikipedia.org/wiki/Dengue_fever, accessed in January 2015.

[12] Wikipedia,http://en.m.wikipedia.org/wiki/weka (machine learning), accessed in January 2015.

[13] Waikato, http://www.cs.waikato.ac.nz/ml/weka,accessed in January 2015.

[14] Wikipedia,en.m.wikipedia.org/wiki/Data_set,accessed in January 2015.

[15] KirkbyR, Frank E, WEKA Explorer User Guide for version 3-4-3, November2004.

[16] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2000.

[17] Varun Kumar and Nisha Rathee," Knowledge discovery from database Using an integration of clustering and classification", (IJACSA) International Journal of Advanced Computer Science and Applications, 2011.

[18] Swasti Singhal, Monika Jena, "A Study on WEKA Tool for Data Preprocessing, Classification and Clustering", International Journal of Innovative Technology and Exploring Engineering(IJITEE), 2013

[19] Arodz,M.Kurdziel, E. O. D. Sevre, and D.A.Yuen, "Pattern recognition techniques for automatic detection of suspicious-looking anomalies in mammograms," Comput. Methods Programs Biomed., vol. 79, pp. 135–149, 2005.

[20] L. Ramirez, N. G. Durdle, V. J. Raso, and D. L. Hill, "A support vector machines classifier to assess the severity of idiopathic scoliosis from surface topology," IEEE Trans. Inf. Technol. Biomed., vol. 10, no. 1, pp. 84–91, Jan. 2006.

[21] A. Swets, R. M. Dawes, and J. Monahan. "Better decisions through science", Scientific American, 283:82–87, October 2000.

[22] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. Numerical Recipes in C. Cambridge University Press, Cambridge, 1988.

[23] J. R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

[24] I. H. Witten and E. Frank. Data Mining - Pracitcal Machine Learning Tools and Techniques with JAVA Implementations. Morgan Kaufmann Publishers, 2000.

[25] Chen, Y.-W., & Lin, C.-J. (2005). Combining SVMs with various feature selection strategies. Available from http://www.csie.ntu.edu.tw/~cjlin/papers/features.pdf.

[26] Cheng-Lung Huang, Hung-Chang Liao b, Mu-Chen Chen c, "Prediction model building and feature selection with support vector machines in breast cancer diagnosis ", Expert Systems with Applications", 2008, 578-587 doi:10.1016/j.eswa.2006.09.041