Volume 3 Issue XII, December 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A Survey on Naïve Bayes Algorithm for Diabetes Data Set Problems

Nilesh Jagdish Vispute[1], Dinesh Kumar Sahu[2], Anil Rajput[3]
[1]M.Tech Scholar, Department of CSE, Sri Satya Sai College of Engineering, Bhopal M.P., India
[2]Ph.D Scholar, Department of Computer Science, Barkatullah University, Bhopal M.P., India
[3]Professor, Department of Comp.Sc.& Mathematics, CSA, Govt. P.G. College, Sehore M.P., India

Abstract- Diabetes Mellitus is one of the growing vitally fatal diseases world-wide. A design of classifier for the detection of Diabetes Mellitus with optimal cost and precise performance is the need of the age. The current project implementation looks further to train self-organizing weka effectively classify a diabetic patient as such. weka are so chosen due to their dynamic nature of learning and future application of knowledge. The proposed method here uses a weka implementation of the Naïve Bayes algorithm for designing of classifier. 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.
Keywords: Weka, Data mining, Classification, Naïve Bayes, Diabetic Disease Prediction.

I. INTRODUCTION

Data Mining represents a process developed to examine large amounts of data routinely collected. The term also refers to a collection of tools used to perform the process. One of the useful applications in the field of medicine is the incurable chronic disease diabetes. Data Mining algorithm is used for testing the accuracy in predicting diabetic status .Classification is one of the most frequently studied problems by DM and machine learning (ML) researchers. Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set not seen before, called training set, which contains the same set of attributes, except for the class label – not yet known. The algorithm analyses the input and produces a prediction. 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.This paper will focus on the analysis of data from a data set called diabetes data set.

II. 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. The 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.

C. 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.

D. Dhamodharan S. has done prediction of liver disease using Bayesian Classification through Naïve Bayes and FT tree algorithms. With the help of data mining techniques they have predicted and analyzed liver diseases using weka tool. They have also compared the outputs obtained from Naïve Bayes and FT tree algorithms and concluded that Naive Bayes algorithm plays a key role in predicting liver diseases.[1]

E. Solanki A.V. has used weka as a data mining technique for classification of sickle cell disease prevalent in Gujarat. They have compared J48 and Random tree algorithms and have given a predictive model for classification with respect to a person's age of different blood group types. From there experimentation it can be inferred that Random tree is better algorithm as it produces more depth decisions respect to J48 for sickle cell diseases.[2]

F. Joshi et al. has done diagnosis and prognosis of breast cancer using classification rules. By comparing classification rules such as Bayes Net, Logistic, Multilayer Perceptron, SGD, Simple Logistic, SMO, AdaBoostM1, Attribute Selected, Classification via Regression, Filtered Classifier, Multiclass Classifier and J48,They have inferred that LMT Classifier gives more accurate diagnosis i.e. 76 % healthy and 24 % sick patients.[3]

G. David S.K. et al. have used classification techniques for leukemia disease prediction. K-Nearest Neighbor, Bayesian Network, Random tree, J48 tree compared on the basis of accuracy, learning time and error rate. According to them Bayesian algorithm has better classification accuracy amongst others.[4]

H. Vijayarani S. and Sudha S. have compared the analysis of classification function techniques for heart disease prediction. Classification was done using algorithms such as Logistic, Multilayer Perception and Sequential Minimal Optimization algorithms for predicting heart disease. In this classification comparison logistic algorithm trained out to be best classifier for heart disease having more accuracy and least error rate.[5]

I. Kumar M.N. used alternating decision trees for early diagnosis of dengue fever. The ADTree correctly classifies 84 % of cases as compared to J48 which can classify only 78% of cases correctly.[6]

J. Durairaj M. and Ranjani V. have compared different data mining applications in healthcare sector. Algorithms such as Naïve, J48, KNN and C4.5 were used for Classification in order to diagnose diseases like Heart Disease, Cancer, AIDS, Brain Cancer, Diabetes, Kidney Dialysis, Dengue, IVF and Hepatitis C. Comparison study analysis revealed high accuracy i.e. 97.77% for cancer prediction and around 70% for IVF treatment through data mining techniques.[7]

K. Sugandhi C. et al. analyzed a population of cataract patient's database by weka tool. In this study, weka has been used to classify the results and for comparison purpose. They have concluded that Random Tree gives 84% classify accuracy which means better performance as compared to other algorithms used for classification accuracy performance of Naïve Bayes, SMO, J48, REP Tree and Random Tree. Thus according to their study Random Tree is the best performance classification algorithm

for cataract patient disease.[8]

L. Yasodha P. and Kannan M. performed analysis of a population of diabetic patient database using weka tool. They have classified the data and then outputs were compared by using Bayes Network, REP Tree, J48 and Random Tree algorithms. Finally the results conclude that these algorithms help to determine and identify the stage or state in which a of disease like diabetes is in by entering patients daily glucose rate and insulin dosages thereby predicting and consulting the patients for their next insulin dosage .[9]

M. Bin Othman M.F. and Yau T.M.S. have compared different classification techniques using weka for Breast cancer. In this study they have used different algorithm methods for simulating results of each algorithm and its training. They have simulated the errors by using Bayes Network, Radial Basis function, Decision Tree and pruning and Single Conjugation Rule Learner algorithms. From their work it can be concluded that Bayes Network performs best for breast cancer data. Its time taken to build model is 0.19 second and accuracy 89.7 % and least error at 0.2140 as compared to other algorithms used .[10]

N. Mihaila C. and Ananiadou S. have compared two data mining tools i.e. weka and CRF Suite on the basis of features like Lexical, Syntactic and semantic with various parameters to compare their impacts on each algorithm. The experiments have been employed in CRF Suite implementation by using Conditional Random Field algorithm and in weka by algorithms like Support Vector machine and Random Forests to identify discourse causality trigger in the biomedical domain. Classification tasks have been performed on the basis of statistics such as F score, precision and recall. As per them CRF is the best performance classifier, achieved F score = 79.35 % by combining three features as compared to other classifier.[21]

O. Thitiprayoonwonsge D. et al. have analyzed dengue infection using data mining decision tree. In this paper two datasets have been used from two different hospitals Srinagarindra Hospital and Songklanagarind Hospital, each having more than 400 attributes. Four classification algorithms have been used in this paper for experimental purpose. The first and second experiment test got an accuracy of 97.6% and 96.6%. The third experiment extracts useful knowledge. Another objective of this paper was to detect day abatement of fever also referred as day0. In fourth experiment of day0 accuracy is very low as compared to other three experiments. Therefore physician need day0 amongst patient in order to treat them.[22]

III. OVERVIEW OF PROPOSED APPROACH

A. 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.

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.

B. 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, , 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, 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, 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.

1) *Correctly Classified Accuracy:* It shows the accuracy percentage of test that is correctly classified.
2) *Incorrectly Classified Accuracy:* It shows the accuracy percentage of test that is incorrectly classified.
3) *Mean Absolute Error:* It shows the number of errors to analyze algorithm classification accuracy.
4) *Time:* It shows how much time is required to build model in order to predict disease.
5) *ROC Area:* Receiver Operating Characteristic19 represent test performance guide for classifications accuracy of diagnostic test

B. Data Mining Techniques

The data mining technique have been used by us to predict diabetes database 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:

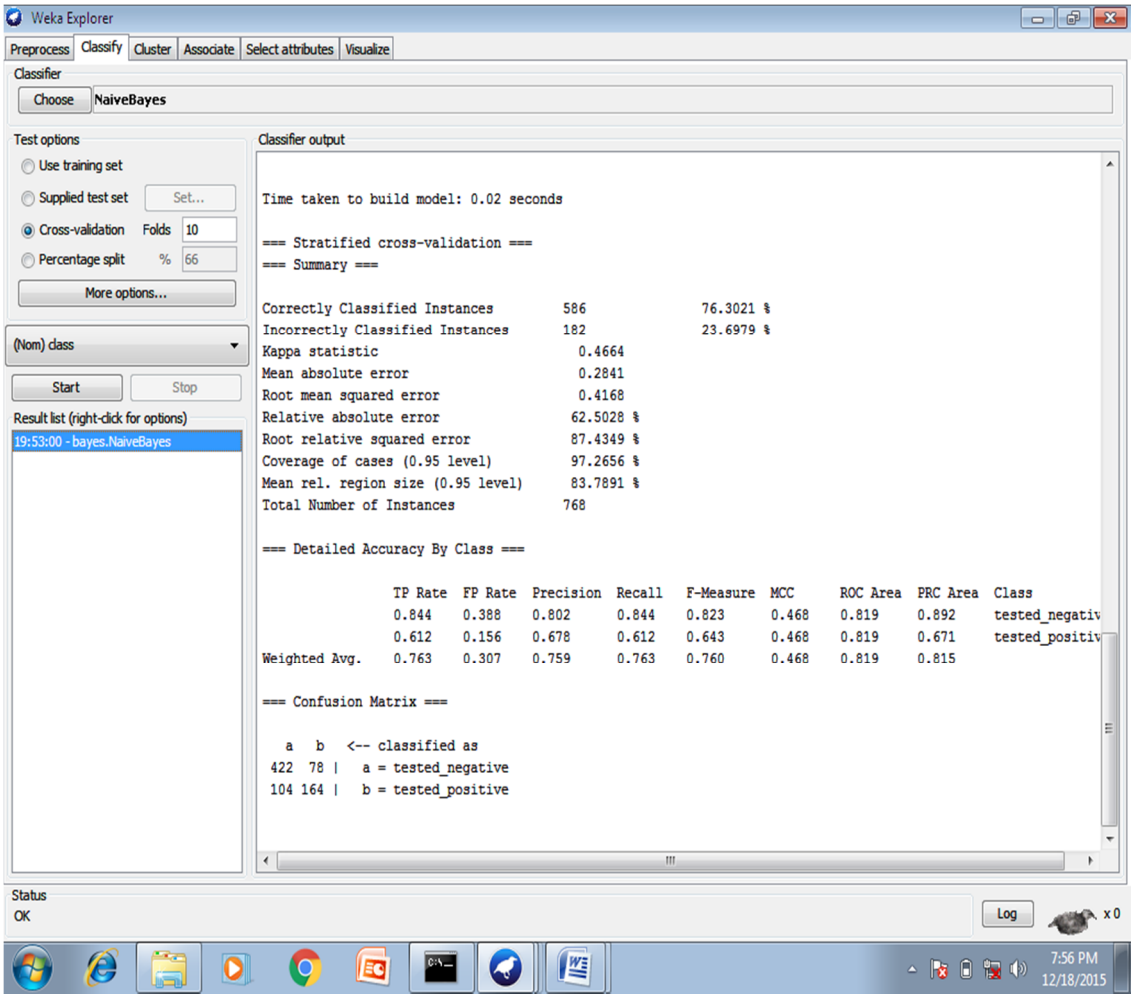
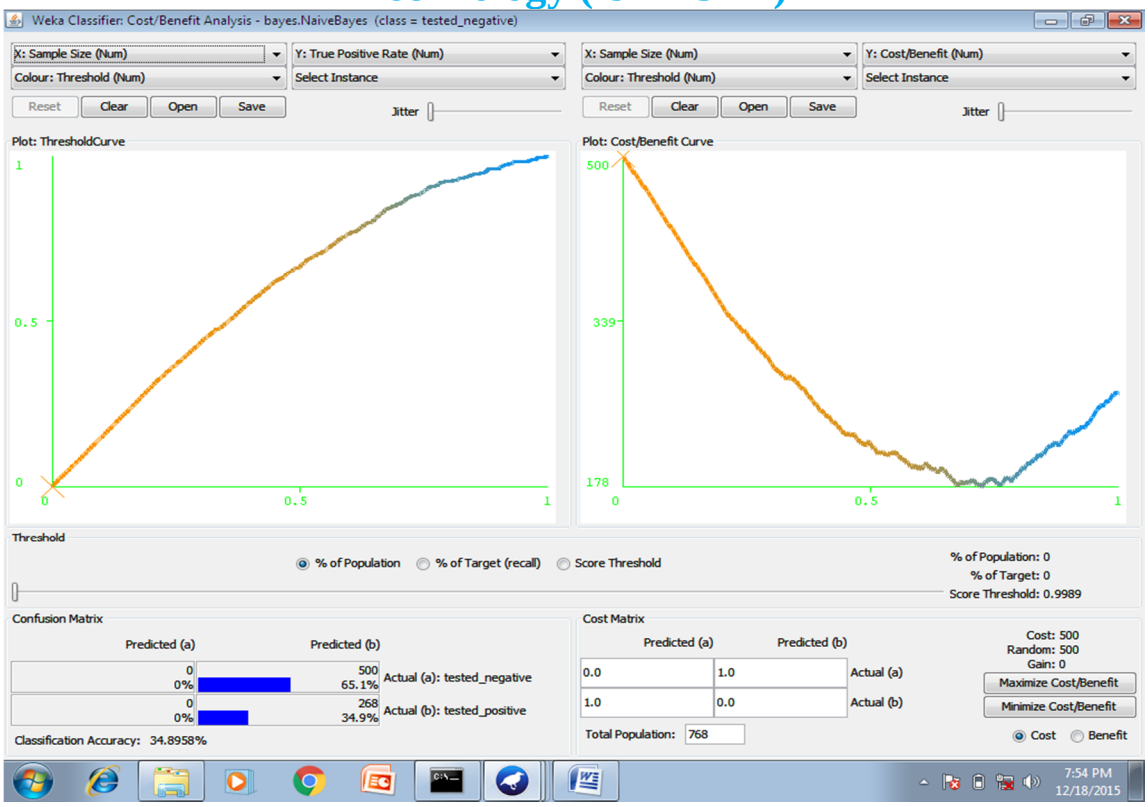


Figure 1. Naïve Bayes Algorithm applied on the Diabetic dataset

The figure shown above is the weka tool containing Naïve Bayes Algorithm applied on the Diabetic dataset.

Volume 3 Issue XII, December 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)



Graph 2. Cost/Benefit Analysis of the Diabetic Dataset

The graph as shown above is the analysis of the cost/benefit curve on the basis of Diabetic dataset Accuracy chosen.

IV. RESULTS/ DISCUSSION

Explorer, the data mining techniques that have been used by us using algorithms Naïve Bayes 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 Naïve Bayes 34.8958 % correctly instances accuracy with minimum Naïve Bayes Mean Absolute Error = 0.2841 having maximum Naïve Bayes ROC =0.819 time taken to build model=0.02 seconds Classification Accuracy=34.8958 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.

V. CONCLUSION / FUTURE WORK

From these experiments, it can be concluded that, results of the proposed system are expected to perform better than those in the current literature survey. Though, the expected results shall be comparable with those of the weka, but with lesser Accuracy of the propose system, and minimum compromise with the quality, such a trade-off is acceptable. 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 interfaces: Explorer. interface has its own classifier algorithms. We have used algorithms i.e. Naïve Bayes, 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 are several scoring algorithms for accuracy but for our experimentation we have used only algorithms. The outputs obtained from Explorer flow are 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 it was inferred that Nave Bayes the best performance classifier algorithms as they achieved an of Accuracy=34.8958 %, 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.

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
[27] N. J. Nilsson, “Introduction to Machine Learning,” 2010. http://ai.stanford.edu/~nilsson/mlbook.html
[28] M. S. Sapna and D. A. Tamilarasi, “Fuzzy Relational Equation in Preventing Neuropathy Diabetic,” Internati- onal Journal of Recent Trends in Engineering, Vol. 2, No. 4, 2009, p. 126.
[29] L. Carnimeo and A. Giaquinto, “An Intelligent System for Improving Detection of Diabetic Symptoms in Retinal Images,” IEEE International Conference on Information Technology in Biomedicine, Ioannina, 26-28 October 2006