INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

(Volume2, Issue5)

Available online at: www.Ijariit.com

DIABETES PREDICTION BY SUPERVISED AND UNSUPERVISED LEARNING WITH FEATURE SELECTION

Rabina[1], Er. Anshu Chopra[2]

Research Scholar (CSE) [1], Assistant Professor (CSE)[2]

rabinabagga@gmail.com[1], anshuchopra19@gmail.com[2]

S.S.C.E.T, Badhani, Pathankot, India.

Abstract: Two approaches to building models for prediction of the onset of Type diabetes mellitus in juvenile subjects were examined. A set of tests performed immediately before diagnosis was used to build classifiers to predict whether the subject would be diagnosed with juvenile diabetes. A modified training set consisting of differences between test results taken at different times was also used to build classifiers to predict whether a subject would be diagnosed with juvenile diabetes. Supervised were compared with decision trees and unsupervised of both types of classifiers. In this study, the system and the test most likely to confirm a diagnosis based on the pre-test probability computed from the patient's information including symptoms and the results of previous tests. If the patient's disease post-test probability is higher than the treatment threshold, a diagnostic decision will be made, and vice versa. Otherwise, the patient needs more tests to help make a decision. The system will then recommend the next optimal test and repeat the same process. In this thesis find out which approach is better on diabetes dataset in weka framework. Also use feature selection techniques which reduce the features and complexities of process.

Keywords— data mining, weka, Prediction, machine learning.

I. INTRODUCTION

Data Mining refers to the process of separating knowledge or other interesting paradigm from large gathering of data [1]. It includes iterative series alike Data Cleaning (removal of noise and inconsistent data), Data Integration (combining data from multiple heterogeneous sources), Data selection (selection of relevant data for analysis), Data Transformation (data is transformed into forms suitable for mining), Pattern Evaluation (identification of interesting patterns) and Knowledge Presentation (use of visualization and knowledge presentation techniques for presenting the mined knowledge to users).

Diabetes is a leading health issue not only in industrial but advancing Countries as well and its incidence is inclining. It is a condition in which the body inadequate to generate or suitably use the hormone called insulin that ‘‘unlocks” the cells of the body and permits glucose to enter and fuel them. There are various factors which required to be investigated to diagnose the diabetic patient, and this makes the physician’s job difficult. So we will carry out an profitable technique for categorization of patients for diabetes with the use of soft computing method. Our considerable establishment is to enhance the accuracy of diabetes dataset. Several methods have been investigated in the past to specify diabetic patients and anticipate the accuracy.

Diabetes mellitus is the most common endocrine disease. The disease is characterized by metabolic abnormalities and by long-term complications involving the eyes, kidneys, nerves, and blood vessels. The diagnosis of symptomatic diabetes is not difficult. When a patient presents with signs and symptoms attributable to an osmotic dieresis and is found to have hyperglycemia essentially all physicians agree that diabetes is present. The two major types of diabetes are Type I diabetes and Type II diabetes. Type I diabetes is usually diagnosed in children and young adults, and was previously known as juvenile diabetes [4]. Type I diabetes mellitus (IDDM)

patients do not produce insulin due to the destruction of the beta cells of the pancreas. Therefore, therapy consists of the exogenous administration of insulin. Type II diabetes is the most common form of diabetes. Type II diabetes mellitus (NIDDM) patients do produce insulin endogenously but the effect and secretion of this insulin are reduced compared to healthy subjects [5].

Classification is the identification of data for its most useful and competent use. In a fundamental prospective to storing computer data, data can be distributed confer to its critical value or how frequent it requires to be promoted, with the most analytical or often-used data safeguard on the fastest media although other data can be saved on slower (and less expensive) media [3]. This kind of categorization influence to developed the use of data storage for multiple purposes - technical, administrative, legal, and economic.
Optimization refers to the process of choosing the best solution to an issue from all possible outcome with respect to reducing or increasing the optimization principle. It is an applied science which analyzed the best values of the parameters of a issues that may take under particular conditions [8]. Optimization, in its most simple way, intention to retrieve the pertinent framework values which facilitate an aim operation to produce the *minimum* or *maximum* value [9]. It has been the utmost objective although solving any specific issues. Optimization is pervasive and instinctive process that forms a combined part of our day-to-day life [10]. Optimization issues arise in several disciplines like engineering designs, agricultural sciences, manufacturing systems, economics, physical sciences, pattern recognition etc. in fact optimization methods are being broadly used in several spheres of human activities, where some difficult decisions have and can be presented by means of a mathematical model.

Optimization can therefore be viewed as one of the major quantitative tools in network of decision making, in which decisions have to be taken to optimize one or more aims in some prescribed set of circumstances. To rational employ optimization for various issues there is a requirement to advance developed and reliable computational algorithms. Today, optimization consists of a broad variety of techniques from Operations Research, Artificial Intelligence and Computer Science, and is used to enhance business processes in practically all industries.

Another current trend is the combination of optimization methods for issues that do not lend themselves easily to one method alone and so far many optimization methods have been used alongside with clustering algorithms to enhance the potency and quality of the final clusters and form more appropriate clusters. The evolutionary and search based approach revealed above have been used so far for optimization of real world issues. Some of them according to their chronological order of occurrence as explained by Abraham et al. [11] Genetic Algorithm (GA), Hill climbing, Simulated Annealing, Particle Swarm Optimization (PSO), Ant Colony Optimization(ACO), Differential Evolution(DE), Harmony Search and Artificial Bee Colony Algorithm.

II. LITERATURE REVIEW

Thiyagarajan c et al (2016), In this work [13], an effective machine learning algorithm is proposed for the classification of type dm patients. This machine learning algorithm used for classification will find the optimal hyper-plane which divides the various classes. By using this machine learning algorithm, the classification accuracy is achieved for classifying the diabetes patients. So, there is an increase in interest by various researchers to set up a medical system which can screen a large number of people for life threatening disease such as cardio vascular disease, retinal disorder in diabetic patients. Several data mining and machine learning methods have been used for the diagnosis, prognosis, and management of diabetes.

Rc kessler et al (2016), this paper represent [14], heterogeneity of major depressive disorder (mdd) illness course complicates clinical decision-making. The report results of model validation in an independent prospective national household sample of 1056 respondents with lifetime mdd at baseline. Ml model prediction accuracy was also compared with that of conventional logistic regression models. area under the receiver operating characteristic curve based on ml (0.63 for high chronicity and 0.71–0.76 for the other prospective outcomes) was consistently higher than for the logistic models (0.62–0.70) despite the latter models including more predictors. These results confirm that clinically useful mdd risk-stratification models can be generated from baseline patient self-reports and that ml methods improve on conventional methods in developing such models.

Seokho kang et al(2015), This paper [15] propose an efficient and effective ensemble of svms, called e3 -svm. The proposed method excludes superfluous data points when constructing a svm ensemble, thereby yielding a better classification performance. The proposed method consists of two phases. The first phase is to select the data points that are likely to be the support vectors by applying data selection methods. The second phase is to construct a svm ensemble using the selected data points. We demonstrated the efficiency and effectiveness of the proposed method using the real-world dataset of the anti-diabetic drug failure prediction problem for type 2 diabetes. Experimental results show that the proposed method requires less training time to achieve comparable success, compared to the conventional svm ensembles. Moreover, the proposed method obtains more reliable prediction results for each independent run of constructing an ensemble. in conclusion, firstly, the proposed method provides an efficient and effective way

to use svm for large-scale datasets. Secondly, we confirmed the suitability of svm for the anti-diabetic drug failure prediction problem with an accuracy of about 80%.

klaus donsa et al (2015), This work [16], cover relevant parameters to personalize diabetes therapy, how cdss can support the therapy process and the role of machine learning in this context. Moreover, we identify open problems and challenges for the personalization of diabetes therapy with focus on decision support systems and machine learning technology. it also supports long-term disease management, aiming to develop a personalization of care according to the patient's risk stratification. Personalization of therapy is also facilitated by using new therapy aids like food and activity recognition systems, lifestyle support tools and pattern recognition for insulin therapy optimization. Therefore, the personalization of the patient's diabetes treatment is possible at different levels. It can provide medication support and therapy control, which aid to correctly estimate the personal medication requirements and improves the adherence to therapy goals.

Davar Giveki (2012)

This paper [17] presents a unique automatic prospect to diagnose Diabetes disease on the basis of Feature Weighted Support Vector Machines (FW-SVMs) and Modified Cuckoo Search (MCS). The model composed of three phases: Firstly, there is a use of PCA to choice an optimal subset of characteristics out of set of all the features. Secondly, Best feature weights are approximated with the use of Mutual Information (MI) on the basis of their degree of significance. Ultimately, MCS is practice to choice the best parameter values from all parameters. An accuracy of 93.58% is retrieved by the prospect MI-MCS-FWSVM technique on UCI dataset. Furthermore, Modified Cuckoo Search is used for speeds up the confluence the algorithm and also to find the optimal values for parameters of SVM. This technique can be united with medical software's to reinforcement physicians so that they can take make more accurate decisions about Diabetes disease.

Meryem SAIDI (2011)

The use of expert systems and artificial intelligence methods in disease diagnosis is expanding slightly. Artificial Immune Recognition System (AIRS) is one of the techniques that is used in medical classification issues. AIRS2 is an altered interpretation of the AIRS algorithm. In this paper [18] an altered AIRS2 called MAIRS2 was used where we exchanged the K- nearest neighbour's algorithm with the fuzzy K-nearest neighbours to enhance the diagnostic accuracy of diabetes diseases. The diabetes disease dataset which taken from UCI machine learning repository. The highest categorization accuracy was retrieved by employ the AIRS2 and MAIRS2 using 10-fold cross-validation, which was 82.69% and 89.10% consequently. MAIRS2 was employ to diabetes disease. From this they examine that an expansion in the number of memory cells led to a better perception rate but side by side, it minimize the degree of data reduction.

Nahla H. Barakat (2010)

In this paper [19] support vector machines (SVMs) are recommended for the diagnosis of diabetes. In this, they used a description module, which is known as "black box" model of an SVM which we used for diagnostic (classification) decision. Results retrieved on diabetes dataset with the use "black box" proves that it is an emerging tool that is delivered by intelligible SVM's for the prediction of diabetes, with prediction accuracy of 94%, sensitivity of 93%, and specificity of 94%. Future work is to conduct a proposed research to further clarify the predictive outcomes retrieved by the proposed rules. (SVM) have been developed by Vapnik (1995) and are gaining popularity due to many attractive features, and promising empirical performance. The formulation embodies the Structural Risk Minimisation (SRM) principle, which has been shown to be superior, (Gunn et al., 1997), to traditional Empirical Risk Minimisation (ERM) principle, employed by conventional neural networks.

Muhammad Waqar Aslam (2010)

This paper [20] carried out genetic programming (GP) and a deviation of genetic programming known GP with relatively partner selection (CPS) for identification of diabetes. The prospective structure involves two phases. In first stage, genetic programming is used to generate an individual from training data that converts the possible characteristics to a single feature such that it has distinct values for healthy and patient (diabetes) data. In the second stage, test data is used for testing of that individual features. The prospective structure was able to attain 78.5±2.2% accuracy. The outcome proves that GP based classifier enforcement better in the diagnosis of diabetes disease. Diagnosis of diabetes depends on many other factors and hence makes the medical practitioners' job difficult, at times.

Santi Wulan Purnami (2010)

This paper [21] presents a unique technique for diabetes disease diagnosis with the use altered spline smooth support vector machine (MS-SSVM) to retrieved optimal accuracy outcomes, firstly uniform Design technique was used for choice of most prevalent characteristics. The enforcement of this technique capability with the use of 10-fold cross validation accuracy, confusion matrix. The retrieved categorization accuracy using 10-fold cross authorization is 96.58% in relatively with other spline SSVM method. The outcome of this research proved that the altered spline SSVM was reliable to identify diabetes disease diagnosis and this is very encouraging outcome relative to the prior reported outcome.

III. RESULTS AND DISCUSSIONS

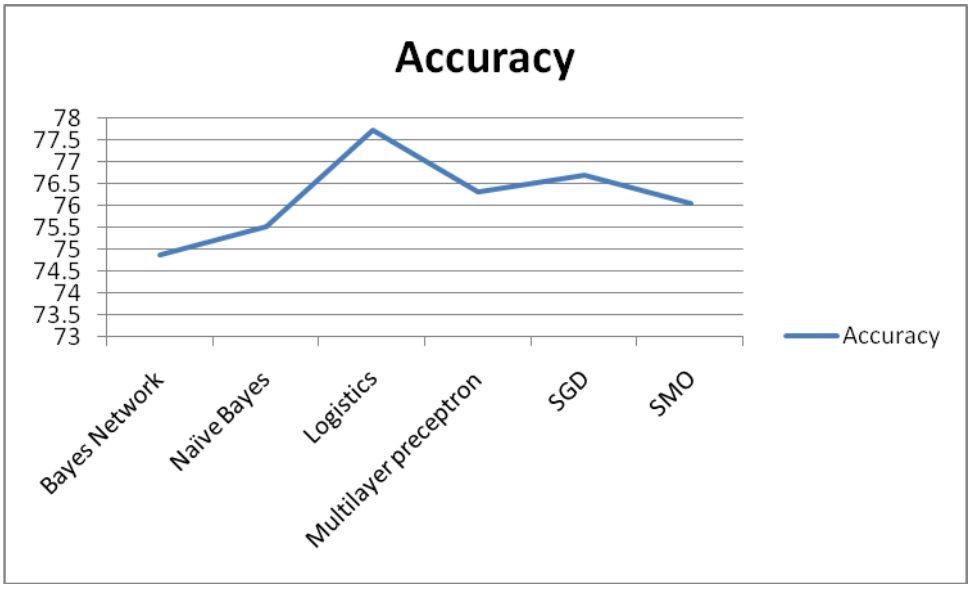

Fig 3.1: Comparison on the base of Accuracy

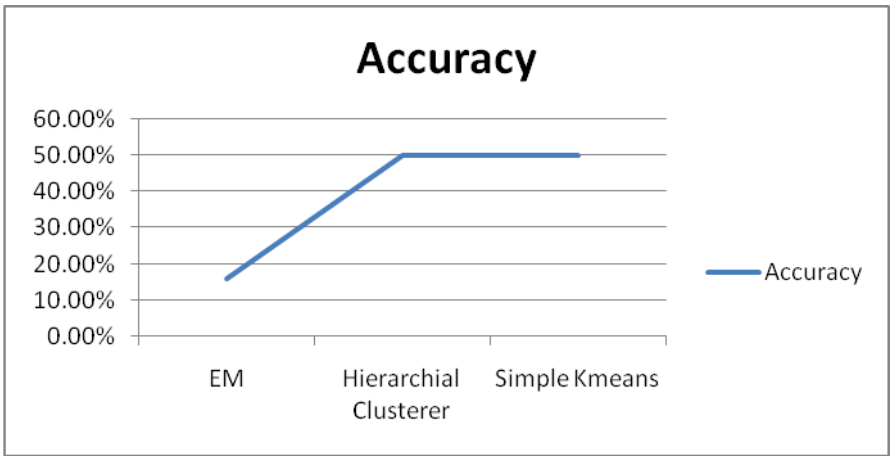

Fig 3.2: Comparison on the base of Accuracy

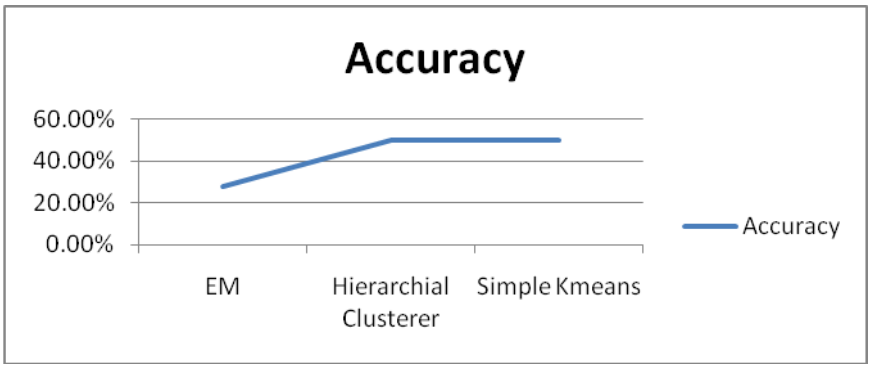

Fig 3.3: CFS(BEST FIRST), accuracy result

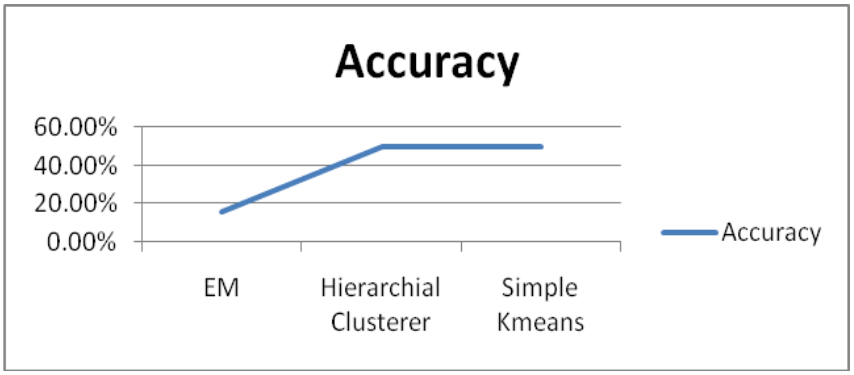

Fig 3.4: Gain Ratio Accuracy Result

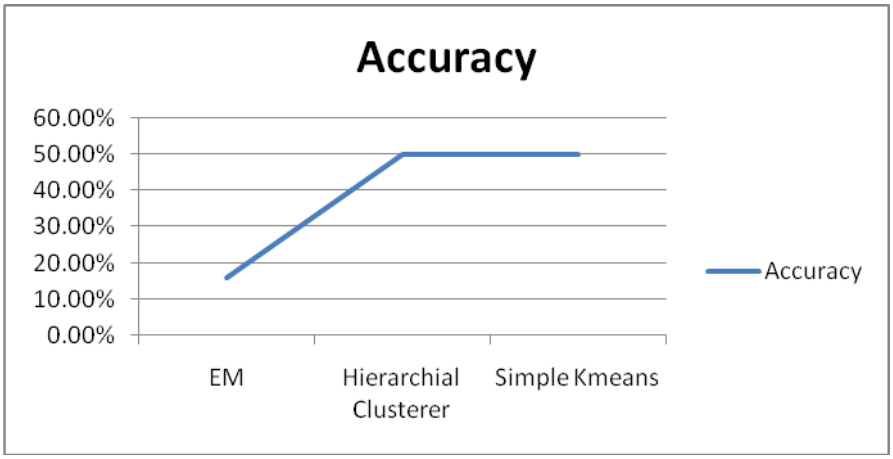

Fig 3.5: Correlation Accuracy Result

IV. CONCLUSIONS

From the results of the experiments conducted, it can be seen that ensemble approaches with decision trees even better diabetes prediction accuracy for juvenile subjects (in the context of our dataset)than cascade correlation based neural networks. Another area where decision tree approaches have potential advantages over neural networks is that they are fast and easy to build and understand. Decision trees have built-in methods to deal with missing attributes (for example allowing examples to go down multiple branches with weights as in our implementation). However, in this case a more explicit representation of missing values was very useful. Moreover, it can also be inferred from the results, that the absence of some medical tests reveal important information which helps to predict the occurrence of Type 1 diabetes in juvenile subjects with better accuracy. Such information needs to be embedded, as done here with indicator attributes associated with each test result attribute.

REFERENCES

[1] S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

[2] J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[3] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.

[4] M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," in *Proc. ECOC'00*, 2000, paper 11.3.4, p. 109.

[5] R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.

[6] (2002) The IEEE website. [Online]. Available: http://www.ieee.org/

[7] M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/

[8] *FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.

[9] "PDCA12-70 data sheet," Opto Speed SA, Mezzovico, Switzerland.

[10] A. Karnik, "Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP," M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.

[11] J. Padhye, V. Firoiu, and D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.

[12] Thiyagarajan, C., K. Anandha Kumar, and A. Bharathi. "A Survey on Diabetes Mellitus Prediction Using Machine Learning Techniques."*International Journal of Applied Engineering Research* 11.3 (2016): 1810-1814.

[13] Kessler, R. C., et al. "Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports." *Molecular psychiatry* (2016).

[14] Kang, Seokho, et al. "An efficient and effective ensemble of support vector machines for anti-diabetic drug failure prediction." *Expert Systems with Applications* 42.9 (2015): 4265-4273.

[15] Donsa, Klaus, et al. "Towards personalization of diabetes therapy using computerized decision support and machine learning: some open problems and challenges." *Smart Health*. Springer International Publishing, 2015. 237-260.

[16] Davar Giveki, Hamid Salimi, GholamReza Bahmanyar, Younes Khademian. "Automatic Detection of Diabetes Diagnosis using Feature Weighted Support Vector Machines based on Mutual Information and Modified Cuckoo Search", pp.312-314, 2012

[17] Mohamed Amine Chikh, Meryem Saidi, Nesma Settouti, "Diagnosis of Diabetes Diseases Using an Artificial Immune recognition System2 (AIRS2) with Fuzzy K-nearest Neighbour", Springer Journal of Medical Systems, 2011

[18] Nahla H. Barakat, Andrew P. Bradley, Mohamed Nabil H. Barakat, "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus", IEEE transaction on Information technology in bioinformatics, Vol. 14, NO. 4, Jul 2010

[19] M.W. Aslam, A.K. Nandi, "Detection of diabetes using genetic programming", 18th European Signal Processing Conference, 2010

[20] Santi Wulan Purnami, Jasni Mohamad Zain and Abdullah Embong, "Data mining techniques for medical diagnosis using a new smooth SVM", Communications in Computer and Information Science, Vol 88, Part 1, pp.15-27,2010

