Diabetes Disease Diagnosis Using Multivariate Adaptive Regression Splines

D. Senthilkumar[1], S. Paulraj [2]
[1] Assistant Professor, Computer Science and Engineering,
Bharathidasan Institute of Technology, Anna University
[2] Associate Professor, Department of Mathematics, Anna University
[1] chandsent@yahoo.co.in
[2] paulraj@annauniv.edu

***Abstract* - Diabetes mellitus is a chronic disease, also associated with an increased risk for heart disease and is emerging as a serious health challenge in India. It requires continuing medical care and patient self-management education to prevent severe problems and to reduce the risk of long-term problems. Healthcare industry stores a large amount of data which is not properly used to discover the hidden patterns and relationships. Disease diagnosis is one of the applications where data mining algorithms are proving successful results. In recent years, several researches have been conducted to develop intelligent clinical decision support system to help the physician in diagnosing the diabetes. This paper provides an introduction to the theory and highlights the importance of Multivariate Adaptive Regression Splines (MARS) in the disease diagnosis through the collected data for diabetes to develop an intelligent decision support system to help the physicians. MARS model obtained better accuracy with minimum number of predictors and outperformed by handling nonlinearities, missing data and interactions among predictors compared to other methods. The proposed approach is easily understandable, provides a better and faster model for diagnosing of diabetes patients.**

Keywords- Data mining, Diabetes mellitus, Decision support system, MARS

1. INTRODUCTION

Now a days health care industry generates huge amounts of data about patients. Medical data analysis is important for medical decision making and management. Analyzing and processing the huge amounts of data generated by healthcare industry are too complex by traditional methods. Data mining is the process of selecting, exploring, and modeling large amounts of data to discover unknown patterns or relationships useful to decision making [1-2]. Data mining methods apply to decision making in many areas like marketing, fraud detection, investment, manufacturing, telecommunication, engineering, medicine, biomedical research [2-5].

Health care industry facing significant challenges in terms of health care service and quality of patient care. A quality service implies, speed up the diagnostic process, reduce overuse of medical tests, save costs, and improve the accuracy of diagnosis [6]. Clinical decisions are often made based on doctors 'intuition and experience rather than on the knowledge-rich data hidden in the database [5]. Machine learning algorithms are used as a tool to extract hidden interesting pattern from the medical database. Interesting patterns will be used to assist the physicians to improve the diagnosis speed, accuracy and/or reliability [7].

Decision support systems (DSS) are defined as a computer based system developed to assist decision makers in the decision making [8]. It reduces the diagnosis time and increase the diagnosis accuracy in complicated diagnosis decision process, as well as the cost of care. Clinical decision support systems (CDSS) are computer systems designed to impact clinician decision making about individual patients [9]. Computerized clinical decision support systems (CCDSSs) are information technology-based systems designed to improve clinical decision-making. Characteristics of individual patients are matched to a knowledge base, and software algorithms generate patient-specific information in the form of assessments or recommendations [10].

The main objective of this research is to develop an intelligent decision support system using multivariate adaptive regression spline (MARS) to help the physicians for better decision making from the historical data of the patients. However, to the best of the authors' knowledge, MARS has not yet been applied for the diagnosis of diabetes.

In the next section, we briefly describe some of the fundamentals of diabetes mellitus, Pima Indians Diabetes Dataset and Data Mining methods. Section 3 introduces the, theory and highlights the importance of MARS algorithm. Section 4, we discussed the evaluation of the MARS model and the experimental results and in Section 5 we present our conclusions.

2. DIABETES MELLITUS AND DIAGNOSIS

Diabetes mellitus is a chronic disease and a greatest health challenge in India. International Diabetes Federation (IDF), has raised a serious alarm for India by saying that nearly 52% of Indians aren't aware that they are suffering from high blood sugar. There are currently 62 million diabetics — an increase of nearly 2 million in just one year and this number is expected to cross the 100 million mark by 2030 [11]. It requires continuing medical care and patient self-management education to prevent severe problems and to reduce the risk of long-term problems [12]. Diabetes is associated to develop various types of disease like cardiovascular disease, peripheral vascular disease, lower limb amputations, blind disease, kidney diseases [13].

Diabetes complications can be prevented or delayed by early identification of people at risk. The National Diabetes Information Clearinghouse (NDIC) categorizes diabetes into type-1 diabetes , which is normally diagnosed in children and young adults, type-2 diabetes, i.e., the most common form of diabetes due to obesity and gestational diabetes, it develops during pregnancy time. To diagnosis diabetes or pre-diabetes, fasting plasma glucose (FPG) or the 75-g oral glucose tolerance test (OGTT) is generally used [13][14].

Prediabetes is a condition in which blood glucose levels are higher than normal but not high enough for a diagnosis of diabetes It is also called as impaired fasting glucose (IFG) or impaired glucose tolerance (IGT). People with prediabetes are at increased risk of developing type 2 diabetes, heart disease, and stroke [13][14].

2.1 Pima Indians Diabetes Dataset and Data Mining Methods

Data used for this research was collected from [15] UCI Machine Learning Repository- Pima Indians Diabetes Data (PIDD) Set. The collected record has 768 samples with eight variables include the No. of times pregnant(NOP), Plasma glucose concentration-(OGTT), Diastolic blood pressure (mmHg)-(DIAS_BP), Triceps skin fold thickness (mm)-(TRI_SFT), 2-Hour serum insulin (mm U/ml)-(SERUM_2HR), Body mass index (weight in kg/height in m2)-(BMI), Diabetes pedigree function-(DIA_PED_FN), age of patient-(AGE) and class variable (0 or 1 – '0' means a negative test for diabetes and '1' means positive test for diabetes). Descriptive statistics about the data set is mentioned in the Table I.

Table I
Descriptive Statistics

Variable	N	N Distinct	Mean	Min	Max
AGE	768	52	33.24089	21.00000	81.00000
BMI	768	248	31.99258	0.00000	67.10000
CLASS	768	2	0.34896	0.00000	1.00000
DIA_PED_FN	768	517	0.47188	0.07800	2.42000
DIAS_BP	768	47	69.10547	0.00000	122.00000
NOP	768	17	3.84505	0.00000	17.00000
OGTT	768	136	120.89453	0.00000	199.00000
SERUM_2HR	768	186	79.79948	0.00000	846.00000
TRI_SFT	768	51	20.53646	0.00000	99.00000

Different data mining methods are applied to classification and diagnosis of diabetes disease related to PIDD in literature. [16] Hasan Temurtas, Nejat Yumusak and Feyzullah Temurtas discussed about Neural network classification accuracy, it ranges from 78% to 82%. [17] Shankaracharya, Devang Odedra, Subir Samanta,and Ambarish S. Vidyarthi reviewed the classification accuracy it ranges from 71% to 98 %. Also discussed the advantages and disadvantage of various methods. [18] G. Magudeeswaran and D. Suganyadevi compared various data mining techniques on diabetes disease diagnosis it ranges from 67 to 78%. [19] Shelly Gupta, Dharminder Kumar and Anand Sharma compared, classification accuracy ranges from 73% to 77%. [20] Joseph L. Breault used rough sets in diabetic databases. Accuracy of the initial random sample was 82.6%, but the mean accuracy was 73.2%. Also, compared with other methods, the mean accuracy was 73.9%. [21]-[26] From the study it is identified that different data mining algorithm are applied to PIDD like neural networks, fuzzy theory, radial basis function, general regression neural network, multimodal evolutionary algorithm, classification and regression tree, support vector machines and various hybrid approaches, etc.,. [27] Neural network algorithms work better on the diabetes diagnosis problem than others where input variables are interrelated. It is identified that results are varied when same algorithm employed in the dataset from the literature. Also MARS has not yet been applied for the diagnosis of diabetes on the PIDD.

3. MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS)

The multivariate adaptive regression splines (MARS) method is a non-parametric flexible piecewise regression modeling of high dimensional problems developed by Friedman [28], where there are many input variables. It shows a great promise for fitting nonlinear multivariate functions [29]. MARS applied for various prediction and data mining applications in recent years [30]–[34]

MARS performs well for predictive modeling of continuous outcomes but also has application in other areas. Discrimination, classification, optimization, predictive modeling of binary outcomes [32],[35], knowledge discovery [36], nonlinear modeling of time series analysis [37] are all recent applications where MARS has shown success or should be considered for use [33].

It is a regression based technique, its outputs are a linear function that is readily understood by the analyst and can be used to explain the model to management. [40] It deals with multidimensional data, evaluating each factor and possible interaction among them. It eliminates a certain number of predictors if they do not contribute to increase the performance of the final model [28],[38]-[39]. [51] MARS model of the form

$$\hat{f}(x) = \sum_{i=1}^{k} c_i B_i(x) \qquad (1)$$

The model is a weighted sum of basis functions $B_i(x)$. Each c_i is a constant coefficient.

Each <u>basis function</u> $B_i(x)$ takes one of the following three forms:

1) a constant 1. There is just one such term, the intercept.

2) a *hinge* function has the form max $(0, x-c)$ or max $(0, c-x)$, where c is a constant called the knot.

3) a product of two or more hinge functions.

Missing data occurs frequently in the large databases. MARS handle missing data by creating a basis function for any variable with missing data. MARS is a competitor to neural networks that does not suffer from any of the limitations of neural networks like nonlinearities, missing data and interactions. It is not a black box, derived model are easily understandable [40]. Training times for this method tend to be much faster than feed-forward neural networks using back-propagation [41].

3.1 Most Important Predictor Identification

To diagnose disease physician has to consider many factors from the data obtained from the patients. However, Factors such as lack of experience by the ex- parts, or their fatigue, may lead to erroneous diagnosis [42]. Most of the researcher's aim is to identify which predictors are used for diagnosis and prediction. The most important predictor is always increasing the predictive accuracy of the model. Generalized cross validation (GCV) developed by Craven and Wahba [43]. Friedman uses the modified form of the generalized cross-validation criterion is used to identify the most important predictor, rank the predictor and eliminate insignificant predictor of the model [38]-[40], [44]. If a variable receives a score 100 and 0, it is the most important predictor and not used in the MARS model respectively [45]. It is defined as

$$GCV(M) = \frac{\frac{1}{N}\sum_{i=1}^{N}[y_i - \hat{f}_M(x)]^2}{[\ 1 - \frac{C(M)^*}{N}]^2} \qquad (2)$$

$$where\ C(M)^* = C(\ M) + \delta \cdot M \qquad (3)$$

N is the number of observations

*C(M)** is a complexity cost function of the model generating *f*, the default is to set equal to a function of the effective number of parameters

M is the number of non constant basis functions in the MARS model and

δ is a cost for each basis-function optimization and is a smoothing parameter for the procedure.

4. EVALUATION OF THE MARS MODEL

A medical diagnosis is a classification process. Using the data mining techniques to perform this classification is becoming more frequent. Various data mining methods are available related to classification and diagnosis of diabetes disease in literature. Based on the previous studies MARS is employed in the Pima Indians Diabetes Data (PIDD) to a diagnosis of diabetes disease, to identify possible interaction among predictors and to identify most important predictors.

4.1 Model accuracy

One of the important tasks in the data mining is to estimate the accuracy of the model is depicted in the Figure 1 . In general the dataset is divided into two parts (training set 70% - to build the model and test set 30% –measure its performance). If the data set is very small, it is not advisable to use a large portion for the testing, not to lose the information from the dataset. An alternative method for training and testing the accuracy of the model for the small data set is cross validation. [46]-[50]

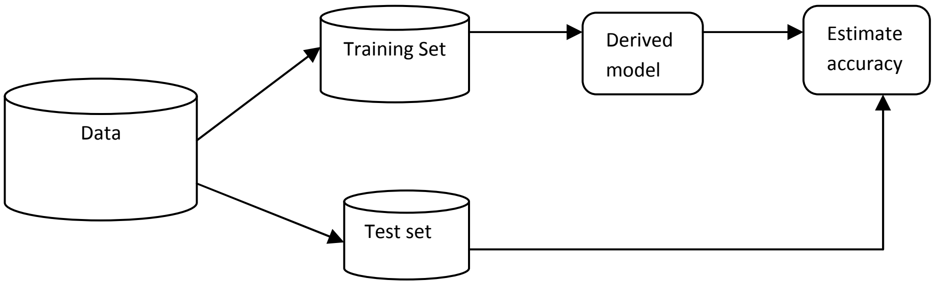

Figure1.Classification accuracy

4.2 Cross Validation

Cross Validation is a method of estimating the performance of a developed model. Divide the data randomly into k subsets or "fold" of equal size.. Train the model on k-1 subsets, use one subset for testing. Training and testing are performed k times. Accuracy estimate is the average performance of the k sets. Stratified 10-fold cross-validation is advised for low bias and variance. A confusion matrix brief the types of errors in the derived model, is estimated by applying the model to test data in the target class already available and is compared with predicted target class. [46]-[50] It is a square matrix with n dimensions, where n is the number of target classes. Confusion matrix for two-by-two is depicted in the Table II.

Table II
Confusion matrix

		Actual Label	
		Normal	Abnormal
Predicted Label	Normal	True Positive (TP)	False Negative (FN)
	Abnormal	False Positive (FP)	True Negative (TN)

The accuracy of a model on a given test set is the percentage of test set that are correctly classified by the classifier. Measures are defined [49],[50] as follows

$$sensitivity = \frac{t\_pos}{pos} \qquad (4)$$

$$specificity = \frac{t\_neg}{neg} \qquad (5)$$

$$precision = \frac{t\_pos}{t\_pos+f\_pos} \qquad (6)$$

where *t_pos* is the number of true positives (*"normal"* tuples that were correctly classified as such), *pos* is the number of positive (*"normal"*) tuples, *t_neg* is the number of true negatives (*"Abnormal"* tuples that were correctly classified as such), *neg* is the number of negative (*"Abnormal"*) tuples, and *f_ pos* is the number of false positives (*"Abnormal"* tuples that were incorrectly labeled as *"normal"*). Accuracy is defined as follows

$$accuracy = sensitivity\frac{pos}{(pos+neg)} + specificity\frac{neg}{(pos+neg)} \qquad (7)$$

4.3 Results and Discussion

In MARS analysis, we used 10-fold cross-validation, basis function is set to 25 and allowed 20 orders of interaction. The most important predictor selection result using MARS is summarized in the Table III and is compared with CART & Random Forest. It indicates that the most important predictor in the diagnosis of diabetes is Plasma glucose concentration, Diastolic blood pressure (mm Hg), age of patient, Body mass index (weight in kg/height in m2), Diabetes pedigree function, 2-Hour serum insulin (mm U/ml). Comparing the algorithms MARS model uses only 6 predictors as important out of 8 predictors All the models rank OGTT as100% most important predictor for the final model. From the Expert Committee statement on the Diagnosis and Classification of Diabetes Mellitus [American Diabetes Association (Diabetes Care 28:S4-S36, 2005) it can be concluded that machine learning algorithms perform well in the diagnosis by identifying OGTT as the most important predictor. If a variable receives a score 100 and 0, it is the most important predictor and not used in the MARS model respectively [45]. Modified Generalized Cross Validation eliminated insignificant predictors Triceps skin fold thickness and No. of times pregnant in the final MARS model.

Table III
Variable Importance

S.NO.	Variable	Score		
		MARS	CART	Random Forest
1	OGTT	100	100	100
2	DIAS_BP	64.23	9.7727	4.4312
3	AGE	59.89	52.9210	35.3079
4	BMI	49.61	39.1668	44.1314
5	DIA_PED_FN	35.30	6.6213	12.6091
6	SERUM_2HR	18.80	18.2706	7.3725
7	NOP	0	13.0335	10.0028
8	TRI_SFT	0	10.0928	4.3091

It is identified from the MARS model, there is an interaction between

1. Age of patient with Diabetes pedigree function, Diastolic blood pressure, Body mass index
2. Diastolic blood pressure and 2-Hour serum insulin
3. Diabetes pedigree function and Diastolic blood pressure
4. 2-Hour serum insulin and OGTT

The final equation for the MARS model for the prediction of diabetes in Pima Indians Diabetes Data Set selected the model with 13 basis functions is depicted in the Table IV. The final MARS model is given below

$$\left.\begin{aligned} Y = {} & 0.241013 + 0.00667071 * BF1 - 0.0155414 * BF3 + 0.0155959 * BF5 - 0.236775 * BF8 + \\ & 0.0613893 * BF9 - 0.000164229 * BF11 - 0.00945052 * BF12 - 5.46596e\text{-}007 * BF14 + \\ & 0.00555851 * BF15 + 1.86536e\text{-}008 * BF17 - 0.0313332 * BF19; \end{aligned}\right\} \quad (8)$$

Table IV
Basis Functions

BF1 = max(0, OGTT - 71); BF3 = max(0, AGE - 46); BF4 = max(0, 46 - AGE); BF5 = max(0, BMI - 20.8); BF7 = max(0, DIA_PED_FN - 1.258); BF8 = max(0, 1.258 - DIA_PED_FN); BF9 = max(0, DIA_PED_FN - 1.127) * BF4; BF11 = max(0, DIAS_BP + 7.62939e-006) * BF4; BF12 = max(0, BMI - 38.5) * BF3; BF14 = max(0, SERUM_2HR - 7.62939e-006) * BF11; BF15 = max(0, DIAS_BP - 72) * BF9; BF17 = max(0, OGTT - 153) * BF14; BF19 = max(0, DIAS_BP + 7.62939e-006) * BF7;

Various measures of predictive accuracy of the models were computed. Accuracy measures Average percent correct, Overall percent correct, Specificity, Sensitivity/Recall, Precision and F1 statistic of four methods are depicted in the Table V and represented in terms of the graph in Figure 2. Sensitivity plays an important role in the correct diagnosis of the disease. In this study, the best sensitivity achieved by Random forest with the value of 88.81 and the second best sensitivity is 81.72 by MARS. The overall percent correct and average percent correct of predicting diabetic status on the data set using MARS, CART, Random Forest and Logit Regression are (78.13and78. 96), (74.09 and 75.08), (70.05 and 74.40) and (76.56 and 76.03) respectively. One of the most surprising observation is that with 6 predictors MARS accuracy and sensitivity are higher than other method.

Table V
Classifier Accuracy Measures

S.No.	Algorithm	Specificity	Sensitivity/ Recall	Precision	F1 statistic	Average percent correct	Overall percent correct
1	MARS	76.20	81.72	64.79	72.28	78.96	78.13
2	CART	71.80	78.36	59.83	67.85	75.08	74.09
3	Random Forest	60.00	88.81	54.34	67.42	74.40	70.05
4	Logit-regression	77.80	74.25	64.19	68.86	76.03	76.56

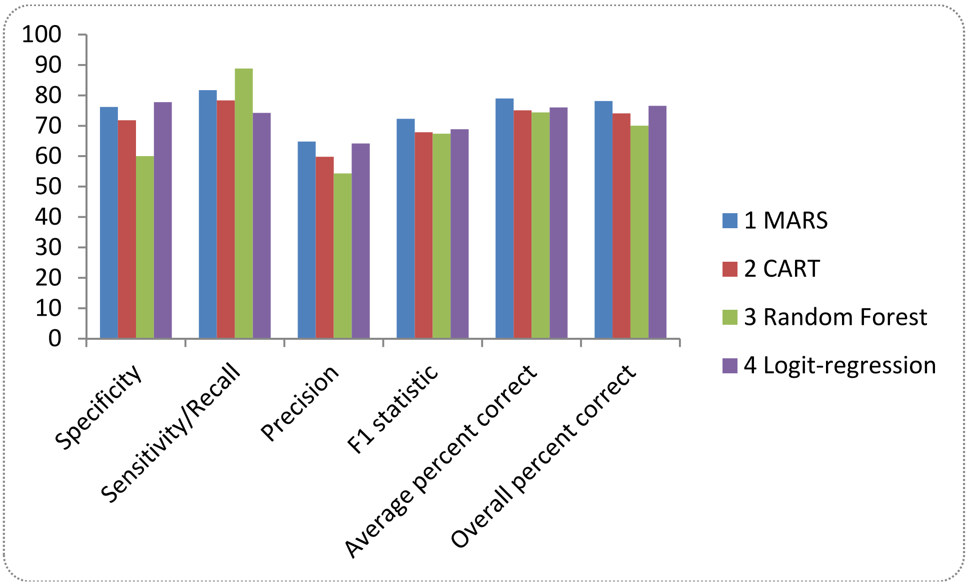

Figure 2. Comparison study graph

The "Receiver Operating Characteristic" (ROC) curve is one of the methods for examining the performance of data mining algorithms that classifies the subject positive or negative. ROC value always lies between [0,1]. ROC curve is a graphical representation of the relationship between both sensitivity and specificity and it helps to decide the optimal model through determining the best threshold for the diagnostic test [50]. The results of ROC of four method MARS, CART, Random Forest and Logit Regression are presented in the Table VI. The ROC curve of the above methods is presented in the Figure 3,4,5 and 6 respectively . From the ROC it indicates that MARS shows the best prediction performance among CART, Random Forest and Logit Regression.

Table VI
ROC Value - (Area Under Curve)

S.NO.	Method	ROC(Area under curve)
1	MARS	0.87376
2	CART	0.83255
3	Random Forest	0.82716
4	Logit Regression	0.83943

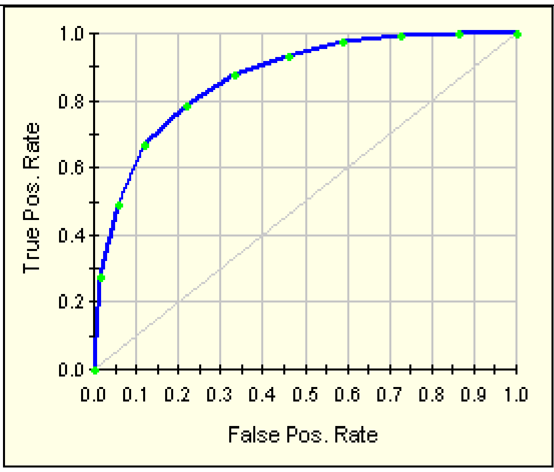

Figure 3. ROC –MARS

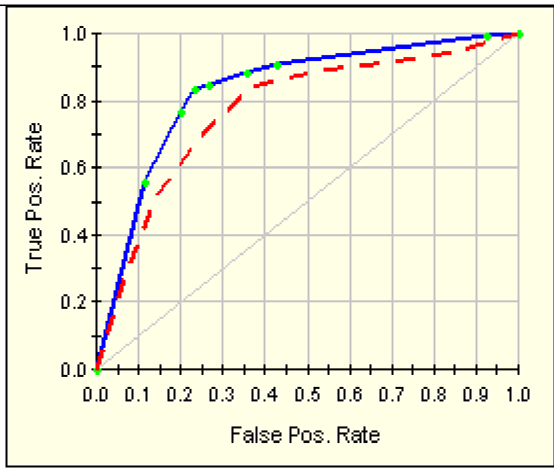

Figure 4. ROC- CART

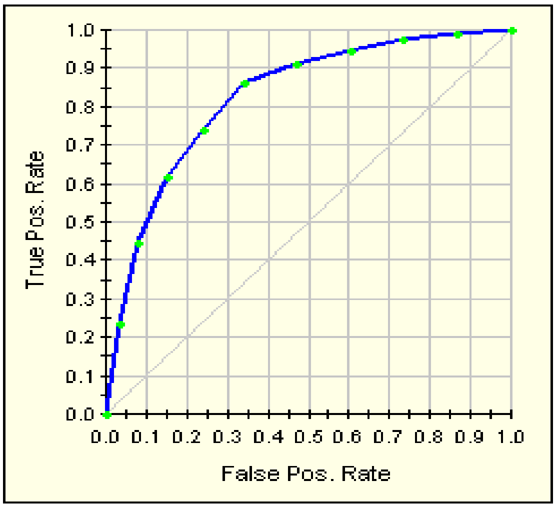

Figure 5. ROC-Random Forest

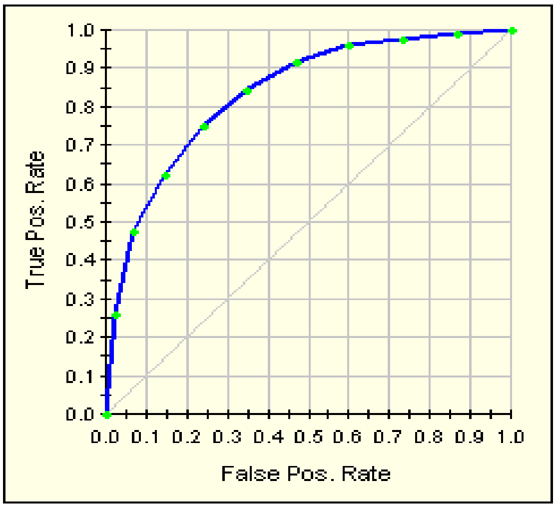

Figure 6. ROC- Logit Regression

5. CONCLUSION AND FUTURE WORK

In this research we have employed MARS for the diagnosis of diabetes. Model accuracy also evaluated using the area under Receiver Operating Characteristic curve (AUC). Four classification models- MARS, Logistic regression, CART, Random Forest are applied in this paper, MARS obtained a better accuracy and sensitivity with the minimum number of predictors and outperformed by handling nonlinearities, missing data and interactions among predictors compared to other methods. We can conclude that the proposed approach is easily understandable, provides a better and faster model for diagnosing of diabetes patients. Also very easy to develop a decision support system with the help of the proposed model. One of the potential future extensions of this work is to conduct a prospective study to further refine the predictive results obtained by the proposed rules by combining artificial intelligence techniques to develop an intelligent decision support system to help the physicians for better decision makings.

REFERENCES

[1] Giudici P, "Applied data mining, statistical methods for business and industry", Wiley & Sons, 2003
[2] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "Data mining and knowledge discovery in databases", Commun. ACM 39, pg.24–26, 1996
[3] Bellazzi R, Zupan B., "Predictive data mining in clinical medicine: current issues and guidelines", International Journal of Medical Informatics 77, pg.81-97, 2008
[4] Bellazzi R, Zupan B., "Towards knowledge-based gene expression data mining", Journal of Biomedical Informatics 40, pg.787–802, 2007
[5] Sellappan Palaniappan and Rafiah,"Intelligent, "Heart Disease Prediction System Using Data Mining Techniques", International Journal of Computer Science and Network Security, vol.8, No.8, pg.343-350, Aug 2008
[6] Medical decision support systems based on machine learning Chih-Lin Chi, University of Iowa, Iowa Research Online- Theses and Dissertations, pg.1-118, 2009
[7] Igor Kononenko, "Machine Learning for Medical Diagnosis: History, State of the Art and Perspective", University of Ljubljana, Artificial Intelligence in Medicine, pg.1-25, 2001

[8] R. Rupnik, M. Kukar, "Decision Support System to support decision process with data mining", Journal of information and organizational sciences, vol.31, No.1, pg:217-232, 2007
[9] Eta S. Berner andTonya J. La Lande, "Overview of Clinical Decision Support Systems", Health Informatics series-Second Edition, pg.3-22, 2007
[10] R Brian Haynes, Nancy L Wilczynski, "Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: Methods of a decision-maker researcher partnership systematic review", Implementation Science, 5:12, pg.1-8, 2010
[11] http://articles.timesofindia.indiatimes.com
[12] American Diabetes Association, "Standards of medical care in diabetes", Diabetes Care, vol.30, no.1, pg. S4–S41, 2007
[13] American Diabetes Association "Diagnosis and Classification of Diabetes Mellitus", Diabetes Care, vol.31, supplement 1, pg.55-60, Jan 2008
[14] "Insulin Resistance and Prediabetes", National Diabetes Information Clearinghouse; NIH Publication No.09–4893, pg.1-8, Oct 2008
[15] UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
[16] Hasan Temurtas, Nejat Yumusak, Feyzullah Temurta, "A comparative study on diabetes disease diagnosis using neural networks", Expert Systems with Applications, vol.36, Issue 4, pg. 8610-8615, May 2009
[17] Shankaracharya, Devang Odedra, Subir Samanta and Ambarish S. Vidyarthi, "Computational Intelligence in Early Diabetes Diagnosis: A Review", The Review of Diabetic Studies, vol.7, No. 4, pg 252-262, 2010
[18] G. Magudeeswaran , D. Suganyadevi, "Forecast of Diabetes using Modified Radial basis Functional Neural Networks", International Conference on Research Trends in Computer Technologies (ICRTCT), Proceedings published in International Journal of Computer Applications (IJCA) (0975 – 8887), pg.35-39, 2013
[19] Shelly Gupta, Dharminder Kumar and Anand Sharma, "Performance analysis of various data mining classification techniques on health care data", International journal of computer science & Information Technology(IJCSIT), vol.3, No.4, pg.155-169, Aug 2011
[20] Joseph L. Breault, MD, MPH, MS, "Data Mining Diabetic Databases: Are Rough Sets a Useful Addition?", Alton Ochsner Medical Foundation, pg.1-10
[21] P. Magni and R. Bellazzi, "A stochastic model to assess the variability of blood glucose time series in diabetic patients self-monitoring," IEEE Transactions on Biomedical Engineering, vol. 53, No.6, pg.977-985, Jun 2006
[22] K. Polat and S. Gunes, "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease", Journal of Digital Signal Processing, vol.17, Issue 4, pg. 702-710, July 2007
[23] S. Sahan, K. Polat, H. Kodaz, and S. Gunes, "The medical applications of attribute weighted artificial immune system (awais): Diagnosis of heart and diabetes diseases," in ICARIS, pg. 456 468, 2005
[24] T. Exarchos, D. Fotiadis, and M. Tsipouras, "Automated creation of transparent fuzzy models based on decision trees application to diabetes diagnosis," in 30th Annual International IEEE EMBS Conference, 2008.
[25] S. Purnami, A. Embong, J. Zain, and S. Rahayu, "A new smooth support vector machine and its applications in diabetes disease diagnosis," Journal of Computer Science, pg.1003-1008, 2009.
[26] T. Jayalakshmi and A. Santhakumaran, "A novel classification method for diagnosis of diabetes mellitus using artificial neural networks", International Conference on Data Storage and Data Engineering, pg.159-163, 2010
[27] T. McConoghy and G. Gielen, "Analysis of simulation-driven numerical performance modeling techniques for application to analog circuit optimization", Proc. IEEE International Symposium on Circuits and Systems, vol.2, pp. 1298–1301, May 2005
[28] Friedman, J.H., "Multivariate adaptive regression splines", The Annals of Statistics, 19, 1,pg. 1-67, 1991
[29] Taylan, P., Weber, G.W., and Yerlikaya, F., "Continuous optimization applied in MARS for modern applications in finance, science and technology", ISI Proceedings of 20th Mini-EURO Conference Continuous Optimization and Knowledge- Based Technologies, Neringa, Lithuania, pg.317-322, 2008
[30] H. Haas and G. Kubin, "A multi-band nonlinear oscillator model for speech", Conference Record of the Thirty-Second Asilomar Conference on Signals, Systems & Computers, vol.1, 1-4, pg. 338 – 342, Nov 1998
[31] S. M. Chou, T.-S. Lee, Y. E. Shao, and I.-F. Chen, "Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines," Expert Systems With Applications, vol. 27, no. 1, pp. 133–142, July 2004.
[32] T. McConoghy and G. Gielen, "Analysis of simulation-driven numerical performance modeling techniques for application to analog circuit optimization," IEEE International Symposium on Circuits and Systems, 23-26, vol.2, pg.1298 – 1301, May 2005
[33] Scott Crino, Member, and Donald E. Brown, "Global Optimization With Multivariate Adaptive Regression Splines", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol.37, Issue.2 , pg. 333 – 340, April 2007
[34] H. Zareipour, K. Bhattacharya, C.A. Canizares, "Forecasting the Hourly Ontario Energy Price by Multivariate Adaptive Regression Splines", IEEE on Power Engineering Society General Meeting, pg.1-7, 2006
[35] K. Cogger, "Modern regression methods: A comparative study", Proc.Washington Statistical Soc. Seminar, pp. 2–23, May 2001
[36] I. Uysal and H. A. Guvenir, "An overview of regression techniques for knowledge discovery", Knowl. Eng. Rev., vol.14, No.4, pg. 319–340, Dec 1999
[37] P. A. Lewis and J. G. Stevens, "Non-linear modeling of time series using multivariate adaptive regression splines (MARS)", J. Amer. Stat. Assoc., vol.86, No.416, pg. 864–877, 1991
[38] J. H. Friedman, "Multivariate adaptive regression splines (with discussion)", Annals of Statistics., vol.19, pg. 1–141, 1991
[39] J. H. Friedman and C. B. Roosen, "An introduction to multivariate adaptive regression splines", Stat. Methods Med. Res., vol. 4, pg. 197–217, 1995
[40] "Martian Chronicles: Is MARS better than Neural Networks?", Louise Francis, FCAS, MAAA, pg.269-306
[41] Fast MARS, Jerome H. Friedman –Technical Report No.110, pg.1-18, 1993
[42] Shankaracharya1, Devang Odedra, Subir Samanta,and Ambarish S. Vidyarthi, "Computational Intelligence in Early Diabetes Diagnosis: A Review", Computer-Based Diabetes Diagnosis The Review of Diabetic Studies, vol. 7, No. 4, pg 252-262, 2010
[43] P. Craven and G. Wahba, "Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation", Numer. Math., vol. 31, pg. 377–403, 1979
[44] "Nonlinear Regression: Modern Approaches and Applications", Salford System, 2013
[45] Steinberg, D., Colla, P. L., and, K. Martin, "MARS User Guide", San Diego, CA: Salford Systems, 1999
[46] Witten, Ian H., and Eibe Frank, "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations", San Diego, CA: Morgan Kaufmann, 2000
[47] Thomas M. Mitchell, "Machine Learning", New York McGraw-Hill, 1997
[48] Oracle®Data Mining Concepts 11g Release 1 (11.1)
[49] "Data Mining: Concepts and Techniques", Second Edition Jiawei Han: Micheline Kamber - Morgan Kaufmann Publishers
[50] Wen Zh, Nancy Zeng, Ning Wang, "Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS® Implementations", NESUG Health Care and Life Sciences, 2010
[51] www.en.wikipedia.org/wiki/Multivariate_adaptive_regression_splines