ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 4, April 2018

Diabetes Prediction using Linear Regression, Decision Tree & Least Square Support Vector Machine

Vaishali [1], Nisha Pandey [2]

M. Tech (Pursuing), Dept. of CSE, SRCEM, Palwal, Haryana, India [1]

Assistant Professor, Dept. of CSE, Dept. of CSE, SRCEM, Palwal, Haryana, India[2]

ABSTRACT: Machine learning algorithms will facilitate us to sight the onset endocrinology or diabetes disorder. Early detection of endocrinology disorder will cut back patient's health risk. Physicians, patients, and patient's relatives may be benefited from the prediction's outcomes. In low resource clinical settings, it's necessary to predict the patient's condition once the onus to portion resources suitably measured and preventions can be demonstrated or exercised. Many articles are revealed analyzing Prima Indian information set applying on numerous machine learning algorithms. However, under this scheme using Linear Regression and LS-SVM Classification techniques to predict the onset of diabetes on Prima Indian polygenic disorder dataset are demonstrated under this approach for such classification the confusion matrix and variance from Least Square Support Vector Machine is reliable approach and can forecast the unforeseen measures and symptoms foe endocrinology disorder. These techniques increase diagnosing accuracy and cut back medical bills and ensure the health living. During this study, the most focus is to analyze differing types of machine learning classification algorithms and show their amalgamated analysis. The aim of this study is to sight the diabetic patient's onset from the outcomes generated by machine learning classification algorithms.

KEYWORDS: Diabetes mellitus, Linear Regression, Decision Tree, Least Square Support Vector Machine.

I. INTRODUCTION

Diabetes mellitus (DM) be an unrelenting ailment, within which individual have elevated glucose measures. Which consequently, influences the capacity of human body to utilize optimum vitality found in sustenance for deep rooted. Once the body ingests straightforward sugar (sucrose) it ordinarily changes over the same into glucose in addition to it force go about as prime foundation energy for the human body. However, the glucose for the most part transports through circulatory systems as well as is in use up by cells. As in medical terms there are three categories of diabetes mellitus. Below the detailed are mentioned for ready reference:

a. Type 1 - At this point the human organ namely pancreas do not generate mandatory quantity of insulin and consequently the glucose measures inside the blood is more and high than the normal range. Human or person anguish commencing Type 1 diabetes mellitus is generally reliant on infusion of artificial developed human insulin.

b. Type 2 –At this occasion the human-cells of the person fall short to utilize the natural-insulin formed since of insulin-resistance.

c. Gestational diabetes – This transpires whilst to child-expectant or pregnant women which carry out not encompass the pre-diabetic or diabetes history will be originating diabetic amid elevated blood-sugar intensity.

Large amount glucose in blood can harm liver, brain, legs, eyes, kidneys, and nerves. It canister also foundation of heart attack, brain ham-rage or stroke, and scarcity in blood stream to legs or other internal organs. Overweight, lack of exercise, family history and stress increased the possible risk of diabetes. In Bangladesh, people are not conscious

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 4, April 2018

about health. There are 62.6 million cases of Diabetes in India. The increasing level of Diabetes is up bound. People do not know about it and they do not go to check it. Subsequently, the proposed technique will classify the diabetes with certain attributes and will forecast either the person in non-diabetic, acute diabetic or chronic diabetic based on three categories of Type1, Type2 and Gestational respectively.

.

II. RELATED WORK

1. Rian Budi Lukmanto etal This scheme proposed the submission of computational intellect by using fuzzy based logic system that executes the detection of diabetes mellitus (DM). This projected process is based on the comprehension achievement process. An accurateness of 87.46% is acquired by the proposed methodology.

2. Cheng-Hsiung Weng et al. In this scheme the various category of neural network classifiers are used for sickness forecast. Foremost we evaluate the presentation of solitary neural network with classifier with numerous neural networks with authentic data set. Secondly use arithmetical testing to scrutinize the divergence in presentation between these classifiers. Manifold neural network supposed to be improved than solitary neural network.

3**. Kamadi V.S.R.P. Varmaet al**. Under this scheme we emergent a decision tree representation to envisage the incidence of diabetes ailment. The large amount enhanced pronouncement regulations can be acknowledged from the data set with the use of the fuzzy conclusion precincts. The adapted Gini index-Gaussian fuzzy decision tree algorithm is projected. This algorithm outperforms supplementary decision tree algorithms to provide the effective and accurate results.

4. **Yue Huang** Under the scheme the authors discussed the characteristic assortment procedure, characteristic collection via supervised representation edifice was used to recognize the significant attribute affecting pancreas glucose control. After collection of appropriate features, three complementary classification techniques (Naïve Bayes, IB1 and C4.5) were practical implemented to the respective datasets which regulate to envisage how fine the state of patient was proscribed. The scheme identified patients' 'age', 'diagnosis duration', the need for 'insulin treatment', 'random blood glucose' measurement and 'diet treatment' as the most important factors which influence blood glucose control. With the usage of this technique, the best predictive accuracy of 95% and sensitivity of 98% was achieved. The factors, like 'type of care' delivered, the use of 'home monitoring', and the importance of 'smoking 'are not so much important in diabetes control. The more important factors that are identified include: 'age of patients', 'diagnosis duration' and 'family history', which are beyond the control of physicians.

5. **Polat and Gunes et.al**. under the proposed paper we confer using principal component analysis (PCA) and adaptive neuro-fuzzy inference system (ANFIS). Motive behind this study is to explore the expansion in the investigative accurateness of diabetes ailment combining PCA and ANFIS. The projected classification has two junctures. In the initial juncture, measurement of pancreas ailment dataset with the intention Eight characteristics, is decreased to four characteristics with the convention of primary constituent investigation. In the second juncture, conclusion of pancreas ailment is conceded out via adaptive neuro-fuzzy inference system classifier. The dataset used in our study is taken from the UCI (from Department of Information and Computer Science, University of California) Machine Learning Database.

III. PROPOSED METHODOLOGY

Under the scheme we proposes the diabetes forecast or prediction system in addition to responsiveness LS-SVM classification, regression and decision tree with determination and put into practice using classification based on machine learning algorithm (amalgamation logistic regression, decision tree and least squares support vector machines). It helps the user to be acquainted with or whether they are diabetic or non-diabetic. It also raises awareness

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 4, April 2018

amongst the patient and helps to maintain track of their health condition the proposed scheme is depicted in below Figure 1:

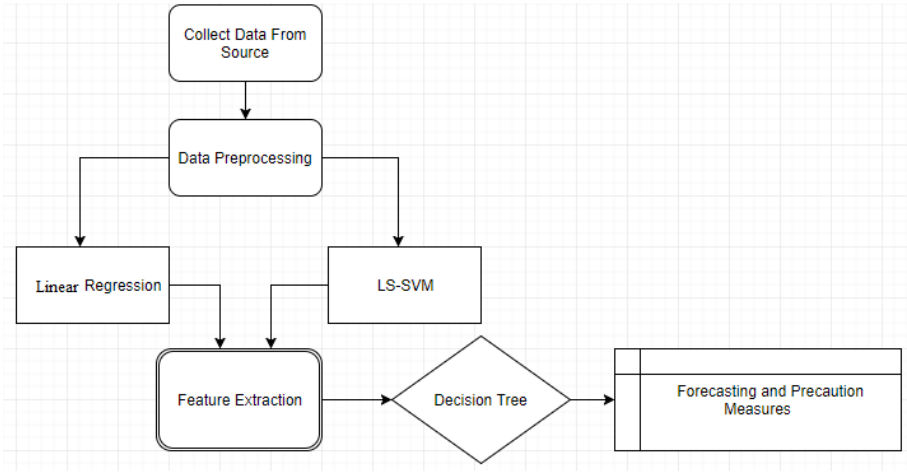


Figure 1: Proposed Scheme

Linear Regression : Linear Regression Technique is basically a Linear Approach which is used in order to model the relationship that exists between scalar dependent variables and also between variables more than pone termed as independent variable. Equation of Linear Regression Form expressed whereas The righteousness of fit character for the model calibrations are obtainable in below equation, and the calibrated coefficients are shown below. However, presents standard error (S_e) calculated as:-

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 4, April 2018

$$S_e = \sqrt{\frac{1}{n-m}\sum(y-\hat{y})^2}$$

where n is the number of observations,
m is the number of coefficients or exponents being calibrated,
y is the observed discharge (from the PeakFQ output), and
$\hat{y}$ is the predicted output calibrated by the regression tool.

Standard deviation (S_y) is calculated as

$$, = \sqrt{\frac{1}{n-1}\sum(y-\bar{y})^2},$$

where $\bar{y}$ is the mean of the discharges for the return period (T).

Explained variance (R^2) is calculated as

$$R^2 = \frac{1}{n^2 \cdot S_e^2 \cdot S_x^2}\left[\sum(y-\hat{y})\cdot(y-\bar{y})\right]^2$$

where

$$S_x = \sqrt{\frac{1}{n-1}\sum(\hat{y}-\bar{\hat{y}})^2},$$

in which $\bar{\hat{y}}$ is the mean of the predicted discharges for the return period

Least Square Support Vector Machine The Least Square Support Vector Machine is considered as an approximation tool in this is based on loss function is taken from the error variable. These modifications greatly simplified the problem and can be specifically described as follows:

$$\min J(w,e)=\frac{1}{2}w^T w+\gamma\frac{1}{2}\sum_{k=1}^{N}e_k^2$$

$$y_k=w^T\varphi(x_k)+b+e_k, \qquad k=1,\ldots,N$$

where e_k are error variables that play a similar role as the slack variables ξ_k in Vapnik SVM formulation and γ is a regularization parameter in determining the trade-off between minimizing the training errors and minimizing the model complexity. The Lagrangian corresponding to can be defined as:

$$L(w,b,e,\alpha)=J(w,e)-\sum_{K=1}^{N}\alpha_k\{w^T\varphi(x_k)+b+e_k-y_k\}$$

where $\alpha_k \in R$ are the Lagrange multipliers. The KKT optimality conditions for a solution can be obtained by partially differentiating with respect to w, b, e_k, and α_k

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 4, April 2018

$$\frac{\partial L}{\partial w}=0\rightarrow w=\sum_{k=1}^{N}\alpha_k\varphi(x_k)$$
$$\frac{\partial L}{\partial b}=0\rightarrow\sum_{k=1}^{N}\alpha_k=0$$
$$\frac{\partial L}{\partial e_k}=0\rightarrow\alpha_k=\gamma e_k,k=1,\ldots,N$$
$$\frac{\partial L}{\partial \alpha_k}=0\rightarrow wT\varphi(x_k)+b+e_k-y_k=0,\qquad k=1,\ldots,N$$

After elimination of the variable w and e_k, the following linear equation can be obtained:

$$\begin{bmatrix}b\\a\end{bmatrix}=\begin{bmatrix}0 & \vec{1}_N\\ \vec{1}_N & \Omega+\gamma^{-1}I_N\end{bmatrix}\begin{bmatrix}0\\y\end{bmatrix}$$

where $y=[y_1,\ldots,y_N]$, $\vec{1}_N=[1,\ldots,1]$ and $y=[\alpha_1,\ldots,\alpha_N]$. The kernel trick is applied here as follows

$$\Omega_{kl}=\varphi(x_k)^T\varphi(x_l)=K(x_k,x_l),\qquad k,l=1,\ldots,N$$

where $K(.,.)$ is the kernel function meeting Mercer's condition. b and α can be obtained by the solution to the linear system

$$b=\frac{\vec{1}_N(\Omega+\gamma^{-1}I_N)^{-1}y}{\vec{1}_N^T(\Omega+\gamma^{-1}I_N)^{-1}\vec{1}_N^T}$$

$$\alpha=(\Omega+\gamma^{-1}I_N)^{-1}(y-\vec{1}_N^T b)$$

Eventually, the resulting LS-SVR model for function estimation can be expressed as:

$$y(x)=\sum_{k=1}^{N}\alpha_k K(x_k,x_l)+b$$

Decision Tree:

Pocedure DecisionTreeClassify(e,X,Y,DT)
Inputs from LS-SVM and Linear Regression
set of input features, X={X1,...,Xn}
Y: target feature
e: example to classify
DT: decision tree
Prediction on Y for example e
Local: S sub branch of DT
S←DT
While S is an internal node of the
Form ⟨Xi=v,T1,T2⟩ do
If val(e,Xi)=v then

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

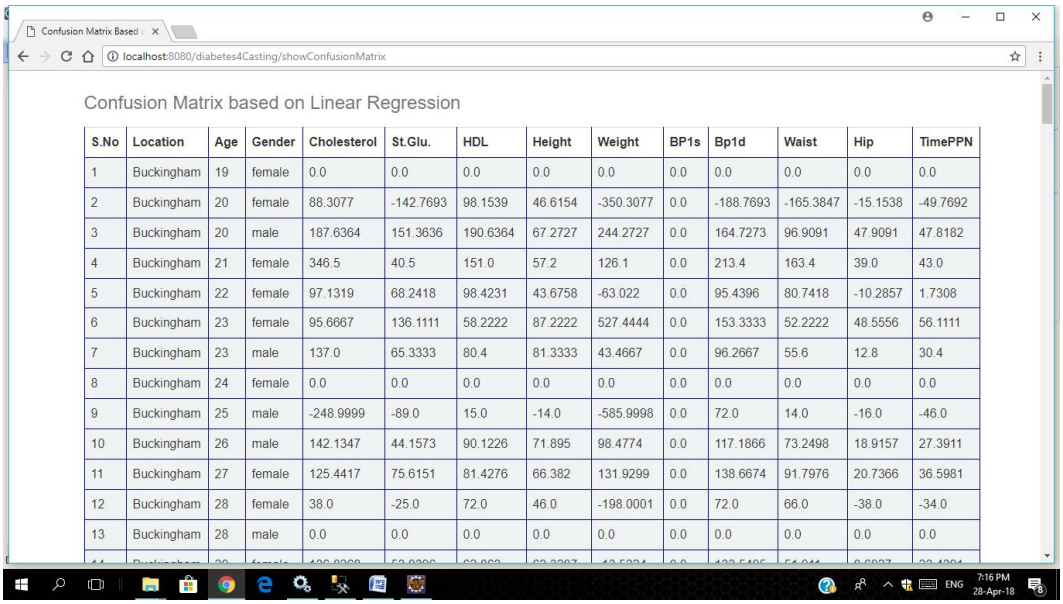
Website: www.ijircce.com

Vol. 6, Issue 4, April 2018

S←T1 else S←T2

IV. SIMULATION

Confusion Matrix pertaining set of predictors among which some are significant in terms of mean and variance than others where estimating the predictors if the same significant predictors before prediction, pose as significant after prediction generated through the linear regression bases on dataset characteristics or attributes i.e Cholesterol, Standard Glucose, HDL, Height, Weight, Blood Pressure and Waist Size.



S.No	Location	Age	Gender	Cholesterol	St.Glu.	HDL	Height	Weight	BP1s	Bp1d	Waist	Hip	TimePPN
1	Buckingham	19	female	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	Buckingham	20	female	88.3077	-142.7693	98.1539	46.6154	-350.3077	0.0	-188.7693	-165.3847	-15.1538	-49.7692
3	Buckingham	20	male	187.6364	151.3636	190.6364	67.2727	244.2727	0.0	164.7273	96.9091	47.9091	47.8182
4	Buckingham	21	female	346.5	40.5	151.0	57.2	126.1	0.0	213.4	163.4	39.0	43.0
5	Buckingham	22	female	97.1319	68.2418	98.4231	43.6758	-63.022	0.0	95.4396	80.7418	-10.2857	1.7308
6	Buckingham	23	female	95.6667	136.1111	58.2222	87.2222	527.4444	0.0	153.3333	52.2222	48.5556	56.1111
7	Buckingham	23	male	137.0	65.3333	80.4	81.3333	43.4667	0.0	96.2667	55.6	12.8	30.4
8	Buckingham	24	female	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	Buckingham	25	male	-248.9999	-89.0	15.0	-14.0	-585.9998	0.0	72.0	14.0	-16.0	-46.0
10	Buckingham	26	male	142.1347	44.1573	90.1226	71.895	98.4774	0.0	117.1866	73.2498	18.9157	27.3911
11	Buckingham	27	female	125.4417	75.6151	81.4276	66.382	131.9299	0.0	138.6674	91.7976	20.7366	36.5981
12	Buckingham	28	female	38.0	-25.0	72.0	46.0	-198.0001	0.0	72.0	66.0	-38.0	-34.0
13	Buckingham	28	male	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 1: Confusion Matrix produced From Linear Regression

LS-SVM with Decision Tree : *k* -fold cross validation method is used for performance evaluation of diabetes diagnosis using LS-SVM. *K-fold* cross validation is a way to improve over the holdout method. The data set is divided into *k* subsets, and method is repeated *k* times. As *k* increases, the variance of the resulting estimate reduces. The downside of this method is that the training algorithm must rerun *k* times from scratch, in other words, it takes *k* times computation to make an evaluation. To randomly divide the data into a test and training set *k* different times is a variant of this method. The advantage of this method is that you can independently choose how large you wish each test set to be and after how many trials you average should be over and decision tree is inculcated for decision support system,.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 4, April 2018

Gender	Age V	Chol V	Stab Glu V	HDL V	Height V	Weight V	BPLS V	BPID V	Waist V	Hip V	PPN V	Mean W
male	58.0	114.67	574.89	128.22	2.89	290.67	296.0	110.22	14.22	0.22	4850.0	104.06
Pre Diabetic - Take Precautions												
female	59.0	150.56	3324.56	123.84	3.76	294.64	40.64	121.6	7.36	6.8	9576.0	120.36
Pre Diabetic - Take Precautions												
male	59.0	1372.22	5304.67	297.56	8.67	68.22	6.22	48.22	0.67	1.56	140600.0	466.73
Chronic Diabetic Meidcal Emergency												
female	60.0	2675.43	7094.24	471.96	7.63	1158.53	322.12	97.63	29.06	22.12	10983.67	195.33
Pre Diabetic - Take Precautions												
male	60.0	1968.67	3544.67	240.89	1.56	190.89	242.67	155.56	6.22	1.56	36950.0	210.13
Acute Diabetic - Consult Doctor												

V. CONCLUSION AND FUTURE WORK

Using the above amalgamated technique the resultant values and derived with more accuracy where as the linear regression produced the confusion matrix thus resulting the compact the precise formation of weights based on characteristics and attributes where as LS-SVM classifies the estimation of probabilistic model where scheme can define the range in which the prediction can be made more perfectly based on type of categorization in diabetes and precautions respectively. For the future work the same the same can be implements on gigantic database vide hadoop where map and reduce will cut short the datasets in small proportions and parallel execution can be performed for quick and prompt results and parallel processing can also be levied for excellent production of information.

.

REFERENCES

1. Rian Budi Lukmanto, Irwansyah E et all “The Early Detection of Diabetes Mellitus (DM)Kamadi V.S.R.P. Varma Using Fuzzy Hierarchical Model.” ELSEVIER, Volume 59, 2015.
2. Cheng-Hsiung Weng, T et all “Disease prediction with different types of neural network classifiers.” ELSEVIER, Volume 33, 2014.
3. Kamadi V.S.R.P. Varma, A et all “A computational intelligence approach for a better diagnosis of diabetic patients.” ELSEVIER, Volume 40, 2014.
4. Bum Ju Lee, Boncho N et all “Prediction of Fasting Plasma Glucose Status Using Anthropometric Measures for Diagnosing of Diabetes.” IEEE, Volume 18, NO.2, 2014.
5. Longfei Han, Beijing T et all “Rule Extraction from Support Vector Machines Using Ensemble Learning Approach: An Application for Diagnosis of Diabetes.” IEEE, Volume 19, 2014
6 J.Pardeep Kandhasamy, S Balamurali “Perforamance analysis of Classifier Models to Predict Diabetes Mellitus.” ELSEVIER, Volume 47, 2014.
7 Ajith Abraham, Radha Thangaraj, Millie Pant, Pascal Bouvry, “Particle swarm optimization: Hybridization perspectives and experimental illustrations”, Applied Mathematics and Computation, 2011
8. Corne D, Dorigo M, Glover F, “New ideas in optimization”, McGraw-Hill, USA, 1999
9. D.Mishra, B.Sahu, “Feature selection for cancer classification: a signal-to-noise ratio approach”, International Journal of Scientific and Engineering Research, vol.2, 2011.
10. Expert Committee on the Diagnosis and Classification of Diabetes Mellitus, “Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus”. Diabetes Care, vol 20, pp. 1183–1197, 1997
11. http://en.wikipedia.org/wiki/Data_mining

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 4, April 2018

12. Carpenter, G. A., & Markuzon, "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases", Neural Networks, vol. 11, pp. 323–336, 1998.
13. Kemal Polat and S. Gunes, "An expert system approach based on principal component analysis and adaptive NeuroFuzzy inference system to diagnosis of diabetes disease", Digital Signal Processing, vol. 17, pp. 702-710, Jul 2007
14. Yue Huang, Paul McCullagh, Norman Black, Roy Harper, "Feature selection and classification model construction on type 2 diabetic patients", Volume 41 Issue 3, pp 251-262, Nov. 2007
15. K. Polat, S. Gunes and A. Aslan, "A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine", Expert Systems with Applications, vol. 34(1), pp. 214–221, 2008
16. T. Hasan, Y. Nejat, T. Feyzullah, "A comparative study on diabetes disease diagnosis using neural networks", Expert Systems with Applications, vol. 36, pp. 8610- 8615, May 2009
17. Santi Wulan Purnami, Abdullah Embong, Jasni Mohd Zain and S.P. Rahayu, "A New Smooth Support Vector Machine and Its Applications in Diabetes Disease Diagnosis", Journal of Computer Science vol.5 (12), pp.1006-1011, ISSN 1549-3636,2009
18. 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

1.