International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 5

Diagnosing Diabetes Using Support Vector Machine in Classification Techniques

T. Padma Nivethitha[1], A. Raynuka[1], Dr. J. G. R. Sathiaseelan[2]
[1]Research Scholar, [2]Head & Associate Professor
Department of Computer Science, Bishop Heber College, Trichy, Tamilnadu, India

ABSTRACT
Data mining is an iterative development inside which development is characterized by exposure, through either usual or manual strategies. In this paper, we proposed a model to ensure the issues in existing framework in applying data mining procedures specifically Classification and Clustering which are connected to analyze the type of diabetes and its significance level for each patient from the data gathered. It includes the illnesses plasma glucose at any rate held value. The research describes algorithmic discussion of Support vector machine (SVM), Multilayer perceptron (MLP), Rule based classification algorithm (JRIP), J48 algorithm and Random Forest. The result SVM algorithm best. The best outcomes are accomplished by utilizing Weka tools.

Keyword: *Data Mining, Diabetes, Classification, Clustering, SVM, MLP, JRIP, J48 and Random Forest, Weka.*

1. INTRODUCTION

The importance and Uses of Data Mining in Medicine inspite of the distinctions and conflict in approaches, the well being part has more requirements for data mining today. There are a considerable amount of contentions that could be refined to help the utilization of data mining in the well being part (Data overload, early discovery as well as shirking of ailments, Evidence-based medicine and avoidance of doctor's facility mistakes. Non-prominent finding and decision support, Policy-production in general wellbeing what's more, extra incentive for cash and value investment funds).

Weka contains an accumulation of representation instruments and algorithms for data investigation and prescient demonstrating, together with graphical UIs for simple access to these functions. Weka bolsters a few standard data mining errands, all the more particularly, data pre-processing, classification, clustering, regression and highlight determination. The majority of Weka's systems are predicated on the suspicion that the data is accessible as one level document or connection, where every datum point is depicted by a settled number of qualities (regularly, numeric or ostensible properties, however some other trait writes are likewise bolstered). Weka furnishes access to profound learning with deep learning. It isn't fit for multi-social data mining; however there is separate programming for changing over an accumulation of connected database tables into a solitary table that is reasonable for handling utilizing Weka. Another vital zone that is at present not secured by the algorithms incorporated into the Weka conveyance is succession demonstrating.

Diabetes is certainly not a recently born disease, it has been with human race from long back be that as it may, came to thought about it in 1552 B.C. Since this period, a considerable lot of Greek also French doctors had chipped away at it and made us mindful of the idea of illness, organs in charge of it and so on. In 1870s, a French doctor had find a connection amongst Diabetes and eating routine intake, and a thought to get ready individual eating regimen design.

Diabetic eating regimen was defined with consideration of milk, oats and other fiber containing sustenances in 1900-1915. Capacity of insulin, its

inclination, along with its utilization began from 1920 - 1923, found by Dr. Banting, Prof. Macleod and Dr .Collip, who were granted the Noble prize. In the time of 1940, it has been found that distinctive organs like kidney and skin are additionally influenced if diabetes is crawling for a long term.

The principle specialized goal in KDD advancement is to plan for Data Mining. Notwithstanding the development, it is additionally planned to address the procedure related issues. It is accepted that the execution of the Data Mining innovation would bargain out, memory and data requesting tasks as in restriction to one that require persistent cooperation with the database.

2. DATA ANALYSIS

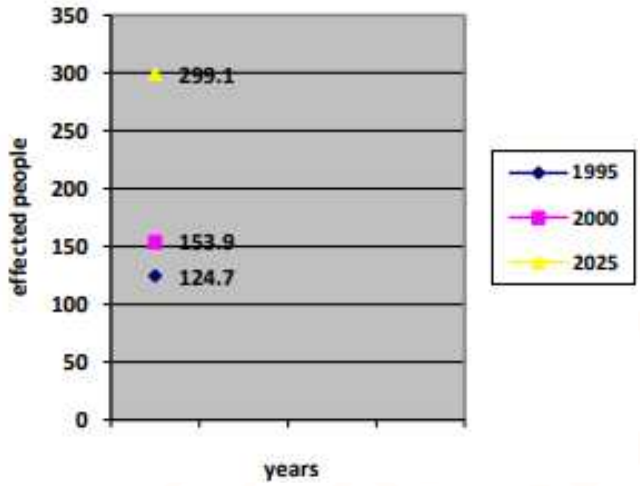
The most essential system use for this paper all through the examination of journals and publications in the field of medicine. The investigate concentrated on later publications. The information consider comprises of diabetes dataset. It incorporates name of the attributes and in addition the clarification of the attributes. The conclusion is 62.4 million individuals live with diabetes in India, and 77.2 million individuals are on the limit, with prediabetes. It calculated in anthropometric parameters like body weight, BMI (weight Index),height and weight limits and furthermore tried fasting glucose after glucose load(known diabetes exempted),and cholesterol for all member.

The events of pre-diabetes (disable fasting glucose as well as disable glucose resistance) was 8.3 percent, 12.8 percent, 8.1 percent, 14.6 percent correspondingly. Nineteen years to the lead of that due date, India has 62.4 million, and further 77.2 million (potential diabetes) in the pre-diabetes period. As indicated by the diabetes chart book of 2009, there were 50.8 million individuals with diabetes in India. Table.1 [1] demonstrates the diabetes effected and estimated details in India.

TABLE 1: INCREASING OCCURRENCE OF DIABETES: INDIA

Diabetes effected and estimated details in India	
Years	No of People effected (In Millions)
1995	124.7
2000	153.9
2025	299.1

Fig.1 [1] Number People Effected



The graph expresses the expansion pervasiveness of diabetes in India, estimate in 2025, 299.1 million impacts in diabetes. The event of diabetes in Tamil Nadu was 10.4 percent, in Maharashtra it was 8.4 percent, in Jharkhand 5.3 percent, and regarding rate, most noteworthy in Chandigarh at 13.6 percent.

Prediabetes is a condition when tolerant glucose level triggers higher than ordinary, yet not all that high that we can approve it as sort 2 diabetes. Gestational diabetes is a type of diabetes which influences pregnant women. It is suspected that the hormones made among pregnancy diminish a women receptivity to insulin, prompting high glucose levels. Gestational diabetes influences on 4% of every single pregnant women.

3. ALGORITHMS USED

3.1 Support Vector Machine (SVM):

Support vector machines are a respectably modern kind of learning algorithm, initially presented. Normally, SVM go for pointed for the hyper plane that most fantastic isolates the classes of data. SVMs have confirmed the capacity not exclusively to precisely isolate substances into revise classes, yet in addition to recognize case whose set up grouping isn't upheld by information. In spite of the fact that SVM are nearly obtuse characterize conveyance of preparing cases of each class.

SVM can be just reached out to perform numerical counts. Two such augmentations, the first is to broaden SVM to execute regression analysis, where the objective is to deliver a direct capacity that can genuinely exact that objective capacity. An additional expansion is to figure out how to rank components instead of creating an order for singular components. Positioning can be decreased to looking at sets of

example and creating a +1 appraise if the combine is in the right positioning request in expansion to −1 something else.

Kernel used:
Linear Kernel: K(x,y) = <x,y>

Classifier for classes: tested_negative, tested_positive

Binary SMO

Machine linear: showing attribute weights, not support vectors.

```
      1.3614 * (normalized) preg
 +    4.8764 * (normalized) plas
 +    -0.8118 * (normalized) pres
 +    -0.1158 * (normalized) skin
 +    -0.1776 * (normalized) insu
 +    3.0745 * (normalized) mass
 +    1.4242 * (normalized) pedi
 +    0.2601 * (normalized) age
 -    5.1761
```

Number of kernel evaluations: 19131 (69.279% cached)

3.2 Multilayer perceptron (MLP):

Multilayer perceptron neural systems (MLP) are the most regularly utilized provide for forward neural systems because of their fast operation, simplicity of usage, and set supplies. The MLPNN comprises of three consecutive layers: input, covered up and yield layers. The covered layer forms and transmits the info data to the yield layer. A MLP show with lacking or unreasonable number of neurons in the covered layer in all probability causes the issues of poor assumption and over fitting. There is no investigative technique for determining the number of neurons in the covered layer.

MLP
==========
Sigmoid Node 0

Inputs	Weights
Threshold	-2.722932253214729
Node 2	1.4723226659593067
Node 3	7.82043049951793
Node 4	2.443868722516107
Node 5	3.201530885671735
Node 6	-3.0281089464151147

Sigmoid Node 1

Inputs	Weights
Threshold	2.722932258743456
Node 2	-1.4723226671089364
Node 3	-7.820430653667116
Node 4	-2.4438687292638233
Node 5	-3.2015308924321246
Node 6	3.028108958788402

Sigmoid Node 2

Inputs	Weights
Threshold	-2.8783754488183835
Attribpreg	-9.079447868879274
Attribplas	-9.396273506987924
Attribpres	2.9422558149576776
Attrib skin	2.0758003770037323
Attribinsu	-8.023036419765178
Attrib mass	-10.734818857856807
Attribpedi	-3.3518466283880084
Attrib age	10.737145998129163

Sigmoid Node 3

Inputs	Weights
Threshold	-7.418945790100152
Attribpreg	1.150673574271605
Attribplas	-1.9072546608611365
Attribpres	3.403282312301751
Attrib skin	-7.795475278795782
Attribinsu	0.5709218287459333
Attrib mass	-6.176985830095681
Attribpedi	11.301811251887385
Attrib age	2.15894737853103

Sigmoid Node 4

Inputs	Weights
Threshold	0.3325483432081033
Attribpreg	0.9103698712206522
Attribplas	-13.824796952946144
Attribpres	-6.39936668126574
Attrib skin	3.3726419474851244
Attribinsu	-3.0955093615542073
Attrib mass	-8.775026662333607
Attribpedi	-5.192368159609418
Attrib age	9.269528793152958

Sigmoid Node 5

Inputs	Weights
Threshold	-3.376747660325633
Attribpreg	9.122015480585343
Attribplas	-12.642913808448162
Attribpres	5.67580447475827
Attrib skin	-0.0608550132015529
Attribinsu	2.307018507010625

Attrib mass -5.232080916356015
Attribpedi -0.7354842913650445
Attrib age -19.265633701797817

Sigmoid Node 6
Inputs Weights
Threshold 0.05437383478943048
Attribpreg 12.83676278178946
Attribplas -6.062276016682769
Attribpres -1.3896840458164514
Attrib skin 0.34997084802061346
Attribinsu -2.211940814726504
Attrib mass -0.9589656235930499
Attribpedi 6.090003751353257
Attrib age -8.833262465394691

Class tested_negative
Input
Node 0

Class testedpositive
Input
Node 1

3.3 Rule Based Classification Algorithm (JRIP):

JRIP is otherwise called Repeated Incremental Pruning to Produce Error Reduction (RIPPER). It is situated in affiliation rules with diminished blunder pruning (REP), an exceptionally normal and compelling procedure found in choice tree calculations. In REP for rules calculations, the preparation information is part into a developing set and a pruning set. Initial, an underlying tenet set is shaped that over ts the developing set, utilizing some heuristic technique. This overlarge run set is then more than once streamlined by applying one of an arrangement of pruning administrators run of the mill pruning administrators is erase any single condition or any single run the show. At each phase of disentanglement, the pruning administrator picked is the one that yields the best decrease of mistake on the pruning set. Rearrangements closes while applying any pruning administrator would expand mistake on the pruning set.

JRIP rules:
==========
(plas>= 132) and (mass >= 30) => class=tested_positive (182.0/48.0)

(age >= 29) and (insu>= 125) and (preg<= 3) => class=tested_positive (19.0/4.0)

(age >= 31) and (pedi>= 0.529) and (preg>= 8) and (mass >= 25.9) => class=tested_positive (22.0/5.0)

=> class=tested_negative (545.0/102.0)

Number of Rules: 4

3.4 J48 Algorithm:

J48 is an expansion of ID3. The extra highlights of J48 are representing missing qualities, decision trees pruning, persistent quality esteem ranges, deduction of standards, and so on. In the WEKA information mining device, J48 is an open source Java usage of the C4.5 algorithm. The WEKA device furnishes various decision related with tree pruning. In instance of potential over fitting pruning can be utilized as an tool for précising. In different algorithms the order is performed recursively till each and every leaf is unadulterated, that is the characterization of the information ought to be as flawless as could reasonably be expected. This algorithm it produces the principles from which specific personality of that data is produced. The goal is dynamically speculation of a decision tree until the point when it picks up balance of adaptability and precision.

J48 pruned tree
----
plas<= 127
| mass <= 26.4: tested_negative (132.0/3.0)
| mass > 26.4
| | age <= 28: tested_negative (180.0/22.0)
| | age > 28
| | | plas<= 99: tested_negative (55.0/10.0)
| | | plas> 99
| | | | pedi<= 0.561: tested_negative (84.0/34.0)
| | | | pedi> 0.561
| | | | | preg<= 6
| | | | | | age <= 30: tested_positive (4.0)
| | | | | | age > 30
| | | | | | | age <= 34: tested_negative (7.0/1.0)
| | | | | | | age > 34
| | | | | | | | mass <= 33.1: tested_positive (6.0)
| | | | | | | | mass > 33.1: tested_negative (4.0/1.0)
| | | | | preg> 6: tested_positive (13.0)
plas> 127
| mass <= 29.9
| | plas<= 145: tested_negative (41.0/6.0)
| | plas> 145
| | | age <= 25: tested_negative (4.0)
| | | age > 25
| | | | age <= 61

```
| | | | | mass <= 27.1: tested_positive (12.0/1.0)
| | | | | mass > 27.1
| | | | | | pres<= 82
| | | | | | | pedi<= 0.396: tested_positive (8.0/1.0)
| | | | | | | pedi> 0.396: tested_negative (3.0)
| | | | | | pres> 82: tested_negative (4.0)
| | | | age > 61: tested_negative (4.0)
| mass > 29.9
| | plas<= 157
| | | pres<= 61: tested_positive (15.0/1.0)
| | | pres> 61
| | | | age <= 30: tested_negative (40.0/13.0)
| | | | age > 30: tested_positive (60.0/17.0)
| | plas> 157: tested_positive (92.0/12.0)

Number of Leaves:  20
Size of the tree:  39
```

3.5 Random Forest:

Random Forest is an adaptable, simple to utilize machine learning algorithm. It is likewise a standout amongst the most utilized algorithm, since it's straight forwardness and the way that it can be utilized for both classification and regression tasks. Random woods, otherwise called irregular decision backwoods, are a main stream group technique that can be utilized to manufacture prescient models for both classification and regression issues. Ensemble strategies utilize numerous learning models to increase better prescient outcomes on account of an random forest, the model makes a whole backwoods of irregular uncorrelated decision trees to touch base at the most ideal answer.

```
Random Forest
==========
Bagging with 100 iterations and base learner
```

4. RESEARCH FINDINGS

4.1 Data mining in the diabetes disease Prediction:

Five different supervised classification algorithms i.e. SVM, MLP, JRIP, J48, Random Forest have been used analyze dataset in. Weka tool has a few standard data mining tasks, all the more particularly, data pre-processing, classification, clustering, regression and feature selection.

4.2 Data source:

The main objective is to conjecture if the patient has been influenced by diabetes utilizing the data mining devices by utilizing the therapeutic data accessible. The characterization kind of data mining has been connected to the Pima Indians diabetes dataset. Table 2 shows a brief description of the dataset that is being considered.

TABLE 2: DATASET DESCRIPTION

Dataset	No of attributes	No of instances
Pima Indians Diabetes Dataset	9	768

The attribute descriptions are shown in table3.

TABLE 3: ATTRIBUTES OF DIABETES DATASET

S. No	Name	Description
1	Preg	No of times pregnant
2	Plas	Plasma glucose concentration a 2 hours in a oral glucose tolerance test
3	Pres	Diastolic blood pressure (mm hg)
4	Skin	Triceps skin fold thickness (mm)
5	Insu	2 hours serum insulin (mm u/ml)
6	Mass	Body mass index (weight in kg/ (height in m)^2)
7	Pedi	Diabetes Pedigree Function
8	Age	Age (years)
9	Class	Class Variable (0 or 1)

4.3 Performance study of algorithms:

The table 4 consists of values of different classification. According to these values the lowest computing time (0.01s) can be determined.

TABLE 4: PERFORMANCE STUDY OF ALGORITHM

Algorithm Used	Time Taken	Accuracy %	Positive Recall	Error Rate
SVM	0.01	79.31	0.554	0.2069
MLP	0.95	74.32	0.675	0.3186
JRIP	0.11	77.01	0.663	0.3579
J48	0.15	76.25	0.566	0.3125
Random Forest	0.23	78.54	0.602	0.3046

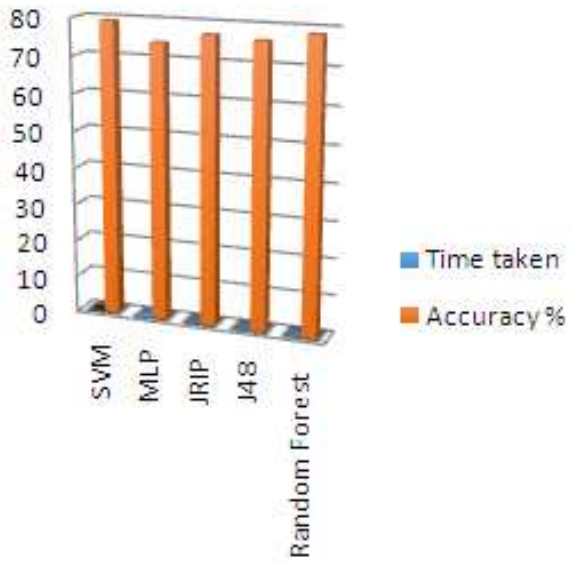
SVM, JRIP, J48 and Random Forest in a lowest computing time that we have experimented with a dataset. A distinguished confusion matrix was obtained to calculate sensitivity, specificity and accuracy [1]. Confusion matrix is a matrix representation of the classification results [1]. Table 5[1] shows confusion matrix.

TABLE 5: CONFUSION MATRIX

	Classified as Healthy	Classified as not Healthy
Actual Healthy	TP	FN
Actual not Healthy	FP	TN

The table 4 consists of values of different classification. According to these values the accuracy was calculated. The figure 2 represents the resultant values of above classified dataset using data mining supervised classification algorithms and it shows the highest accuracy [1]. It is logical from chart that compared on basis of performance and computing time.

Fig. 2 Predicted Accuracy



5. CONCLUSION:

In this paper five classification techniques in data mining to predict diabetes disease in patients. They names are: SVM, MLP, JRIP, J48 and Random Forest. These techniques are compared by using disease among patients. In this paper five classification validation error rate (True Positive, True Negative, False Positive and False Negative) and Accuracy. Our studies first filtered five algorithms based on lowest computing time SVM, JRIP, J48, Random Forest and MLP. The second one was highest accuracy above 78.54%. The SVM algorithm best among five with highest accuracy of 79.31%. In future to improve the performance of these classification.

REFERENCES

1. Karthikeyani, Parvin Begum, Tajudin, Shahina Begam," Comparative of Data Mining Classification Algorithm (CDMCA) in Diabetes Disease Prediction" International Journal of Computer Applications (0975 – 8887) Volume60– No.12, December 2012.

2. Suresh Kumar, Umatejaswi," Diagnosing Diabetes using Data Mining Techniques" International Journal of Scientific and Research Publications, Volume 7, Issue 6, June 2017 705 ISSN 2250-3153.

3. Aiswarya Iyer, Jeyalatha, Ronak Sumbaly," Diagnosis of Diabetes using Classification Mining techniques" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.

4. S. Kumari and A. Singh, "A Data Mining Approach for the Diagnosis of Diabetes Mellitus", Proceedings of Seventh international Conference on Intelligent Systems and Control, 2013, pp. 373375.

5. Neeraj Bhargava, Girja Sharma, Ritu Bhargava and Manish Mathuria," Decision Tree Analysis on J48 Algorithm for Data Mining". Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013.

6. Sankaranarayanan. S and Dr Pramananda Perumal. T, "Predictive Approach for Diabetes Mellitus Disease through Data Mining Technologies", World Congress on Computing and Communication Technologies, 2014, pp. 231-233.

7. Rajesh K, Sangeetha V. "Application of data mining methods and techniques for diabetes diagnosis", International Journal of Engineering and Innovative Technology (IJEIT). 2012; 2(3):224–9.

8. Ananthapadmanaban K R, Parthiban G. "Prediction of chances - diabetic retinopathy using data mining classification techniques" Indian Journal of Science and Technology. 2014 Oct;

7(10):1498–503.

9. Patil B M, Joshi R C, Toshniwal D. “Association rule for classification of type2 diabetic patients” 2nd International Conference of IEEE on Machine Learning and Computing; 2010. p. 67. DOI 10.1109/ICMLC.
10. Jali M V, Hiremath M B.” Diabetes” Indian Journal of Science and Technology. 2010 Oct; 3(10).
11. Lanord M, Stanley J, Elantamilan D, Kumaravel TS. Prevalence of Prehypertension and its Correlation with Indian Diabetic Risk Score in Rural Population. Indian Journal of Science and Technology. 2014 Oct; 7(10):1498–503.
12. Joseph L. Breault., “Data Mining Diabetic Databases: Are Rough Sets a Useful Addition?” Jamia Hamdard University, New Delhi, Proceedings of the 4th National Conference, INDIA Com-2010 Computing for Nation Development, February 25-26, 2010 Bharati Vidyapeeth’s Institute of Computer Applications and Management, New Delhi.
13. Chau , Shin, “A Comparative study of Medical Data classification Methods Based on Decision Tree and Bagging algorithms”, Proceedings of IEEE International Conference on Dependable, Autonomic and Secure Computing 2009, pp.183-187.
14. Jiawei Han and Micheline Kamber, “Data Mining Concepts and Techniques”, second edition, Morgan Kaufmann Publishers an imprint of Elsevier.
15. Dayle, L., Sampson, Tony J., Parker, Zee Upton, Cameron, P., Hurst, September, 2011. “Comparison of Methods for Classifying Clinical Samples Based on Proteomics Data: A Case Study for Statistical and Machine and SIMCA classification, Journal of Chemo metrics”, 20(8–10), 341–351.
16. Palaniappan, S., Awang, R., “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, Proceedings of IEEE/ACS International Conference on Computer Systems and Applications 2008,pp.108-115.
17. Chau, M., Shin, D., “A Comparative study of Medical Data classification Methods Based on Decision Tree and Bagging algorithms”, Proceedings of IEEE International Conference on Dependable, Autonomic and Secure Computing 2009, pp.183-187.