International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

A Novel Integrated Type 2 Diabetes Prediction Model for Indian Population using Data Mining Techniques

Omprakash S. Chandrakar, Jatinder kumar R. Saini

Abstract: ***Late diagnosis and undiagnosed type 2 diabetes are the two major concerns for India, which is going to be a diabetes capital shortly. Several diabetes risk score (DRS) tools have been proposed and deployed for detecting the persons with high risk. These DRS tools have been developed using the multiple logistic regression model. But this model is both imperfect and subject to misuse. Another major issue with the DRS tools developed for Indian population is that they are based on the very limited urban population that does not represent the population of India. The objective of current research work is to develop a classification model for type 2 diabetes prediction. Along with this, the building of a novel integrated model for type 2 diabetes risk prediction is discussed consisting of the aggregate classification model and Indian weighted diabetes risk score model. The dataset used to develop and validate the model is obtained from the Annual Health Survey comprising of nearly 0.7 million and nearly 75 thousand adult participants respectively from around 400 districts of India. The proposed integrated diabetes risk prediction model predicts diabetes with 69.89% sensitivity, 56.58% specificity. The positive predictive value of the proposed integrated model is 15.88%, which is a significant improvement as the prevalence of diabetes is only 3.68% for the study population. Developing countries such as India, where undiagnosed diabetes and limited financial resources are a significant concern, the proposed integrated model for diabetes risk prediction can be useful as a cheaper tool useful for mass-screening, which can save up to 30% of the total screening cost.***

Keywords : Indian Weighed Diabetes Risk Score; Aggregate Classification Model; Feature Selection; Semantic Discretization, Diabetes Mass Screening Test.

I. INTRODUCTION AND RELATED LITERATURE

World Health Organization (WHO)'s latest report on world diabetes states that the count of diabetic people, in 2014, has noted an increase of 314 million in addition to 108 million in the year 1980. There were 1.6 million of deaths for which diabetes was responsible worldwide in 2016, hence becoming the seventh major cause of death. According to another estimate by WHO, in 2012, more than 2.2 million deaths owed to high levels of blood glucose [1]. Indian scenario is even worse. According to WHO estimation, there were 69.2 million people with diabetes in 2015, and out of these, 36 million people remained undiagnosed [2]. Diabetes risk score (DRS) tools can be used as a cost effective tool for the mass screening test in detecting people with high risk for diabetes. Three significant research works on Indian DRS tools carried out by Mohan et al. [3], Ramachandran et al. [4], and Chaturvedi et al. [5]. The summary of the findings of these studies is presented in Table I.

Table-I: Comparative Study of Three Diabetes Risk Scores for Indian Population

Method & Researcher	Sample size & Location	Prevalence [%]	Year	Sensitivity [%]	Specificity [%]	Population at high risk [%]
Multiple Logistic Regression [3]	2350, Chennai	15.5	2001 to 2003	72.5	60.1	42.9
Multiple Logistic Regression [5]	4044, Delhi	10.88	1991 to 1994	79	56	Not disclosed by author
Multiple Logistic Regression [4]	4993, **Six Metro Cities	5.14	2005	76.6	59.9	38

**Six Metro cities: Hyderabad, Chennai, Bengaluru, Kolkata, New Delhi, and Mumbai

Following three issues identified in the above three research studies motivated us to develop a new diabetes risk prediction model for the population of India.

1. The DRS tools developed with very small population size, and also there is a lack of diverse demographic participants. All the participants belong to an urban area, where the prevalence of diabetes was quite high than in the rural area. A DRS tool, used for a particular set of people may not be effectively useful for predicting diabetes in other population.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
Omprakash S. Chandrakar*, Associate Professor, Uka Tarsadia University, Bardoli, Gujarat, India.
Jatinderkumar R. Saini, Professor, Symbiosis Institute of Computer Studies and Research, Symbiosis International Deemed University, Pune, Maharashtra, India.

Rotterdam predictive model gives better accuracy when it is tested with the Rotterdam population in comparison to the population such as Denmark, Spain, USA, etc. [6, 7]. India is a country of diversity. So the DRS tools derived for the urban population may not be effectively applied for the Indian population at large.

2. Koopman et al. [8] have studied how the average age of diagnosis of type 2 diabetes mellitus in the USA is changing. Based on the data obtained from the two National Health and Nutrition Examination Surveys in the year 1988–1994 and 1999–2000, they observed that mean age at diagnosis for type 2 diabetes gets decreased to 46 years from 52 years in US population. The three studies presented in Table I were conducted 15 to 20 years back. So the models developed for predicting diabetes based on these studies may not be applied accurately on the present population.
3. Logistic regression analysis is used to develop risk score tools for various disease or diagnostic algorithms. Anderson et al. [9] have observed that the logistic regression model is being used extensively in the clinical medicine is imperfect and prone to misuse. Lee J. et al. [10] found that building a logistic regression model is not a fixed exercise. It cannot always be reproduced. They have observed that different researchers derived different risk model with the same dataset. Logistic methods suffer from overfitting. Generalizability of such models is limited. In the over-fitted model, low risks are under estimated while the high risks are estimated overly. Another major issue with the logistic model is its assumption of linear probability, as shown in the following equation. This assumption may not be valid for all risk factors.

Patient's risk of disease is defined as exp (PRS) ÷ (1+ exp (PRS)). Here, the PRS for the disease = intercept + (βAge × Age) + (βBMI × BMI) + (βBP_S×BP_S) + (βBP_D×BP_D) +(βPulse_Rate×Pulse_Rate)+(βRural_Urban×Rural_Urban)) Here, βAge, βBMI, βBP_S, βBP_D, βPulse_Rate and βRural_Urban are all of the regression coefficients. They describe the effect of patient's values on the risk. PRS stands for Patient's Risk Score.

Additionally, the related literature shows that the researchers suggested the best technique for classification [18], IWDRS [19], validation of IWDRS [20] and semantic discretization [21] for the type 2 diabetes. This paper addresses the above challenges and proposes a solution. To resolve issue 1 and 2, a comprehensive dataset is used to build prediction models. The third issue is addressed by offering an integrated model consists of Indian weighted diabetes risk score model and aggregate classification model. The remainder of the paper is structured as follows. The components of the proposed integrated diabetes prediction model are discussed with an architectural diagram in the research methodology section. The comprehensive dataset, along with the process of the building model, is presented in the experiment section. The paper is concluded with the analysis of experimental results and drawing a conclusion from them.

II. RESEARCH METHODOLOGY: THE PROPOSED INTEGRATED MODEL FOR DIABETES RISK PREDICTION

Two distinct methods have been proposed by the researchers for diabetes risk prediction. One is based on the Indian weighted DRS tool, and another is an aggregate classification model [6]. This paper presents an integrated approach to building a reliable and robust model, which is depicted in Fig. 1. The proposed integrated model consists of 4 components, which will be described now.

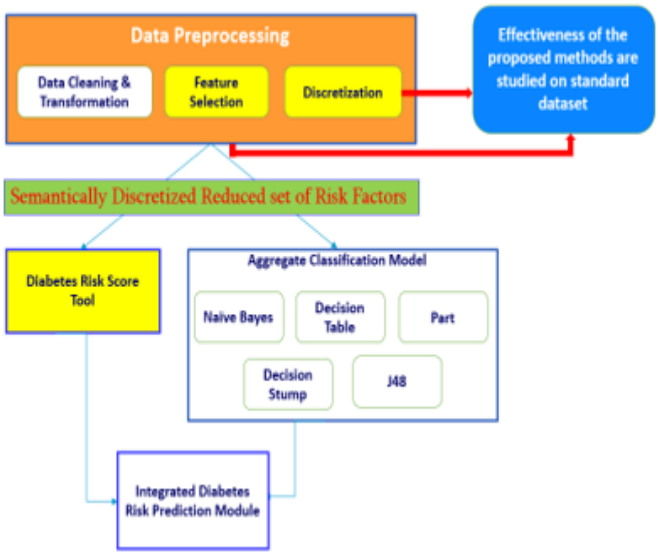


Fig. 1. Block diagram of the proposed integrated diabetes risk prediction model

A. Data Processing

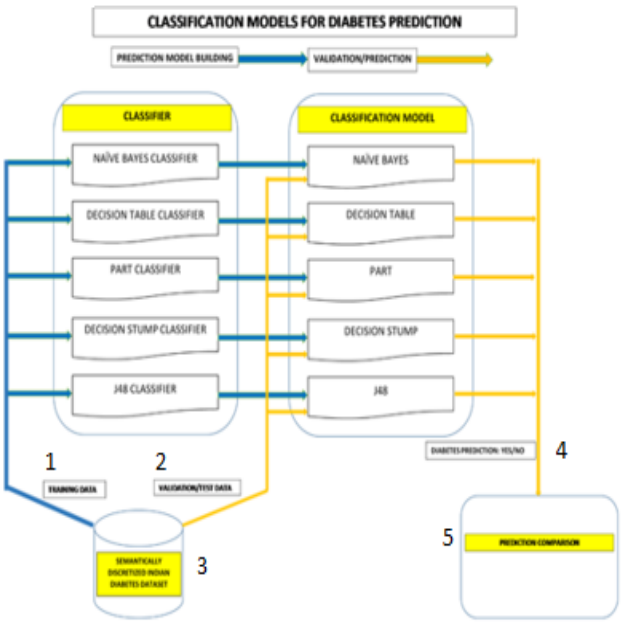
The data processing component produces the semantically discretized and a reduced set of risk factors, based on which diabetes risk score tool and classification model are derived [16]. The component uses a novel Majority Vote Based Iterative Feature Selection method for finding a reduced set of risk factors. The data is further discretized using the novel semantic discretization method to improve the accuracy and efficiency of data mining algorithms used in the next two components. The impact of these two methods has been studied on Pima Indian Diabetes dataset, which is taken from the UCI Machine Learning Repository and frequently used to evaluate the performance of classification models. An average improvement of 2.05% (range 0.13% to 4.17%) in classification accuracy is observed after applying the above two methods [16-17].

B. Diabetes Risk Score Tool

This component uses a novel method of calculating diabetes risk score. It uses semantically discretized dataset to improve the prediction model. This component produces Indian Weighted Diabetes Risk Score (IWDRS) tool [6].

C. Aggregate Classification Model

Five diabetes classification models are built using the semantically discretized reduced dataset. Naive Bayes, Decision Table, Decision Stump, PART, and J48 classifiers are used for building model. The result of all five models is given to an aggregation module. The aggregation module produces the final prediction based on the prediction of the majority of the five classification models. The Fig. 2 depicts the architecture of the aggregate classification model.



1: Training Data
2: Validation/Test Data
3: Semantically Discretized Indian Diabetes Dataset
4: Diabetes Prediction (Yes/No)
5: Prediction Comparison

Fig. 2. The architecture of aggregate classification model for diabetes prediction

D. Integrated Diabetes Risk Prediction Model

The aggregate classification model and the IWDRS models are integrated using a consultation module. The consultation module receives diabetes prediction from both the models and generates a final prediction based on the decision table shown in Table II.

Table-II: Decision Table for Integrated Diabetes Risk Prediction Model

Aggregate Classification Model	Risk Score Model	Final Prediction
No	Low	Low
No	High	Moderate
Yes	Low	Moderate
Yes	High	High

III. EXPERIMENT AND RESULTS

A. Data Source

Data for Indian diabetes dataset is taken from the Annual Health Survey (AHS) [11-12] which was conceived by Dr. Manmohan Singh, the then prime minister of India. The Office of the Reg. Gnrl., India has been assigned the responsibility for the project looking to its experience and expertise in handling such survey. The survey was carried out across 9 States, where half the population of India resides, namely, Odisha, Assam, Madhya Pradesh, Chhattisgarh, Bihar, Jharkhand, Rajasthan, Uttarakhand and Uttar Pradesh. We used AHS CAB 2014 dataset [13], which was released in April 2016 and downloaded from the data repository of Government of India [14].

B. Clinical, Anthropometric and Bio-chemical (CAB) Survey

A special biomarker factor was used as a supplement to AHS survey for collection of data pertinent to Empowered Action Group (EAG) States and the state of Assam. The Clinical, Anthropometric, and Bio-chemical (CAB) Survey [11-14] is meticulously designed to bridge the gaps of data on lifestyle diseases like hypertension, diabetes and anemia, as well as nutritional status.

C. Sample Size and Sample Unit

The Survey has taken into 384 districts of the 9 states of India. A total of as high as 0.34 million of house-holds as well as 1.65 million of people were covered under the survey. Census Enumeration Blocks (CEBs) of 2011 census in urban areas and villages in rural areas are used as sample units [11-14].

D. Quality Control Mechanism

Due care was taken to ensure non-dilution of quality. In addition to training, the manuals with instructions were used for the process [13].

E. Finding Semantically Discretized Reduces Set of Risk Factors

After going through the research literature and consultation with domain experts, 8 parameters have been selected as initial risk factors for diabetes. They are Pulse Rate, Gender, Age, Hemoglobin (Hb), BMI, BP Systolic and Diastolic and Residential Area (Rural/Urban). After applying the Majority Vote Based Iterative Feature Selection Algorithm, 6 parameters have been selected as a reduced set of risk factors. They are Pulse Rate, BMI, Age, BP Systolic and Diastolic and Residential Area. Further, the semantically discretized data set is obtained by applying semantic discretization method.

F. Derivation of Indian Weighted DRS

The risk scores have been derived for all the six risk factors [6], which is shown in Table III. The total IWDRS of a person may vary from 0 to 79, which indicates the lowest and highest diabetes risk, respectively. Positive Predictive Value (PPV), Sensitivity, Negative Predictive Values (NPV), Specificity, and Accuracy for predicting diabetes are found through calculations at the optimal cut-off scores (TWDRS>=24). Table IV shows the performance statistics for the training dataset and validation dataset.

G. Building Classification Models

Classification models are build using 5 different classifiers, namely, Naïve Bayes, Decision Table, Decision Stump, PART, and J48. The Table V shows the evaluation parameters of each classification models. Weka has been used for carrying out data mining tasks [15]. The data represented in Table V is presented pictorially through Fig. 3.

H. Aggregation of Classification Models

The aggregate classification module receives class (Diabetes =
Yes or No) from each of five classification models and produces a final result according to the classification result given by three or more classification models. The following Table VI shows the results of the aggregate classification model.

Table-III: Indian Weighted DRS (IWDRS) [6]

Risk Factors	Criteria	Nominal Value	Risk Score
Age	<=28	Low	0
	<=43	Moderate	3
	<=58	High	13
	>58	Extreme High	26
Blood Pressure Systolic	<=109	Low	0
	<=124	Moderate	3
	<=145	High	14
	>145	Extreme High	18
Blood Pressure Diastolic	<=68	Low	0
	<=80	Moderate	1
	<=93	High	3
	>93	Extreme High	11
BMI	<=19	Low	0
	<=22	Moderate	0
	<=25	High	2
	>25	Extreme High	10
Pulse Rate	<=73	Low	0
	<=82	Moderate	1
	<=93	High	4
	>93	Extreme High	10
Rural_Urban	= Rural	Low	0
	=Urban	High	4

Table-IV: Performance statistics at optimal cut off score [6]

Dataset	Proportion [%]	Sensitivity [%]	Specificity [%]	Accuracy [%]	PPV [%]	NPP [%]
Training	41.82	73.29	59.39	59.9	6.47	98.31
Validation	41.85	72.43	59.32	59.8	6.37	98.25

Table-V: Diabetes Prediction Result for Classification Model

Classifier	Sensitivity [%]	Specificity [%]	Accuracy [%]	PPV [%]	NPP [%]
Naïve Bayes	61.72	69.48	69.19	7.18	97.94
Decision	66.11	66.72	66.69	7.06	98.09
Table					
Part	65.56	67.67	67.6	7.2	98.09
Decision Stump	90.02	29.41	31.64	4.65	98.72
J48	67.46	65.85	65.91	7.02	98.15

I. Integrated Diabetes Risk Prediction Model

The DRS and aggregate classification model are integrated, and it produces the final prediction based on the decision table shown in Table II. Diabetes Prediction result for Integrated Diabetes Risk Prediction Model is shown in Table VII.

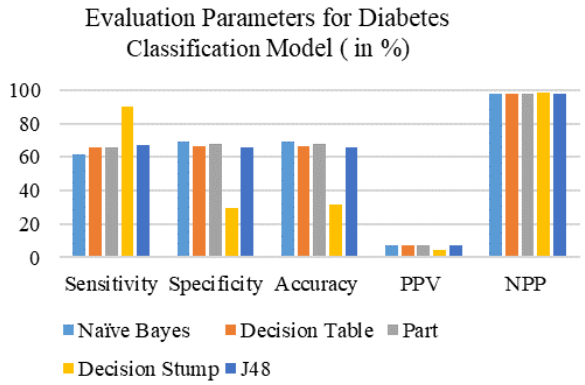


Fig. 3. Prediction evaluation parameter for various classifiers

Table-VI: Performance of Aggregate Classification Model

Sensitivity [%]	Specificity [%]	Accuracy [%]	PPV [%]	NPP [%]
68.3	64.93	65.06	6.93	98.17

Table-VII: Performance of Integrated Diabetes Risk Prediction Model

Sensitivity [%]	Specificity [%]	Accuracy [%]	PPV [%]	MPP [%]	NPV [%]
69.89	56.58	57.79	15.88	9.09	94.94

IV. RESULT ANALYSIS

This paper discussed the development of a novel integrated model for diabetes risk prediction consisting of Indian weighted diabetes risk score and aggregate classification model. Dataset used to develop and validate the model is taken from Annual Health Survey data, a very comprehensive dataset representing the population 384 districts of India. This made the model reliable and robust, which can be used to predict the diabetes risk of any Indian. Performance of the proposed integrated model is summarized along with two of its constituent models and three other significant DRS models in Table VIII. Moreover, it is evident from Fig. 4 as well as Fig. 5 that PPV is good for the integrated model.

Retrieval Number: B4220129219/2019©BEIESP
DOI: 10.35940/ijeat.B4220.129219
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

V. CONCLUSION

Indian weighted diabetes risk score model predicts the diabetes person with 72.43% sensitivity, 59.32% specificity. 41.85% population has been kept under the high-risk category by this model. The aggregate classification model predicts with 68.3% sensitivity and 64.93% specificity. Highest accuracy is obtained with this model. Fig. 4 and Fig. 5 depict that sensitivity, specificity, and negative predictive value are almost similar for all three models, but the positive predictive value is more than double for the integrated model than other two models. This is a very significant improvement because the prevalence of diabetes is only 3.68% in the study population. The proposed integrated model can be used in such a scenario where high positive predictive value is desirable. For example, the proposed integrated model can be useful in two ways. First, the risk of a person is predicted using the model, along with the pathology test. If the person is pathologically tested negative but predicted at high risk by the model, the person should be advised to take preventive measures. Second, it can be used as an inexpensive mass screening tool. In the first phase, the diabetes risk for the whole population is predicted using the model. In the second phase, only those persons can be pathologically tested who have been already predicted at high risk in the previous phase itself.

Table-VIII: Summary of the performance of all three proposed Diabetes Risk Prediction Models along with the other three DRS models

Prediction Model	Sensitivity [%]	Specificity [%]	PPV [%]
Mohan [3]	72.5	60.1	5.7
Chaturvedi [5]	79	56	6.1
Ramachandran [4]	76.6	59.9	6.2
Proposed Models			
Risk Score Model	72.43	59.32	6.37
Aggregate Classification Model	68.3	64.93	6.93
Integrated Model	69.89	56.58	15.88

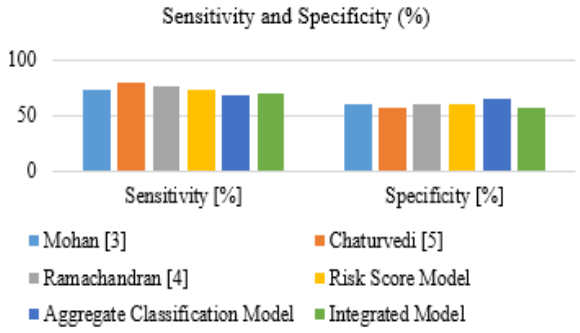


Fig. 4. Comparative study: All three proposed Diabetes Risk Prediction Models along with the other three DRS models

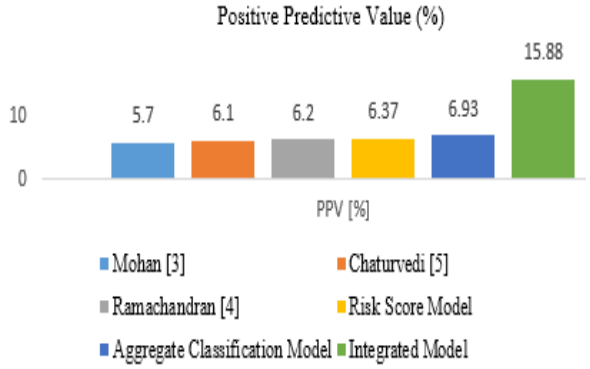


Fig. 5. Comparative study of Positive predictive value: All three proposed Diabetes Risk Prediction Models along with the other three DRS models

The two-phase approach for screening the masses can save up to 30% of the total screening cost. This will be of great use for developing countries such as India, where undiagnosed diabetes and limited financial resources are a major concern [2].

REFERENCES

1. WHO, "Diabetes, Key Facts," 30-10-2018. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes
2. WHO, "World Health Day", accessed on 02-02-2019. Available: http://www.searo.who.int/india/mediacentre/events/2016/en/
3. Mohan V., Deepa R., Deepa M., Somannavar S., Datta M., "A simplified Indian Diabetes Risk Score for screening for undiagnosed diabetic subjects," J Assoc Physicians India. 2005 Sep; 53:759-63
4. A. Ramachandran, C. Snehalatha, V. Vijay, N.J. Wareham, S. Colagiuri, "Derivation and validation of diabetes risk score for urban Asian Indians," Diabetes Research and Clinical Practice, October 2005
5. Chaturvedi V., "Development of a clinical risk score in predicting undiagnosed diabetes in urban Asian Indian adults: a population-based study." CVD prevention and control3.3 (2008): 141-151.
6. Chandrakar O., Saini J.R.. "Derivation of a Novel Diabetes Risk Score Using Semantic Discretization for Indian Population." Ambient Communications and Computer Systems. Springer, Singapore, 2018. 331-340.
7. Charlotte Glümer, Dorte Vistisen, Knut Borch-Johnsen, Stephen Colagiuri, "Risk Scores for Type 2 Diabetes Can Be Applied in Some Populations but Not All", Diabetes Care Feb 2006, 29 (2) 410-414; DOI: 10.2337/diacare.29.02.06.dc05-0945
8. Koopman, R. J., Mainous, A. G., Diaz, V. A., & Geesey, M. E. (2005). Changes in Age at Diagnosis of Type 2 Diabetes Mellitus in the United States, 1988 to 2000. Annals of Family Medicine, 3(1), 60–63. http://doi.org/10.1370/afm.214.
9. Richard P. Anderson, Ruyun Jin, Gary L. Grunkemeier, "Understanding Logistic Regression Analysis in Clinical Reports: An Introduction", The Annals of Thoracic Surgery, Volume 75, Issue 3, March 2003, Pages 753-757, Published by Elsevier Science Ink Menelaos Pavlou, Gareth Ambler, Shaun R Seaman, Oliver Guttmann, Perry Elliott
10. Lee J, "An insight on the use of multiple logistic regression analysis to estimate association between risk factor and disease occurrence," International Journal of Epidemiology 1986, 15: 22-29.
11. Office of the Registrar General & Census Commissioner, India, "Annual Health Survey Report - A Report on Core and Vital Health Indicators Part I," Ministry of Home Affairs, Government of India, New Delhi, 2016.
12. Office of the Registrar General & Census Commissioner, India, "Annual Health Survey Report - A Report on Core and Vital Health Indicators Part II," Ministry of Home Affairs, Government of India, New Delhi, 2016.
13. ORGI, "Annual Health Survey: Clinical, Anthropometric & Bio-chemical (CAB) Survey," https://data.gov.in/keywords/cab, accessed on 10-06-2016. ORGI.

14. Government of India, “Datasets”, accessed on 10-06-2018. Available: https://data.gov.in
15. Bouckaert et al., WEKA Manual for Version 3-8-1 (December 2016), University of Waikato, Hamilton, New Zealand.
16. Chandrakar O., Saini J.R. "Knowledge based Semantic Discretization using Data Mining Techniques." Int. J. Advanced Intelligence Paradigms, Inderscience Publication 340 (2017).
17. Lichman, M. (2013), “Pima Indian Diabetes Dataset,” UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
18. Chandrakar O.S., Saini J.R. “Empirical Study to Suggest Optimal Classification Techniques for Given Dataset” proc. of IEEE Int. Conf. on Computational Intelligence & Communication Technology (CICT-2015), Ghaziabad, India; Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7078662
19. Chandrakar O.S., Saini J.R. “Development of Indian Weighted Diabetic Risk Score (IWDRS) using Machine Learning Techniques for Type-2 Diabetes” proc. of The 9th Annual ACM India Conf. (Compute-2016), Ahmedabad, India; Available: http://dl.acm.org/citation.cfm?id=2998497
20. Chandrakar O.S., Saini J.R., Barik L.B. “Validation of Semantic Discretization based Indian Weighted Diabetes Risk Score (IWDRS)” Int. J. of Advanced Computer Science and Applications, 2017, vol. 8(10): 436-439
21. Chandrakar O.S., Saini J.R., Bhatti D.G. “Novel Semantic Discretization Technique for Type-2 Diabetes Classification Model”, proc. of 6th Int. Conf. on Innovations in Computer Science and Engineering (ICICSE-2018), Hyderabad, India; Available: https://doi.org/10.1007/978-981-13-7082-3_17

AUTHORS PROFILE

Omprakash S. Chandrakar is working as Associate Professor at Uka Tarsadia University, Bardoli, Gujarat, India. He is also pursuing PhD from Uka Tarsadia University, Bardoli. To his credit, he has a number of publications in reputed international and national journals as well as conferences. His publications have been widely cited and indexed by Scopus as well as Web of Science among others. He has also presented a number of papers in the conferences.

Jatinderkumar R. Saini is working as Professor at Symbiosis Institute of Computer Studies and Research, Symbiosis International Deemed University, Pune, Maharashtra, India. He has completed PhD in Computer Science and MCA with Gold Medals in all 3 years. He has many research publications to his credit. His papers have been cited in almost all parts of the world. He has been awarded for best papers as well as outstanding reviewer by the research journals bearing Thomson Reuters impact factors.