Application of Data Mining Techniques in Deriving Waist Circumference-Age Index for Diabetes Risk Score

Omprakash Chandrakar
Associate Professor,
Uka Tarsadia University
Bardoli, Gujarat, India

Jatinderkumar R. Saini, PhD
Professor & I/C Director,
Narmada College of
Computer Application
Bharuch, Gujarat, India

ABSTRACT
Diabetes Risk Score (DRS) tools are computational tools, used to assess the risk of a person's getting diabetes. DRS tools are generally used as a simple, inexpensive and non-invasive mass screening tool to detect diabetes. Various DRS tools are reported in literature and being used successfully. The accuracy of the DRS tools highly depends on the parameters used to derive it. Total Diabetic Risk Score is calculated by adding individual parameter's risk scores. This approach won't work, if any pair of parameters is negatively correlated with diabetes risk. In such cases, it reduce the total diabetes risk score when one parameter is kept constant and other is decreased, while they are actually expected to increase it. In this research study, researchers propose a new parameter Waist Circumference Age Index (WAI), to address the above issue. This paper also discusses the derivation of criteria for determining high and low risk for diabetes based on WAI using machine learning technique. The outcome of this research study can be used to develop a new Diabetes Risk Score tool.

General Terms
Machine Learning, Clustering, Diabetes Risk Score

Keywords
Association Rule Mining, Clustering, Discretization, Diabetes Risk Score, Indian Weighted Diabetes Risk Score, Machine Learning, Type -2 Diabetes

1. INTRODUCTION
Diabetes mellitus (DM) has become the most common non-communicable diseases across the globe. According to the latest report of International Diabetes Federation (IDF) [1], around 415 million people in the world, which is 8.8% of adults aged 20 to 79 years, are estimated to have diabetes. Around three fourth of them come from low and middle income countries like India. It is projected that there will 642 million people with diabetes by 2040. The Indian diabetes scenario is worse than global scenario. About 98.18 million people with diabetes live in India. This is 9.3% of its total population aged 20 to 79 years. In other words 1 out of 10 people is diabetic in India. Even the worst part is that out of 98.18 million, 36 million diabetes cases are undiagnosed [2].

Computational methods such as association rule mining [3], classification [4], [26] clustering [5], [6], [7], neural network [8] and other machine learning techniques [9] have been used in predicting diabetes. Diabetes Risk Score (DRS) tools are computational tools. It uses various health parameters, socio-economic indicators, family and personal history to assess the risk of a person's getting diabetes. DRS tools are being used as an inexpensive and efficient method for mass screening test for detecting undiagnosed diabetes. DRS tools are derived using statistical methods, machine learning techniques and artificial intelligence. DRS tools assign specific scores to the specific range of parameters according to the risk associated with it. In order to access the diabetic risk of a person, all individual parameter's risk scores, scored by that person is added to get total diabetes risk score. Then, based on the criteria, risk of that person is determined with the help of total diabetes risk score [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23]. DRS tools are useful in two ways [24].

1. DRS tools can be used as an inexpensive mass diabetic screening tool. In the first phase DRS will be calculated for all participants. In second phase only those participants will be tested for diabetes pathologically, who scored high on DRS.
2. Participants who scored high on DRS but tested negative on pathological test, will be advised to undergo certain life style change, to avoid or at least minimize, diabetes risk [25].

Chandrakar and Saini [24], [27] have derived and validated novel Indian Weighted Diabetes Risk Score (IWDRS) tool for Indian population. Using machine learning techniques, they have derived criteria to categorize continuous range of each of 9 parameters into low, moderate and high diabetes risk. They considered the following 9 parameters; Age, Family history, Personal history, BMI, Waist circumference, Diet, Stress and Physical activity. For this study, researchers considered two of them; Age and Waist circumference.

As they reported in their work, the criteria for diabetes risk for age and waist circumference are given in fig. 1. These rules categorized age and waist circumference measurement to low, moderate & high according to their diabetes risk.

(Age in years, Waist circumference in inches)

For Female

If Age<42 then diabetes risk = Low

If Age >= 42 and Age<=58 then diabetes risk = High

If Age>58 then diabetes risk = Moderate

If Waist circumference<33 then diabetes risk = Low

If Waist circumference >= 33 and <=38 then diabetes risk = High

If Waist circumference >38 then diabetes risk = Moderate

For Male

If Age<41 then diabetes risk = Low

If Age >= 41 and Age<=61 then diabetes risk = High

If Age>61 then diabetes risk = Moderate

If Waist circumference<36 then diabetes risk = Low

If Waist circumference >= 36 and <=43 then diabetes risk = High

If Waist circumference >43 then diabetes risk = Moderate

Fig. 1: Diabetes Risk Criteria for Age and Waist Circumference

Table 1 shows Indian weighted diabetic risk score (IWDRS) for Age and Waist circumference.

Table 1. Indian weighted diabetes risk score (IWDRS) for Age (years) and Waist circumference (inches)

		Indian Weighted Diabetic Risk Score		
No	**Attribute**	**Low**	**Moderate**	**High**
1	Age	10	27	63
5	Waist circumference	15	41	44

Total Indian Weighted Diabetic Risk Score (TIWDRS) is calculated by summing up these two individual parameter's risk scores along with other 7 parameters.

2. RESEARCH GAPE

Straight forward summation of all individual parameter's risk scores for getting TIWDRS and based on TIWDRS, determining overall diabetes risk [24] is perfectly right, only if there is no pair of parameters which are negatively correlated. If such pair exist, then this approach will neutralize each other's impact, while they are actually supposed to enhance it. Consider the following illustration of 3 male persons.

Table 2. Illustration of 3 male people's cases

	Age	**IWDRS**	**Waist Circum**	**IWDRS**	**TIWDRS**
Person 1	35 (Low)	10	34 (Low)	15	25
Person 2	35 (Low)	10	38 (Moderate)	41	51
Person 3	45 (Moderate)	27	38 (Moderate)	41	68

TIWDRS (considering only two parameters, age & waist circumference), is 25, 51 and 68 respectively for person 1, person 2 and person 3 according to table 2. Risk score of person 1 and person 3 is logically right, but risk score of person 2 should be more than of person 3 because age and waist circumference is negatively correlated with diabetes risk. In other words, keeping waist circumference constant, if age goes down, diabetes risk is expected to go high, but the existing, method of simple summation brings it down.

To address the above issue, researchers introduce a new parameter Waist circumference – Age Index (WAI).

Rest of the paper is organized as follows.

1. Section 3 briefly outlines the research methodology.
2. In section 4, researchers describe how relationship between age and waist circumference is discovered using association rule mining.
3. In section 5, a new indicator Waist Circumference – Age Ratio (WAR) is introduced.
4. Section 6 discusses the derivation of criteria for diabetes risk based on WAR.
5. Section 7 concludes the paper.

3. RESEARCH METHODOLOGY

1. Determine if any correlation between waist circumference and age does exist by applying association rule mining on diabetes dataset.
2. Average Age adjusted Waist circumference – Age Ratio (AAWAR) is used to correlate both parameters. A pattern is identified to categorize a person into low or high diabetes risk based on AAWAR.
3. Diabetes dataset is divided into 4 datasets based on gender and age.
4. Age adjusted WAR is calculated for each record on the all 4 datasets obtained in step 4 and included into corresponding dataset as a new attribute Age adjusted WAR (AWAR).
5. Clustering is performed on Age adjusted WAR for all 4 datasets obtained in step 5.
6. Cluster centroids and standard deviation obtained for datasets A and C is used to determine the criteria for low or high diabetes risk based to Age adjusted WAR is described.
7. Cluster centroids obtained for datasets B and D is compared with A & C respectively.
8. A new parameter Waist Circumference-Age Index (WAI) is introduced based on Age adjusted WAR for simplicity.

4. DETERMINING CORRELATION BETWEEN WAIST CIRCUMFERENCE AND AGE

Association rule mining is applied to find out if any correlation exists between waist circumference and age. Weka 3.6, a well-known data mining tool used for research purpose, is used to carry out the experiments in this study [28].

4.1 Dataset

For this study researchers used diabetes dataset that contain 844 records of diabetes patient. The dataset contains, gender, age (in year) and waist circumference measurement (in inch) along with the other attributes.

4.2 Data Preprocessing

Apriori algorithm can be performed on the dataset that contains only nominal attributes. So before applying apriori algorithm all attributes are needed to be transformed into nominal attribute. In diabetes dataset, age and waist circumference both contains continuous integer value. Their values are replaced with corresponding risk level (Low, Moderate and High) using rules depicted in fig. 1 [24].

4.3 Knowledge Extraction & Interpretation

Apriori algorithm is applied on the diabetes dataset which is obtained after data pre-processing step 4.2, with an objective to discover any association exists between age and waist circumference. 6 relevant rules are discovered. They are shown in fig. 2.

1. AGE_T=M 246 ==> WAIST_T=M 152 conf:(0.62)
2. AGE_T=H 446 ==> WAIST_T=M 270 conf:(0.61)
3. AGE_T=L 152 ==> WAIST_T=L 72 conf:(0.47)
4. AGE_T=L 152 ==> WAIST_T=M 64 conf:(0.42)
5. AGE_T=M 246 ==> WAIST_T=L 80 conf:(0.33)
6. AGE_T=H 446 ==> WAIST_T=L 140 conf:(0.31)

Fig. 2. Applying association rule on age and waist circumference

Researchers observed that when

1. Age is moderate than either waist circumference is moderate (confidence = 62%) or low (confidence = 33%) [Rule # 1 & 5]
2. Age is high than either waist circumference is moderate (confidence = 61%) or low (confidence = 31%) [Rule # 2 & 6]
3. Age is low than waist circumference is low (confidence = 47%) or moderate (confidence = 42%) [Rule # 3 & 4]

From observation 3, it can be concluded that if individually age falls into low diabetes risk category and waist circumference falls into moderate diabetes risk category, but if taken both parameters collectively, it falls into high risk category as we argued in section 2.

To address this issue, researchers introduce a new parameter that correlates Age and Waist circumference as single parameter.

5. AVERAGE AGE ADJUSTED WAR (AAWAR)

Researchers introduce a new parameter Age Adjusted Waist circumference – Age Ratio (AWAR), which is defined as follows;

Age adjusted WAR (AWAR) =

{Waist Circumference (in inch) /Age (in year) for age <=35

{Waist Circumference (in inch)/35 for age >35 ---------- (2)

Table 11 shows Average Age adjusted WAR which is calculated as follows

Average Age adjusted WAR(AAWAR)

$= \sum_{n=1}^{N} (\boldsymbol{Waist\ Circumference_n\ /Age_n})/\boldsymbol{N}$ if Age<=35

$= \sum_{n=1}^{N} (\boldsymbol{Waist\ Circumference_n\ /35})/\boldsymbol{N}$ if Age>35

---------- (3)

A significant pattern is observed. For the age less than or equal to 35, Average Age adjusted WAR is more than 1.

Waist circumference measurement for male and female is significantly different, so researchers separately calculated Average Age adjusted WAR for male and female and shown in table 3 and 4. It is observed that Average Age adjusted WAR is more than 1 for the age <=35 for female and male.

Table 3. Average Age Adjusted WAR for female

Age Group (Year)	Average Age (Year)	Average Waist Circumference (Inch)	Average Age adjusted WAR
0-25	21.4	33.2	1.59
26-30	28.75	36.75	1.28
31-35	33	37.67	1.15
36-40	38.38	35.06	0.97
41-45	43.75	35.05	0.97
46-50	48.6	35.86	1
51-55	53.41	35.03	0.97
56-60	58.19	37.36	1.04
61-65	62.82	34.82	0.97
66-70	68.09	35.45	0.98
71-75	72.75	34.5	0.96

*No records found for age over 75 years.

Table 4. Average Age adjusted WAR for male

Age Group (Year)	Average Age (Year)	Average Waist Circumference (Inch)	Average Age adjusted WAR
0-25	27.5	33	1.51
26-30	28.82	36.75	1.28
31-35	33.08	34.23	1.03
36-40	37.95	35.5	0.99
41-45	43.08	36.08	1
46-50	48.09	35.85	1
51-55	53.05	36.05	1
56-60	58.19	37.36	1.04
61-65	63.18	35.21	0.98
66-70	68.53	35.76	0.99
71-75	73.43	34.71	0.96
76-80	77	36	1
81-85	82	43	1.2

*No records found for age over 85 years.

6. DERIVATION OF WAIST CIRCUMFERENCE – AGE INDEX (WAI)

This section explains derivation of criteria for low and high diabetes risk based on Age adjusted WAR.

6.1 Dataset

Dataset described in section 4.1 is used for the further study that contains 844 records of diabetes patient.

6.2 Data Preprocessing

Following transformations are applied on the dataset.

1. Two datasets are created, separating patients with age <=35 and age >35.
2. Further, each of the above dataset is bifurcated into two datasets, separating female and male patient. 4 dataset are obtained as follows;
    1. Dataset with female patient and age <=35
    2. Dataset with female patient and age >35
    3. Dataset with male patient and age <=35
    4. Dataset with male patient and age >35
3. Age adjusted WAR is calculated for each records for all 4 datasets and it is included into corresponding datasets as a new attribute.

6.3 Experiment: Applying Distance based Clustering

Now the objective is to determine criteria for low and high diabetes risk using Age adjusted WAR.

Distance based clustering on Age adjusted WAR is applied for all the four dataset obtained in previous section. Clustering results are shown in table 5 and 6 for female and male respectively.

Table 5. Clustering on Age adjusted WAR for female with age <=35 and age>35

	Age adjusted WAR Female with age<=35	**Age adjusted WAR Female with age>35**
Cluster centroids	1.185	0.9621
Std Dev	0.0817	0.0327
Instances %	67	90

Table 3 shows that a cluster with cluster centroids is 1.185 is found for the dataset that contains female with age<=35. 67% of the instances fall in this cluster. Here standard deviation is 0.0817. Subtracting standard deviation from cluster centroid, we get 1.1033 (=1.185-0.0817). Here we can infer that, for the maximum female patient at or under age 35, Age adjusted WAR will be >=1.1033. In other words, for a female person whose age is less than or equal to 35, is on high risk for diabetes if her Age adjusted WAR is >=1.1033.

If [(Gender = Female) AND (Age<=35) AND (Age adjusted WAR >= 1.1033)]

Then

Diabetes Risk = High ---------- (4)

Similarly, from table 4, we can derive following criteria for male.

If [(Gender = Male) AND (Age<=35) AND (Age adjusted WAR >= 0.984)]

Then

Diabetes Risk = High ---------- (5)

Table 6. Clustering on Age adjusted WAR for male with age <=35 and age>35

	Age adjusted WAR Male with age<=35	**Age adjusted WAR Male with age>35**
Cluster centroids	1.079	0.9491
Std Dev	0.095	0.0467
Instances %	81	73

Researchers propose a new parameter Wait Circumference-Age Index (WAI) based on the above experimental results and discussion as follows

WAI = { Waist Circumference (in inch)*100/Age (in year)

for age <=35

Waist Circumference (in inch) *100/35 for age >35

Researchers obtained following criteria for determining diabetes risk by substituting WAI in equation 4 and 5. In similar way criteria for the person over 35 years is obtained.

For the female person age <=35

If WAI>=110 then Diabetes Risk Category is High

Else Low

For the male person age <=35

If WAI>=98 then Diabetes Risk Category is High

Else Low

For the female person age >35

If WAI>=93 then Diabetes Risk Category is High

Else Low

For the male person age >35

If WAI>=90 then Diabetes Risk Category is High

Else Low

Fig 3. Criteria for determining diabetes risk based on WAI

7. CONCLUSION

Diabetes risk of a person is calculated by adding diabetes risk scores of all parameters. This approach does not work if there is some pair of parameters which are negatively correlated with diabetes risk. In this paper, we identified two such parameters in IWDRS, age and waist circumference. Keeping waist circumference constant, if age goes down, diabetes risk score should goes high. But the approach used in calculating TIWDRS, bring down the predicted risk score instead of enhancing it. To overcome this limitation, researchers introduced a new parameter Wait Circumference-Age Index (WAI). Criteria for diabetes risk, based on WAI, is derived by applying clustering on diabetes dataset, which is depicted in fig. 3. The outcome of this research study may be used to derive a more accurate diabetes risk score tool for Indian population.

8. REFERENCES

[1] http://www.diabetesatlas.org, Accessed on 5th December 2016.

[2] IDF DIABETES ATLAS Seventh Edition 2015, ISBN: 978-2-930229-81-2, © International Diabetes Federation, 2015

[3] Karthikeyan, T., & Vembandasamy, K. (2015). A Novel Algorithm to Diagnosis Type II Diabetes Mellitus Based on Association Rule Mining Using MPSO-LSSVM with Outlier Detection Method. Indian Journal of Science and Technology, 8(S8), 310-320.

[4] Saravananathan, K., & Velmurugan, T. (2016). Analyzing Diabetic Data using Classification Algorithms in Data Mining. Indian Journal of Science and Technology, 9(43).

[5] Venkataraman, S., Sivakumar, S., & Selvaraj, R. (2016). A Novel Clustering based Feature Subset Selection Framework for Effective Data Classification. Indian Journal of Science and Technology, 9(4).

[6] Nagarajan, S., & Chandrasekaran, (2015). Design and Implementation of Expert Clinical System for Diagnosing Diabetes Using Data Mining Techniques. Indian Journal of Science and Technology, 8(8), 771-776.

[7] Sharmila, K., & Vetha Manickam, S. (2016). Diagnosing Diabetic Dataset using Hadoop and K-means Clustering Techniques. Indian Journal of Science and Technology, 9(40). doi:10.17485/ijst/2016/v9i40/101618

[8] Singh Gill, N., & Mittal, P. (2016). A Novel Hybrid Model for Diabetic Prediction using Hidden Markov Model, Fuzzy based Rule Approach and Neural Network. Indian Journal of Science and Technology, 9(35).

[9] Kalaiselvi, C., & Nasira, G. (2015). Prediction of Heart Diseases and Cancer in Diabetic Patients Using Data Mining Techniques. Indian Journal of Science and Technology, 8(14).

[10] American Diabetes Association, http://main.diabetes.org/dorg/PDFs/risk-test-paper-version.pdf

[11] http://www.diabetes.fi

[12] https://www.diabetes.org.uk

[13] http://care.diabetesjournals.org/content/24/6/1120

[14] http://www.health.gov.au/internet/main/publishing.nsf/content/diabetesriskassessmenttool

[15] http://leicesterdiabetescentre.org.uk/The-Leicester-Diabetes-Risk-Score

[16] http://healthycanadians.gc.ca/diseases-conditions-maladies-affections/disease-maladie/diabetes-diabete/canrisk/index-eng.php

[17] http://www.diabetesqld.org.au/healthy-living/who-is-at-risk/assess-your-risk.aspx

[18] Mohan V1, 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.

[19] Shashank R Joshi, Indian Diabetes Risk Score, JAPI • VOL. 53 • SEPTEMBER 2005, www.japi.org

[20] Heikes KE1, Eddy DM, Arondekar B, Schlessinger L., Diabetes Risk Calculator: a simple tool for detecting undiagnosed diabetes and pre-diabetes, Diabetes Care. 2008 May; 31(5):1040-5. Epub 2007 Dec 10.

[21] Charlotte Gl et al, "A Danish Diabetes Risk Score for Targeted Screening", Diabetes Care, Volume 27, Number 3, March 2004

[22] Lanord Stanley, J. M. et al. Evaluation of Indian Diabetic Risk Score for Screening Undiagnosed Diabetes Subjects in the Community. Indian Journal of Science and Technology, [S.l.], p. 2798-2799, jun. 2012. ISSN 0974 - 5645.

[23] Lanord Stanley J, M., Elantamilan, D., & Kumaravel, T. (2013). Prevalence of Prehypertension and its Correlation with Indian Diabetic Risk Score in Rural Population. Indian Journal Of Science And Technology, 6(8), 5163-5166

[24] Omprakash Chandrakar, Jatinderkumar R. Saini, "Development of Indian Weighted Diabetic Risk Score (IWDRS) using Machine Learning Techniques for Type-2 Diabetes", COMPUTE '16 Proceedings of the 9th Annual ACM India Conference, Pages 125-128, ACM New York, NY, USA ©2016, ISBN: 978-1-4503-4808-9, doi 10.1145/2998476.2998497

[25] Tuomilehto J, Lindstrom J, Eriksson JG, Valle TT, Hämäläinen H, Ilanne-Parikka P, et al.Finnish Diabetes Prevention Study Group. Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. N Engl J Med. 2001; 344:1343–50.

[26] Omprakash Chandrakar, Dr. Jatinderkumar R. Saini, Comparative Analysis of Prediction Accuracy of General and Personalized Datasets Based Classification Model for Medical Domain, International Journal of Advanced Networking Applications (IJANA) ISSN No. : 0975-0290 25.

[27] Omprakash Chandrakar, Dr. Jatinderkumar R. Saini, Validation of Indian Weighted Diabetes Risk Score (IWDRS), International Journal of Computer Applications, ISSN No. 0975 – 8887, Volume 158, January 2017.

[28] www.cs.waikato.ac.nz/ml/weka