RESEARCH ARTICLE OPEN ACCESS

Adaptive Indian Diabetes Risk Score(AIRDS)- A Screening Tool For Predicting Diabetes Using Machine Learning

Dr.Suruchi Pimple[1]
[1]*Sadabai Raisoni Women's College, Nagpur*

Dr.Sunil Gupta[3]
[3]*Sunil's Diabetes Care n Research Center Pvt. Ltd., Nagpur*

Dr.Gopal Sakarkar[2]
[2]*Dr.Vishwanath Karad MIT World Peace University, Pune*

Priyanka Mandve[4]
[4]*Sadabai Raisoni Women's College, Nagpur*

ABSTRACT

The objective of this study was to develop a simple non-invasive risk score to identify individuals having increased risk of undiagnosed type 2 diabetes mellitus. A backward stepwise variable selection model building approach was used for calibrating the risk score using linear regression. The data collected from local hospital had 1069 instances and 20 social, medical and anthropometric features. Akaike Information Criterion was used to evaluate the three models. An Adaptive Indian Diabetes risk score was formulated on significant features obtained from the best fit model 2 using *p*-scores. The study inferred that individuals who were infected by Covid19, had smoking habits, consumed of alcohol or had suffered from gestational diabetes were more prone to becoming diabetic. These significant features not considered in earlier researches.

Keywords: Diabetes, Machine learning, risk score

1. INTRODUCTION

The global incidence of type 2 diabetes is on the rise across all populations, posing a significant risk for mortality and a plethora of nonfatal complications that can impose a substantial burden on patients, their families, and healthcare systems [1]. Recent intervention studies unequivocally demonstrate that lifestyle modifications effectively prevent type 2 diabetes in individuals at high risk [2]. Consequently, the primary challenge for public health authorities lies in pinpointing those who would benefit from intensive lifestyle counseling [3].

The use of blood glucose screening has been suggested as a potential tool to identify individuals with a heightened risk of diabetes or those with asymptomatic diabetes. There is an ongoing debate about whether screening for fasting glucose alone is adequate or if an oral glucose tolerance test is necessary to detect asymptomatic diabetes. Both fasting and postprandial blood glucose measurements are invasive, costly, and time-consuming. Additionally, blood glucose levels exhibit significant random variations and only offer insights into the individual's current glycemic status.

However, the ultimate goal of primary prevention is to identify high-risk individuals while they are still prediabetic and intervene with measures that deter their progression from prediabetic to impaired glucose tolerance and, ultimately, diabetes [4].

A risk score, also known as a risk prediction or risk assessment score, is a numerical value assigned to an individual based on various factors to estimate the likelihood of developing a particular disease or condition. The calculation of a risk score involves using statistical models and data from population studies to identify risk factors associated with the disease in question. The goal is to stratify individuals into different risk categories, helping healthcare professionals make informed decisions about prevention, early detection, and intervention. The risk scores are estimates and probabilities, not definitive predictions [5]. They serve as a valuable tool

to assist healthcare workers in making informed decisions and tailoring interventions to individual's specific risk profiles.

2. RELATED WORK

The Framingham diabetes risk prediction model encompasses factors such as overweight, obesity, impaired fasting glucose, low HDL cholesterol, high triglycerides, elevated blood pressure, and a parental history of diabetes [6]. It employs a point scoring algorithm, where the risk of incident diabetes is associated with an individual's cumulative point score. The ARIC Study derived its diabetes risk prediction model, which considers height, waist circumference, black race/ethnicity, systolic blood pressure, fasting glucose, HDL cholesterol, triglycerides, and parental history of diabetes.

In the San Antonio diabetes risk prediction model, age, sex, Mexican-American ethnicity, fasting glucose, systolic blood pressure, HDL cholesterol, body mass index, and family history of diabetes (parent or sibling) are considered [7]. The MESA model, designed for a more diverse population in terms of age and race/ethnicity, includes additional variables [8].

The researcher [9] formulated a risk assessment score specifically for individuals in the United Arab Emirates through the utilization of a stepwise forward regression model. The study encompassed a total of 872 UAE citizens, revealing an overall diabetes prevalence of 25.1% among adult citizens in the Northern Emirates. Noteworthy risk factors associated with diabetes included age (≥35 years), a family history of diabetes mellitus, hypertension, a body mass index (BMI) of ≥30.0, and a waist-to-hip ratio of ≥0.90 for males and ≥0.85 for females. The model's performance exhibited a moderate level of sensitivity (75.4%, 95% CI 68.3 to 81.7) and specificity (70%, 95% CI 65.8 to 73.9).

The risk assessment score was established through data from the Chennai Urban Rural Epidemiology Study (CURES) with 26,001 individuals. Every tenth participant was invited to partake in Phase 3 for diabetes screening using the World Health Organization. The response rate stood at 90.4% (2350/2600). The Indian Diabetes Risk Score (IDRS) was formulated based on the outcomes of multiple logistic regression analysis, with internal validation conducted on the same dataset. IDRS utilized four key risk factors: age, abdominal obesity, family history and physical activity.

The score of value of ≥60 was found accurate for identifying undiagnosed diabetes [10].

3. METHODOLOGY

A backward stepwise variable selection model building approach was used for developing a risk score using linear regression. Three models of linear regression were applied on preprocessed data collected from local hospital having 1069 instances and 20 features. Table 1 depicts the social and physiological features of individuals of dataset. The first regression model included all the features, whereas second model contained 8 features found significant from first model. The third model included five significant features. OLS Regression Model of Statsmodel library was used for analyzing results. Ordinary least-squares (OLS) models presuppose the modeling of a relationship between one or more explanatory variables and a continuous, or at least interval, outcome variable.

Table 1. Social, medical and anthropometric features of individuals of dataset

Variable	**Non-Diabetic**	**Diabetic**	***p*-value**

Body Mass Index			
<25	163	277	0.18
25-29	155	259	
>=29	79	136	
Hypertension Status			
No	288	488	0.27
Yes	109	184	
Gender			
Male=0	155	378	0.35
Female=1	242	294	
Age			
<35	114	181	0.08
35-49	135	196	
>=50	148	295	
Physically Active			
Yes	224	375	0.06
No	173	297	
Consume Alcohol			
No	353	589	0.39
Yes	44	83	
Infected by Covid			
No	379	557	0.18
Yes	18	115	
Gestational Diabetes			
No	268	601	0.6
Yes	129	71	
Family History			
No	120	185	0.3
Yes	277	487	

A backward stepwise variable selection model building approach was used for developing a risk score using linear regression. Three models of Linear regression were applied on preprocessed data collected from local hospital having 1069 instances and 20 features. Table 1 depicts the social and physiological features of individuals of dataset. The first regression model included all the features, whereas second model contained 8 features found significant from first model. The third model included five significant features. OLS Regression Model of Statsmodel library was used for analyzing results. Ordinary least-squares (OLS) models presuppose the modeling of a relationship between one or more explanatory variables and a continuous, or at least interval, outcome variable.

The model aims to minimize the sum of squared errors, where an error signifies the disparity between the actual and predicted values of the outcome variable. The prevalent analytical technique employing OLS models is linear regression, encompassing both single and multiple predictor variables. Linear Regression mathematically models the unknown or dependent variable and the known or independent variable as a linear equation.is implemented by using following equation.

$$Y_i = f(X_i, \beta) + e_i$$

(1)

Where Y_i = dependent variable, f = function, Xi = independent variable, B = unknown parameters

e_i = error terms

Performance Metrics: The proposed models were evaluated on Akaike Information Criterion(AIC), Root Mean Square, R-Squared and Adjusted R-Squared. AIC factor

was used for model selection to distinguish among a set of possible models describing the relationship between selected features. A difference of 2 or more in AIC values indicates a significantly different model fit [11]. AIC score can be calculated by following equation.

$$AIC = -2\ ln\ ln\ (\sigma_\varepsilon^2) + 2k \quad (2)$$

Where 2k = penalty, σ_ε^2 =variance of residuals

Root mean square is the arithmetic mean of the squares of a given set of values. RMS values can be both positive and negative. It is used to check how far a typical value is from the average value.

$$x_{rms} = \sqrt{\frac{1}{n}}(x_1^2 + x_2^2 + \cdots + x_n^2 \quad (3)$$

where xi are values and n are total number of observations

R^2 is the coefficient of determination that tells us that how much percentage variation independent variable can be explained by independent variable. The maximum possible value of R2 can be 1, means the larger the R^2 value better the regression. It is calculated using following equation.

$$R^2 = 1 - \frac{SSR}{SST} \quad (4)$$

where SSR is sum squared regression and SST is total sum of squares.

Adjusted R^2 analyses the number of predictors in the model and penalizes excessive variables, providing a more accurate measure of the model's goodness.

$$Adjusted R^2 = 1 - \frac{(1-R^2)(N-1)}{N-p-1} \quad (5)$$

where R^2 coefficient of determination, N is total sample size and p is number of independent variables.

Algorithm Risk Score Calculation: The following steps were used for formulating the Adaptive Indian Diabetes Risk Score.

1. Input dataset
2. Apply linear regression to find p values
3. Extract new attributes based on p<0.05
4. Assign risk score to features by calculating int (beta coefficient * 10)
5. Formulate the Adaptive Indian Diabetes Risk Score
6. Set the Score cutoff for high risk value (>60% of total score)

4. EXPERIMENTAL RESULTS

The study was conducted in two phases. In the first phase regression model on three set of features was implemented. In phase 2, the best fit model was identified using parameters AIC Score, Mean Square Error and R^2. Next, the significant features identified from Best Fit Model 2 were used to calibrate a risk score for predicting diabetes. The proposed regression models were trained and tested on a dataset collected from a local hospital. The first regression model included all the 20 social, medical and anthropometric features of the dataset. In the second regression model, features with p>0.05 were eliminated. 60% of the earlier feature set was reduced to 40%. The features selected for second model were age, height, Body Mass Index, smoking, infected by Covid19, pregnancies, gestational diabetes and consumption of alcohol. The features used in model 3 were gender, age, post meal blood glucose, H1AC value. In the third regression model features with p>0.05 were eliminated. The significant features found were post meal blood glucose and H1AC value. Figure 1 depicts the three regression models with feature sets.

Experiment Results of Model 1: Features used in Linear Regression were- one dependent variable: CLASS (1 for Diabetic, 0 for non-diabetic) and other independent variables: gender, age, height, weight, Body mass index, hypertension, physical activity, family history, no. of hours of sleep, smoking, snoring, infected with Covid19, no. of pregnancies, gestational diabetes, consumption of alcohol, Systolic Blood pressure, Diastolic blood pressure, H1AC value. Table 2 compares the three regression models on parameters AIC, Mean Square Error and R2.

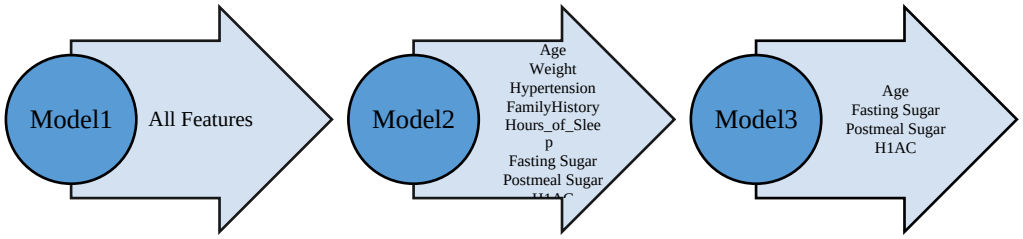

Fig. 1. Models with selected features

Table 2. Comparative study for all the models

	MSE	**RSQUARED**	**ADJUSTED_R^2**	**AIC**
MODEL1	0.207	0.124	0.1118	1370
MODEL2	0.2064	0.1233	0.1167	1356
MODEL3	0.227	0.02996	0.025	1458

Experiment Results of Model 2: Features used in Linear Regression were- one dependent variable: CLASS(1 for Diabetic,0 for non-diabetic) and other independent variables: age, height, Body mass index, smoking, infected with Covid19, no. of pregnancies, gestational diabetes, consumption of alcohol. Features with p>0.05 were eliminated which were gender, weight, hypertension, family history, no. of hours of sleep, fasting blood glucose, post meal blood glucose, H1AC value. The features considered significant were age, height, Body mass index, smoking, infected with Covid19, no. of pregnancies, gestational diabetes, consumption of alcohol.

Experiment Results of Model 3: Features used in Linear Regression were one dependent variable: CLASS(1 for Diabetic,0 for non-diabetic)and other independent variables: age, fasting blood glucose, post meal blood glucose, H1AC value. Features with p>0.05 were eliminated were gender and age. The features considered significant were post meal blood glucose, H1AC value. Table 3 indicated the values of beta coefficients and the calculated scores for the four most significant features of the dataset.

Fig. 2. Best Fit model on AIC scores

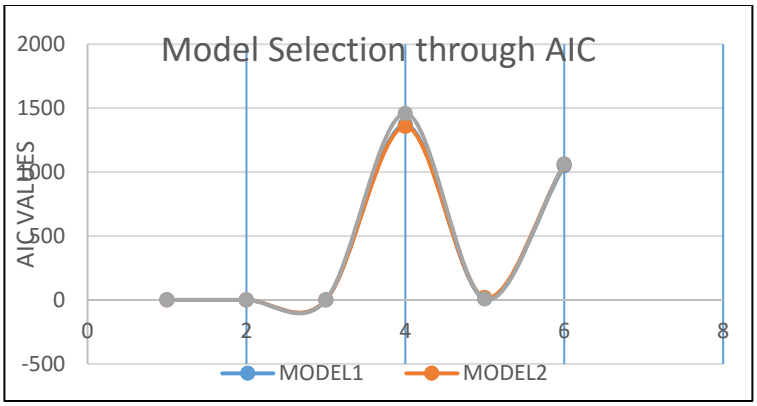

Table 3. Score calculated for most significant features

SRNO.	FEATURE	Beta Coefficient	SCORE
1	SMOKING		
	N Y	0.1184	0 1
2	INFE_COVID		
	N Y	0.2102	0 2
3	GESTDIAB		
	N Y	-0.1348	0 1
4	CON_ALCOHOL		
	N Y	-0.2673	0 3

Model2 was found to have the best fit as AIC, Mean Square Error and R^2 is lowest as compared to Model 1 and Model 3. The highest Adjusted R^2 of Model 2 implies that the overall regression was meaningful. Figure 2 depicts the best fit model on basis of AIC scores.

Risk Score Formulation: The risk score were calculated using the b coefficients of Model 2 using following formula. The beta coefficient of features with *p value* <0.05 were multiplied by 10 and round off to nearest integer.

SCORE = round (abs (β coefficient *10)) (6)

As per Standard Risk Scores, if score is >60% of total score, then risk of acquiring Diabetes is high or else the risk of acquiring diabetes is low. In Adaptive Indian Diabetes Risk Score, the total score=07, so risk score more than 04 will indicate high risk of acquiring diabetes.

5. CONCLUSION

This study developed a risk score model that is specifically designed and appropriate for Indian citizens to assess diabetes risk. Three regression models were proposed in the study wherein Regression Model2 was found to have the best fit model. Akaike Information Criterion, Mean Square Error and R^2 were found lowest as compared to Model 1 and Model 3. The highest Adjusted R^2 of Model 2 implied that the overall regression was meaningful. The risk score Advanced Indian Diabetes Risk Score was developed based on

significant features identified from Model 2. The research suggests that individuals who had contracted Covid-19, engaged in smoking, consumed alcohol or had suffered from gestational diabetes were at a higher risk of developing diabetes. The Adaptive Indian Diabetes Risk Score(AIDRS) will act as a non-invasive, inexpensive and safe tool for identifying individuals at high risk of having undiagnosed diabetes.

References:

1. Saxena, R., Sharma, S. K., Gupta, M., & Sampada, G. C. (2022). A novel approach for feature selection and classification of diabetes mellitus: Machine learning methods. *Computational Intelligence and Neuroscience*, *2022*, 1–11. https://doi.org/10.1155/2022/3820360
2. Suruchi Dive, & Gopal Sakarkar. (2022). Machine Learning For Non- Invasive Diagnostics Of Glucose Metabolism Disorder. International Journal of Next-Generation Computing, 13(5). https://doi.org/10.47164/ijngc.v13i5.9684.
3. Yilmaz, A. (2022). Prediction of type 2 diabetes mellitus using feature selection-based machine learning algorithms. *Health Problems of Civilization*, *16*(2), 128–139. https://doi.org/10.5114/hpc.2022.114541
4. Bild, D. E., Bluemke, D. A., Burke, G. L., Detrano, R., Diez Roux, A. V., Folsom, A. R., ... & Tracy, R. P. (2002). Multi-ethnic study of atherosclerosis: objectives and design. American journal of epidemiology, 156(9), 871-881.
5. Senan, E. M., Abunadi, I., Jadhav, M. E., & Fati, S. M. (2021). Score and correlation coefficient-based feature selection for predicting heart failure diagnosis by using machine learning algorithms. *Computational and Mathematical Methods in Medicine*, *2021*, 1–16. https://doi.org/10.1155/2021/8500314
6. Kannel, W. B., & McGee, D. L. (1979). Diabetes and cardiovascular risk factors: the Framingham study. Circulation, 59(1), 8-13.
7. Wei, M., Gaskill, S. P., Haffner, S. M., & Stern, M. P. (1998). Effects of diabetes and level of glycemia on all-cause and cardiovascular mortality: the San Antonio Heart Study. Diabetes care, 21(7), 1167-1172.
8. THE ARIC INVESTIGATORS. (1989). The Atherosclerosis Risk in communit (Aric) study: Design and objectives. American Journal of Epidemiology, 129(4),
9. Sulaiman, N., Mahmoud, I., Hussein, A., Elbadawi, S., Abusnana, S., Zimmet, P., & Shaw, J. (2018). Diabetes risk score in the United Arab Emirates: a screening tool for the early detection of type 2 diabetes mellitus. In BMJ Open Diabetes Research & Care (Vol. 6, Issue 1, p. e000489). BMJ. https://doi.org/10.1136/bmjdrc-2017-000489
10. Mohan, V., Deepa, R., Deepa, M., Somannavar, S., & Datta, M. (2005). A simplified Indian Diabetes Risk Score for screening for undiagnosed diabetic subjects. The Journal of the Association of Physicians of India, 53, 759–763.
11. Bevans, R. (2023, June 22). *Akaike Information Criterion | When & How to Use It (Example)*. Scribbr. Retrieved January 23, 2024, from https://www.scribbr.com/statistics/akaike-information-criterion

Author CV

Suruchi Dive (Pimple) is a Research Scholar at School of Science, G.H.Raisoni University, Saikheda, Madhya Pradesh. She is working an Assistant Professor at Sadabai Raisoni Women's College, Nagpur, India. She has an experience in teaching of 19 years. Her areas of interest are Artificial Intelligence, Machine Learning and Biomedical sciences. She is a life member of the CSI.

Dr.Gopal Sakarkar is working as an Associate Professor at Dr.Vishwanath Karad MIT World Peace University, Pune, India. He has done his PhD in Computer Science and Engineering. His expertise lies in the domain of Artificial Intelligence, Machine Learning and Data Science. He has filed seven patents and over 50 publications in reputed international journals and conferences. He has authored three textbooks and published three book chapters. He currently holds the position of publication chair at 10th International Conference (ICSCC 2024) scheduled to take place in Bali,Indonesia. He is a life member of ISTE(Indian Society of Technical Education).