International Journal of Diabetes in Developing Countries
https://doi.org/10.1007/s13410-025-01531-9

ORIGINAL ARTICLE

Machine learning identifies waist-height ratio (WHtR) as the strongest determinant of diabetes and prediabetes in children and adolescents: A comprehensive national nutrition survey

Kirti Chauhan[5] · Thozhukat Sathyapalan[1] · Usman Malabu[4] · Shashank Rameshchandra Joshi[2] · Shri Kant Singh[3] · Harshal Deshmukh[4]

Received: 9 October 2024 / Accepted: 1 July 2025

Abstract

Background Given the increasing incidence of prediabetes and type 2 diabetes (T2D) in the adolescent population in India, it is essential to identify the risk factors associated with these conditions. Understanding the risk factors associated with prediabetes and T2D can lead to timely interventions to prevent and potentially avert long-term health complications.

Objective This study aims to use machine learning algorithms to identify the best anthropometric and demographic characteristics associated with prediabetes and diabetes in Indian children and adolescents ages 10–19.

Methods The study utilizes the Comprehensive National Nutrition Survey conducted in 2016–2018 in India. The study sample includes children and adolescents aged 10–19 years. We used nine supervised machine learning algorithms to classify, assess, and identify the best model for ascertaining the risk of diabetes among adolescents in India. Various indices were used to evaluate the classification algorithms, such as the 'accuracy score', 'F1 score', 'recall score', 'precision score', and 'area under the curve' (i.e., AUC). Results were obtained based on the model with higher precision and accuracy in predicting the risk of diabetes among study subjects. Cutoff points for prediabetes were between 5.7 and 6.4 mmol/l and diabetes greater than 6.4 mmol/l.

Results The study comprised 12,318 children and adolescents (6333 males and 5985 females). The prevalence of diabetes and prediabetes in the study population was 11% ($n=1888$), while the prevalence of diabetes alone was 0.6% ($n=233$). WHtR was the most crucial feature in predicting prediabetes/diabetes, with an optimum cutoff of 0.62, a sensitivity of 0.93, and an AUC of 0.79.

Conclusions The findings derived from our machine learning analysis underscore the significance of WHtR as a cost-effective and valuable tool for diabetes and prediabetes screening among adolescents in India.

Keywords Waist-height ratio · Diabetes · Prediabetes · Machine learning · Random forest

✉ Harshal Deshmukh
harshal.deshmukh@jcu.edu.au

Kirti Chauhan
kirti070197@gmail.com

Thozhukat Sathyapalan
thozhukat.sathyapalan@hyms.ac.uk

Usman Malabu
usman.malabu@jcu.edu.au

Shashank Rameshchandra Joshi
shashank.sr@gmail.com

Shri Kant Singh
sksingh31962@gmail.com

[1] Department of Diabetes, Endocrinology and Metabolism, Allam Diabetes Centre, Hull University Teaching Hospitals, Hull HU3 2PA, UK

[2] Department of Endocrinology, DM, FACE, MD, FRCP, FACP, Joshi Clinic, Mumbai, Maharashtra, India

[3] Department of Survey Research and Data Analytics, International Institute for Population Sciences, Mumbai, Maharashtra, India

[4] James Cook University, Australia, Townsville, QLD, Australia

[5] Translational Research and Biostatistics, All India Institute of Ayurveda, New Delhi, India

Published online: 30 July 2025

Introduction

Diabetes is a leading cause of mortality and morbidity, affecting millions worldwide, including individuals as young as children and adolescents. The Centers for Disease Control and Prevention (CDC) estimated that the number of individuals under the age of 20 years, the prevalence of type 2 diabetes (T2D), has increased by 95% from 2001 to 2017 alone in the USA [1]. At the same time, roughly 41,600 new cases were diagnosed with T2D among children and adolescents in 2021 worldwide. The burden of diabetes has increased considerably in India over the last three decades, with 77 million diabetics in 2019 [2]. A remarkable figure of this increase in the incidence of diabetes in India is that one third of all the new cases being diagnosed in the adolescent population [3].

Early-onset T2D, generally acknowledged as T2D when diagnosed as early as below the age of 20 years, is an emerging chronic condition in children and adolescents [4]. Such early onsets can lead to significant and hazardous health consequences in the long run, leading to coronary heart diseases, kidney failure, etc., hence affecting an individual's quality and quantity of life [5, 6]. Given the prevailing condition and the increasing burden of T2D among adolescents with unalterable consequences it causes, it becomes crucial to understand the condition vertically (i.e., in a more profound manner) before it spreads its arms horizontally.

Finding the risk factors that are linked to prediabetes and type 2 diabetes is crucial, as these conditions are becoming increasingly prevalent in India's adolescent population. Understanding the associated risk factors can lead to timely interventions to prevent or mitigate these conditions, lessening the strain on healthcare systems and potentially averting long-term health complications. Furthermore, such insights drive research and policy initiatives to combat the diabetes epidemic effectively, ultimately contributing to healthier futures for adolescents and society. Prior investigations into the risk factors for diabetes among Indian adolescents have been hindered by small sample sizes and inadequate statistical modeling [7–12].

The present study used machine learning classifiers such as Decision Tree, Random Forest, Gradient Boosting Machine (GBM), Support Vector Machine (SVM), Naïve Bayes (NB), Logistic Regression, K-Nearest Neighbour (KNN), and ensemble learning to identify risk factors associated with diabetes and prediabetes in the adolescent population in India [13–24]. Therefore, the objective of this study was to use machine learning algorithms to obtain the relative influence of all plausible risk factors in predicting prediabetes/diabetes and obtain optimum cutoffs for the most important anthropometric features using the best-performing classification algorithm.

Materials and Methods

Data

The present study utilized the "Comprehensive National Nutrition Survey (CNNS)", a nationally representative survey dataset conducted in 2016–2018. This survey was carried out under the supervision of the Ministry of Health and Family Welfare (MOHFW), Government of India (GOI), in partnership with UNICEF, and conducted by the Population Council India. The CNNS represents the first nationwide survey in India focusing on the nutritional status of children and adolescents. Additionally, it includes data on NCDs among individuals aged 10 to 19 years. The survey dataset is available in the public domain and can be obtained upon reasonable request from the Population Council India. The report can be downloaded from the website: https://www.unicef.org/india/media/2646/file/CNNS-report.pdf.

Final study population after data cleaning and data pre-processing

The study population for the present research included children and adolescents aged 10–19 years, as per the definition by the World Health Organization (WHO) [25]. Therefore, a sample of 12,318 adolescents aged 10–19 years, with 6333 male and 5985 female adolescents, was used to measure HbA1c levels.

Derivation of the outcome variable

Diabetes

Adolescents with prediabetes/diabetes were determined by using glycosylated hemoglobin (HbA1c) (WHO,2011). By the American Diabetes Association's recommendation, glycated hemoglobin (HbA1c) can be used as an alternative to fasting blood glucose for diagnosing diabetes [9, 26]. Cutoff points will be prediabetes (lying between 5.7 mmol/l and 6.4 mmol/l) and diabetes (greater than 6.4 mmol/l) [27]. The outcome variable was then categorized into 0 "no prediabetes or diabetes" and 1 "prediabetes/diabetes."

It may be noted that, considering the age group under study (i.e., children and adolescents), the prevalence of diabetes is naturally low, and most glycemic abnormalities present as prediabetes. This reflects the early stages of metabolic dysregulation in this age group, where prediabetes and diabetes represent a continuum rather than discrete categories. Therefore, combining them into a single outcome group allowed for a more meaningful analysis of risk patterns and early predictors of glycemic abnormalities in these ages.

Derivation of features used for evaluation of different classification algorithms

Obesogenic diet (Diet type) The obesogenic diet is based on seventeen questions asked to the respondents about what and how frequently they eat a specific group of food items in a week. In CNNS, 17 questions about dietary practices were asked: how frequently they are consuming a particular food group, which includes cereals/milk/pulses/greens/roots/vegetables/fruits/eggs/fish/chicken/nuts and oilseeds/fats and oils/sugar jaggery/fried foods/junk foods/sweets/aerated drinks. Those who consume fried/aerated/junk food more frequently consider an obesogenic diet. In contrast, those who consume less frequently, say cereal, were considered an obesogenic diet; lastly, after recoding them in a unidirectional manner, where "0" corresponds to those who reported "frequently consume, say, cereal," a good eating habit. Regarding junk/sweet/fried food, "0" for those who reported "never consume junk food," which is good. Similar re-coding was done for the remaining 15 items, and the k-means clustering algorithm was then used after checking the clustering tendency or randomness of the data [28].

Overweight/Obesity Body Mass Index (BMI)-for-age was calculated and expressed as a z-score, which was used in its continuous form for the analysis [29].

Lipid anomaly Lipid profile, which includes serum Triglycerides (assessed by spectrophotometry and enzymatic endpoint method in CNNS), was categorized into "low" (< 130 mg/dl) and "high/borderline" (≥130 mg/dl), TC was recoded as "low" (< 200 mg/dl) and "high/borderline" (≥200 mg/dl) (assessed by spectrophotometry using cholesterol oxidase esterase peroxidase), LDL-C (assessed by spectrophotometry and direct measure cholesterol oxidase) was categorized as "low" (< 130 mg/dl) and "high/borderline" (≥130 mg/dl), and HDL-C (assessed by spectrophotometry and direct measure polyethylene glycol-modified cholesterol oxidase) was recoded as "borderline/low" (< 40 mg/dl) and "high" (≥40 mg/dl) [30]. Fasting blood samples were collected from the selected study sample after verbal and written consent was obtained from the respondents' parents if the respondents were under 18 years old.

Further, lipid anomalies were derived from the combination of whether an individual suffers from either high TC, LDL-C, triglycerides, or low HDL-C "1": were categorized as "1": any lipid anomalies else "0": no lipid anomalies [5].

Micronutrient levels Vitamin A deficiency will be categorized as "yes" (serum retinol concentration <20 µg/dl) or "no". Vitamin B12 deficiency will be categorized into "yes" (serum vitamin B12 < 203 pg/ml) or "no". Vitamin D deficiency will be categorized into "yes" (serum 25 (OH) concentration <12 ng/ml (30 nmol/l)) or "no". Folate deficiency will be categorized into "yes" (serum erythrocyte folate < 151 ng/ml) or "no". Iron deficiency will be categorized into "yes" (serum ferritin <15 µg/l) or "no". Zinc deficiency will be categorized into "yes" (Serum zinc concentration < 70 µg/dl (morning fasting) and < 66 µg/dl (morning non-fasting) in non-pregnant females and < 74 µg/dl in males). Iodine status will be categorized into "Adequate" (median urinary iodine concentrations (mUIC) > 100 µg/l and < 300 µg/l) and "Suboptimal" (mUIC < 50 µg/l) [29].

Other variables The age considered was considered as a continuous variable, and it was 10–19 years. Information on stunting, i.e., low height-for-age (HAZ) was categorized as "yes" if it was less than negative two standard deviations [29]. Sex as "male" and "female". Place of residence as "urban" and "rural". Religion is categorized into "Hindu', "Muslim", and "Others"; Caste has been categorized into "Scheduled Caste/Scheduled Tribe", "Other backward Class", and "Others". Further, the wealth index, a composite score computed based on various household amenities and income variables and categorized into 'Poorest", "Poor", "Middle", "Rich", and "Richest", which was already provided in the dataset. Lastly, India, has been divided into six regions "North" (The states of Chandigarh, Delhi, Haryana, Himachal Pradesh, Jammu and Kashmir, Punjab, Ladakh, Uttarakhand), "Central" (Chhattisgarh, Madhya Pradesh, Uttar Pradesh,), "East" (Bihar, Jharkhand, Odisha, West Bengal), "Northeast" (Assam, Arunachal Pradesh, Manipur, Meghalaya, Mizoram, Nagaland, Sikkim, Tripura) "West" (Gujarat, Dadra and Nagar Haveli, Daman and Diu, Maharashtra, Goa, Rajasthan), "South" (Andhra Pradesh, Karnataka, Kerala, Tamil Nadu, Puducherry, Andaman and Nicobar Islands, Telangana). All the mentioned features were considered for predicting prediabetes/diabetes among the study sample.

Data analyses

Data pre-processing

The data pre-processing involved the steps mentioned in Fig. 1. It started with cleaning the data, removing any missing values, and, where required, imputing missing values with the mean. It also involved performing feature scaling and finally upsampling the sample size by balancing the classes in the diabetes variable.

The study employed a dataset split into a training dataset (70% of the sample) and a testing dataset (30% of the sample) to develop models for identifying risk factors of prediabetes and type 2 diabetes (T2D) among adolescents in India. The upsampling technique used was random upsampling of the minority class using the resample() function in scikit-learn's

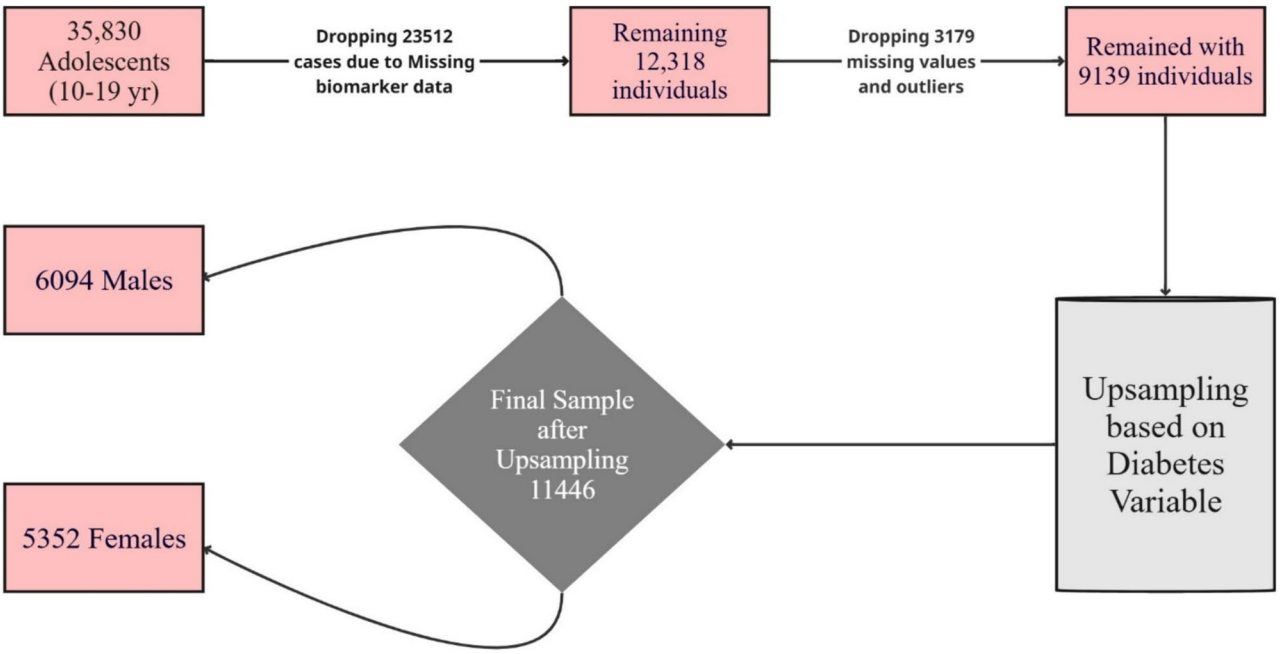


Fig. 1 Workflow data pre-processing

resample() function. The cleaned dataset comprised 9139 observations, with 1357 (unweighted 14.85%) belonging to the minority class (individuals with diabetes) and 7782 to the majority class (non-diabetics). The minority class was upsampled to 3664 instances, resulting in a new total of 11,446 records. This corresponds to an after-upsampling minority class proportion of 32.01%. Therefore, the minority class was upsampled from 1357 to 3664 instances (i.e., roughly 14% to 32% of the diseases). This level of upsampling was chosen to improve class balance while limiting the risk of overfitting associated with excessive duplication.

Model training and validation

To optimize model hyperparameters and assess performance, we employed fivefold cross-validation (cv=5) using scikit-learn's GridSearchCV on the training dataset for hyperparameter tuning (i.e., utilizing the best parameters for modelling). The final model's performance was then assessed on the remaining 30% test dataset, which was not used during the training or tuning stages.

Various classification algorithms were employed for this purpose, which are Decision Tree, Random Forest, Logistic regression, Support Vector Machines, K-Nearest Neighbours (KNN), Naïve Bayes, Gradient Boosting Classification, XGBoost, and AdaBoost. Detailed descriptions of all the machine learning algorithms are listed in supplementary materials. The study assessed each algorithm's performance using accuracy, precision, recall, and F1-score, utilizing a confusion matrix to evaluate the models' effectiveness. The chosen algorithm, determined based on performance evaluation, was used for final predictions. Furthermore, the Receiver Operating Characteristic curve (AUC-ROC) was employed to measure overall model performance by calculating the area under the curve. Lastly, the study identified optimal cutoffs and key anthropometric features using the AUC-ROC approach with the selected classification algorithm. Since the CNNS dataset employs a complex sampling design, national weights for biomarkers have been used in the exploration and computational analysis. STATA(SE) version 16.0 software has been used for data wrangling, and "sklearn" library in python was used to model various machine learning classifiers.

Supplementary material

All the machine learning classification algorithms that were used in the study, such as Decision Tree, Random Forest, Logistic regression, Support Vector Machines, K-Nearest Neighbours (KNN), Naïve Bayes, Gradient Boosting Classification, XGBoost, and AdaBoost, are described in the supplementary material. Furthermore, all the details regarding the performance matrix used to evaluate the model's effectiveness, including accuracy, precision, recall, and F1-score, are also described in the material provided in the article. Variation inflation factor (VIF) values are also reported to

Table 1 Characteristics of the study sample and the prevalence of pre-diabetes and diabetes among children and adolescents aged 10-19 years in India (CNNS 2016-18)

Demographic variables	Unweighted frequency (proportion) Weighted proportions (95% CIs)	No diabetes	Prediabetes + diabetes	Chi-square (*p* values)
Gender				0.006
Male	6333 (51.41)	50.01 (49.03, 50.99)	57.97 (55.19, 60.70)	
Female	5985 (48.59)	49.99 (49.01, 50.97)	42.03 (39.30, 44.81)	
Place of residence				0.734
Urban	5296 (42.99)	25.84 (24.99, 26.72)	26.90 (24.49, 12.63)	
Rural	7022 (57.01)	74.16 (73.28, 75.51)	73.10 (70.55, 75.51)	
Wealth index				0.037
Poorest	984 (7.99)	17.83 (17.09, 18.59)	19.89 (17.76, 22.22)	
Poor	1587 (12.88)	20.52 (19.73, 21.32)	20.36 (18.20, 22.70)	
Middle	2442 (19.82)	21.07 (20.28, 21.88)	21.12 (18.92, 23.49)	
Rich	3268 (26.53)	21.11 (20.32, 21.92)	17.68 (115.65, 19.91)	
Richest	4037 (32.77)	19.48 (18.72, 20.27)	20.95 (18.77, 23.32)	
Caste				< 0.001
Schedule caste/scheduled tribes	4665 (37.87)	32.41 (31.50, 33.34)	36.61 (33.96, 39.34)	
Other backward castes (OBCs)	3840 (31.17)	42.25 (41.29, 43.23)	34.02 (31.43, 36.72)	
Others	3813 (30.95)	25.34 (24.49, 26.20)	29.37 (26.89, 31.98)	
Region				< 0.001
North	2868 (23.28)	12.91 (12.26, 13.58)	10.84 (9.22, 12.70)	
Central	1051 (8.53)	28.73 (27.85, 29.63)	17.03 (15.04, 19.24)	
East	2445 (19.85)	30.21 (29.31, 31.12)	37.95 (35.28, 40.70)	
North-East	1300 (10.55)	10.19 (9.61, 10.80)	18.45 (16.38, 20.72)	
West	1702 (13.82)	15.49 (14.80, 16.22)	11.94 (10.24, 13.87)	
South	2952 (23.96)	2.47 (2.19, 2.80)	3.79 (2.85, 5.01)	
Religion				< 0.001
Hindu	8882 (71.11)	82.63 (81.87, 83.36)	78.02 (75.62, 80.25)	
Muslim	1403 (11.39)	13. 46 (12.81, 14.15)	17.57 (15.54, 19.80)	
Others	2033 (16.50)	3.91 (3.55, 4.31)	4.41 (3.40, 5.71)	
Diet type				< 0.001
Comparatively healthy	2368 (19.22)	18.08 (17.34, 18.85)	21.81 (19.59, 24.21)	
Plant-based	2434 (19.76)	20.52 (19.73, 21.32)	22.65 (20.39, 25.07)	
Obesogenic diet	2650 (21.51)	24.59 (23.20, 24.88)	21.59 (19.38, 23.98)	
Western	2137 (17.35)	17.53 (16.79, 18.29)	15.39 (13.48, 17.52)	
Convenient diet	2729 (22.15)	19.84 (19.07, 20.63)	18.56 (16.49, 20.83)	
Anthropometric variables	**Total**	**No diabetes**	**Prediabetes + diabetes**	***t* test (*p* values)**
	Mean (standard deviation)			
Age (in years)	14.26 (2.77)	14.29 (2.78)	14.09 (2.71)	0.0028
Waist circumference (cm)	63.95 (9.00)	63.91 (8.92)	64.20 (9.41)	0.1845
Height (cm)	150.92 (13.14)	151.08 (12.98)	150.02 (14.01)	0.0013
Weight	41.94 (12.10)	41.90 (12.04)	42.15 (12.51)	0.3967
WHtR (waist-to-height ratio)	0.52 (0.67)	0.52 (0.66)	0.55 (0.73)	0.0654
BMI (kg/m^2)	18.06 (3.51)	18.01 (3.48)	18.29 (3.66)	0.0015
Mid-upper arm circumference (cm)	21.99 (3.58)	21.98 (3.56)	22.08 (3.66)	0.2413
Triceps skinfold thickness (cm)	1.06 (0.50)	1.06 (0.50)	1.08 (0.51)	0.1831
Subscapular skinfold thickness (cm)	1.07 (0.50)	1.06 (0.50)	1.09 (0.51)	0.1831

Table 1 (continued)

Demographic variables	Unweighted frequency (proportion) Weighted proportions (95% CIs)	No diabetes	Prediabetes + diabetes	Chi-square (*p* values)
Clinical variables (continuous)	**Total**	**No diabetes**	**Prediabetes + diabetes**	***t* test (*p* values)**
	Mean (standard deviation)			
Average systolic BP	111.87 (9.77)	111.67 (9.88)	113.01 (9.02)	< 0.001
Average diastolic BP	72.94 (7.35)	72.86 (7.45)	73.41 (6.76)	0.0027
Total cholesterol	141.36 (31.85)	140.64 (31.28)	145.36 (34.60)	< 0.001
Triglycerides	96.04 (48.31)	94.76 (46.36)	103.11 (57.43)	< 0.001
Low-density lipoprotein	84.40 (23.76)	84.00 (23.48)	86.64 (25.13)	< 0.001
High-density lipoprotein	47.28 (10.24)	47.17 (10.15)	47.86 (10.70)	0.0072
Clinical variables (categorical)	**Unweighted frequency (proportion)**	**No diabetes**	**Prediabetes + diabetes**	**Chi-square (*p* values)**
Anemia				0.006
Non-anemic	9280 (75.34)	72.29 (71.40, 73.16)	69.20 (66.56, 71.72)	
Mild	1808 (14.68)	17.44 (16.70, 18.20)	17.67 (15.64, 19.91)	
Moderate	1105 (8.97)	9.40 (8.84, 9.99)	11.22 (9.58, 13.11)	
Severe	125 (1.01)	0.88 (0.71, 1.08)	1.91 (1.28, 2.85)	
Any lipid anomaly	8661 (70.31)	68.93 (68.02, 69.84)	72.21 (69.64, 74.64)	
Stunting	2760 (22.41)	24.89 (24.05, 25.75)	24.42 (22.10, 26.90)	0.255
Folate deficiency	3831 (31.10)	30.31 (29.41, 31.22)	34.90 (32.29, 37.61)	0.443
Zinc deficiency	3738 (30.35)	26.05 (25.20, 26.92)	27.80 (25.36, 30.37)	0.005
Iron deficiency	2147 (17.43)	16.06 (15.35, 16.80)	16.47 (14.50, 18.65)	0.017
Iodine deficiency	433 (3.52)	4.54 (4.14, 4.96)	4.26 (3.27, 5.55)	0.648
Vitamin A deficit	1054 (8.56)	10.05 (9.49, 10.66)	8.77 (7.31, 10.48)	0.226
Vitamin B12 deficiency	2461 (19.98)	25.01 (24.17, 25.88)	18.94 (16.84, 21.22)	< 0.001
Vitamin D deficiency	3425 (27.80)	21.93 (21.13, 22.76)	20.89 (18.71, 23.26)	0.541
Folic supplement	1450 (11.77)	9.37 (8.81, 9.96)	10.77 (9.15, 12.62)	0.091
Multivitamin supplement				0.236
Never	10,829 (87.91)	91.10 (90.52, 91.64)	91.01 (89.28, 92.49)	
Occasionally	465 (3.77)	2.83 (2.53, 3.18)	3.17 (2.32, 4.31)	
Weekly	590 (4.79)	3.80 (3.44, 4.19)	329 (2.43, 4.45)	
Daily	434 (3.52)	2.27 (2.00, 2.59)	2.53 (1.79, 3.57)	
Total	**12,318 (100)**	**88.98 (88.39, 89.55)**	**11.02 (10.04, 11.61)**	

Continuous variables: *** if *p* value for t test <0.001; ** if *p* value for t test <0.01; * if *p* value for t test <0.5 Categorical variable: *** if *p* value for chi-squared-test <0.001; ** if *p* value for chi-squared test <0.01; *** if *p* value for chi-squared test <0.05

demonstrate the collinearity between the features used in the model (Table S1). Confusion matrix obtained from various classifiers has been reported in Table S2.

Results

Figure 1 shows the study flow diagram. The study consisted of 35,830 adolescents, of whom 12,318 had HbA1c levels. After removing outliers, implausible data, and other missing data for covariates, we had complete data on 7865 adolescents, comprising 5842 males and 5150 females. We also unsampled our data to handle the imbalanced information on the diabetes variable.

Table 1 shows the demographic characteristics of the study population and a comparison of those without diabetes/prediabetes and those living with diabetes/prediabetes. The prevalence of diabetes and prediabetes in the study population was 11% (n=1888), while the prevalence of diabetes alone was 0.6% (n=233). Those with prediabetes and diabetes exhibited significantly higher BMI (p=0.001) and elevated systolic blood pressure (p<0.0001). Moreover, they displayed an unfavorable lipid profile, characterized by elevated triglycerides (p<0.001) and low-density lipoproteins (p<0.001). In addition to these health markers,

Table 2 Performance metrics for different classification algorithms in predicting pre-diabetes/diabetes

Classification	Accuracy	Precision	Recall	F1-score
Decision tree	0.71	0.55	0.42	0.48
Random forest	0.95	0.99	0.85	0.91
Logistics regression	0.69	0.56	0.80	0.14
Support vector machine	0.96	0.99	0.89	0.94
KNN classifier	0.84	0.68	0.93	0.79
Naive Bayes classifier	0.66	0.46	0.31	0.37
Gradient boosting machine	0.89	0.82	0.86	0.84
XGBoost algorithm	0.90	0.89	0.77	0.82
AdaBoost algorithm	0.71	0.59	0.26	0.36

those with prediabetes and diabetes were notably deficient in essential nutrients, with deficiencies in zinc (p=0.005), iron (p=0.01), and vitamin B12 (p<0.001) as compared to the control group.

Table 2 shows the performance matrix of nine classification algorithms in predicting diabetes and prediabetes. SVM and the random forest model were top-performing models; however, their complex interpretability can sometimes become computationally expensive and memory-intensive, particularly in handling large datasets. The random forest model was chosen to obtain the results. The ROC curve for the random forest was 0.91 (Fig. 2.2). Therefore, random forest was used to obtain further results due to its computational efficiency and interpretability.

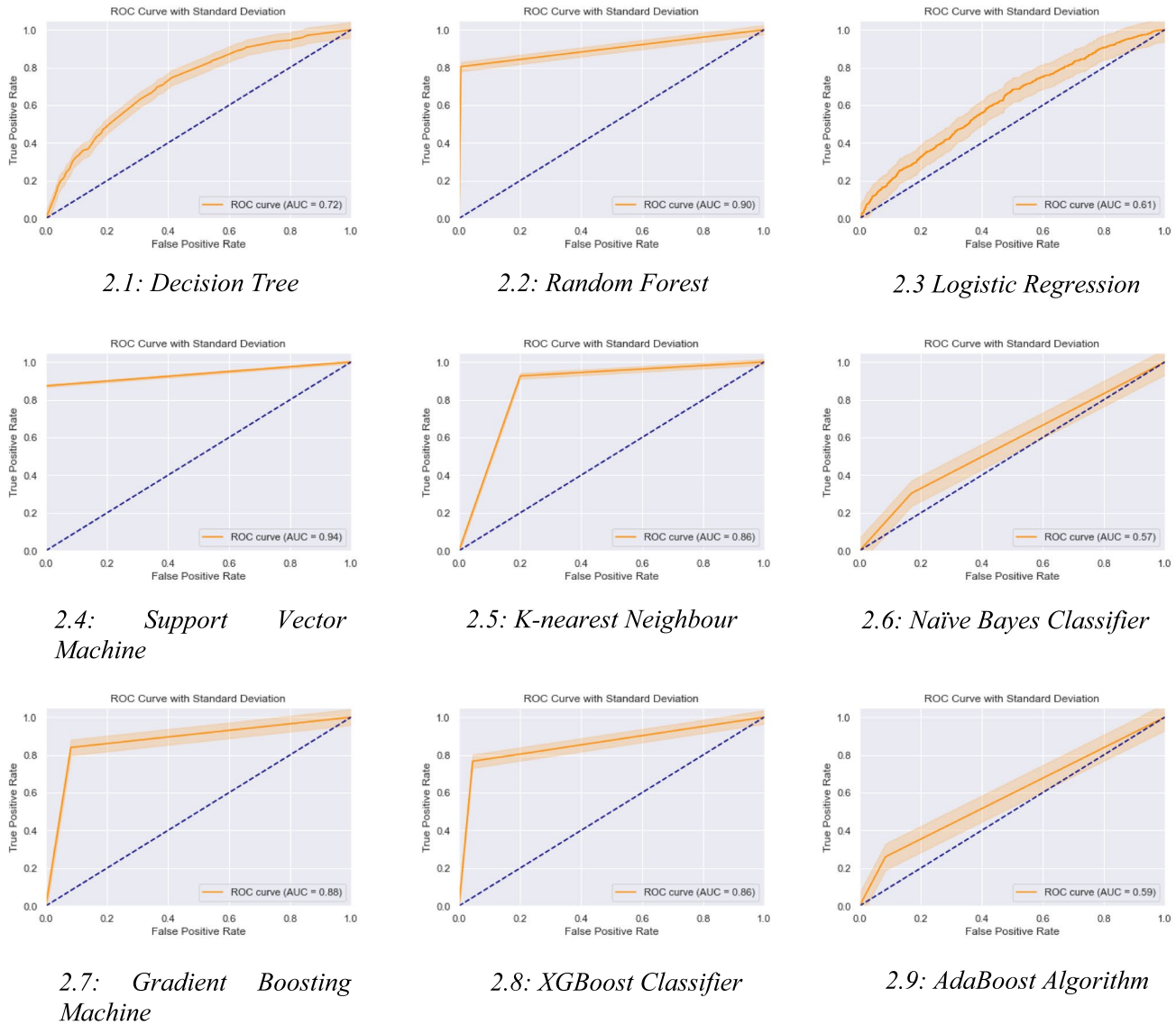


Fig. 2 ROC curves for different classification algorithms with area under the curve (AUC)

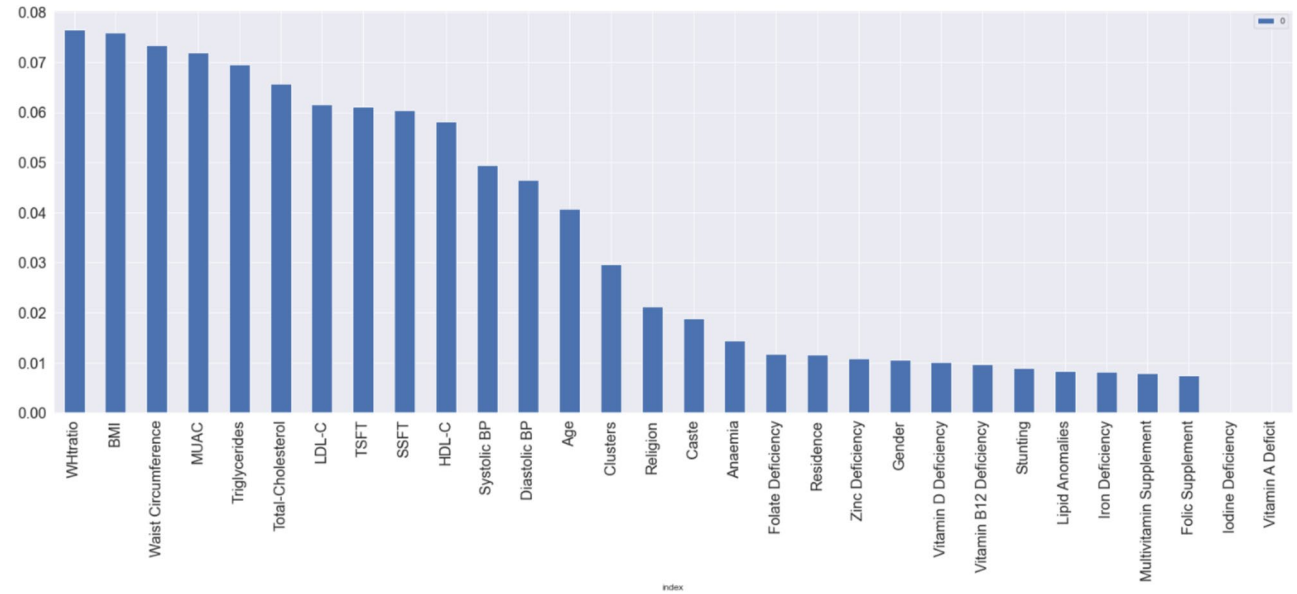


Fig. 3 Relative influence plot of selected features in predicting prediabetes/diabetes using the random forest for classification algorithm among adolescents aged 10–19 years in the sampled population in India CNNS (2016–2018)

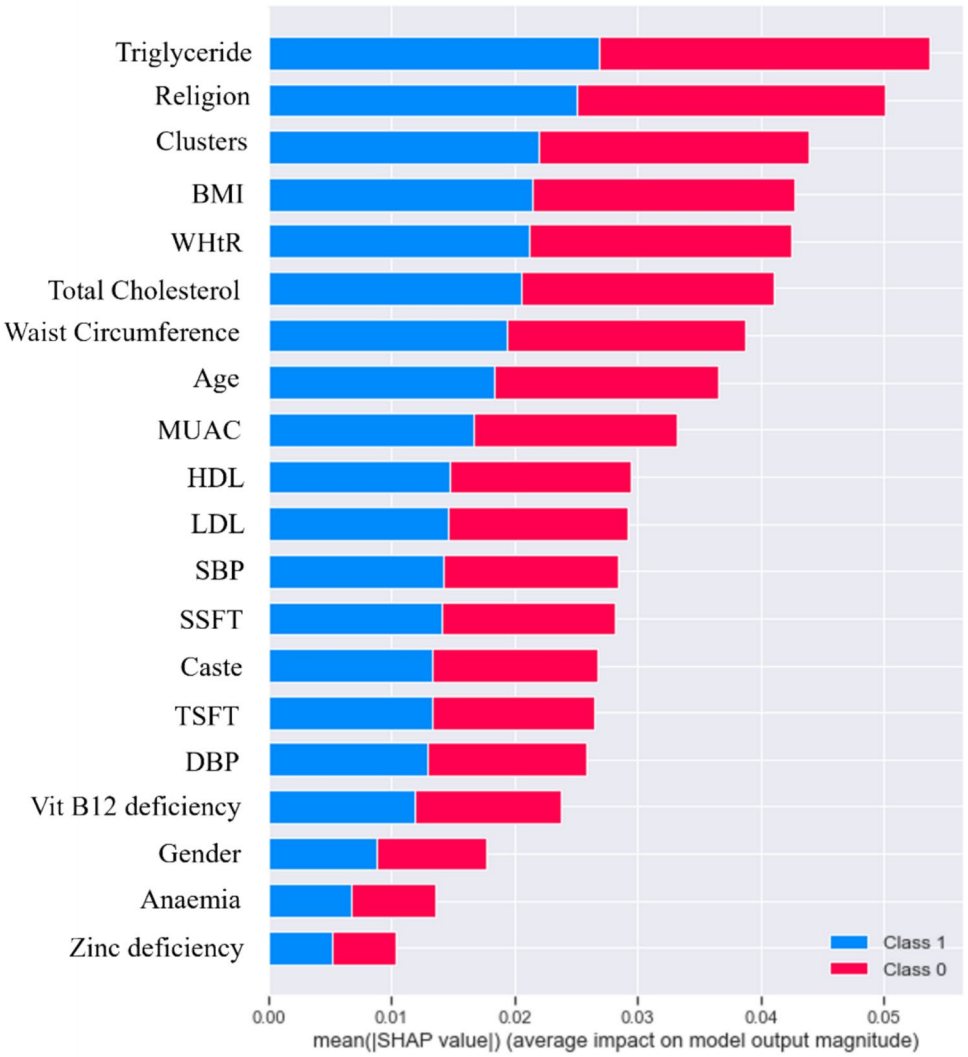
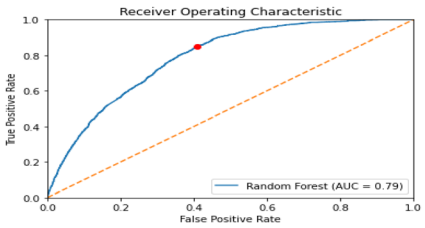
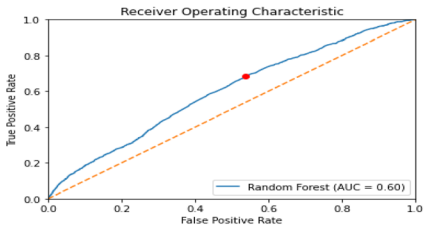


Fig. 4 Class-wise SHAP values from random forest model predicting prediabetes/diabetes among adolescents aged 10–19 years in the sampled population in India CNNS (2016–2018)

Since random forest performed comparatively better, results were obtained on the training dataset using a random forest classifier using the required Python library. The best hyperparameters, i.e., maximum depth of 30, minimum sample leaf of 1, 5 minimum sample splits, and 400 number of estimators, were used after hyperparameter tuning for the random forest. Relative influence (RI) or feature importance (FI) was obtained and plotted in Fig. 3, and SHAP values in Fig. 4. It can be seen from the table and the figure that the features that were most important or had the highest weightage were more significant in predicting prediabetes and diabetes among the study population. All the anthropometric variables were seen at the top with waist-to-height ratio (RI=0.077), BMI (RI=0.076), waist-circumference (RI=0.073), MUAC (RI=0.072), Triglycerides (RI=0.070), Total cholesterol (RI=0.066), LDL (RI=0.062), TSFT (RI=0.061), SSFT (RI=0.060), HDL (RI=0.058), average systolic BP (RI=0.049), average diastolic BP (RI=0.046), and age (RI=0.41). Lastly, the cluster regarding their eating habits has an importance of 0.029 (Fig. 3). From Fig. 4, it may also be noted that the waist-to-height ratio and BMI are among the top five contributing features in predicting prediabetes/diabetes.

Since WHtR and BMI were the most important features in predicting prediabetes/diabetes, the optimum cutoffs were obtained. Estimated optimum cutoff values for WHtR and BMI were obtained as 0.62 and 34.89, respectively. That is, adolescents with WHtR greater than or equal to 0.62 can be anticipated to have a higher risk for prediabetes/diabetes. Similarly, individuals with a BMI greater than or equal to 34.89 can be anticipated to have a higher risk of developing prediabetes/diabetes. The diagnostic performance shows that the AUC was 0.79 and 0.60, and sensitivity was obtained as 0.93 and 0.68 for WHtR and BMI, respectively. It is worth noting that due to the high sensitivity and low specificity, the trade-off between the two may result in a high number of false positives. However, the proposed cutoff may be utilized for initial-level screening and might provide some insights into the imbalanced glucose levels in individuals. Therefore, final diagnostics will be done based on a proper medical examination(Table 3).

Table 3 Diagnostic performance of BMI and waist-to-height ratio in detecting pre-diabetes/diabetes using optimal BMI and waist-to-height ratio cut-off values based on the shortest distance in ROC curves in adolescents aged 10-19 years in India (2016-18)

	WHtR	BMI
AUC	0.79	0.60
Optimum threshold	0.15	0.15
Optimal Cut-off	0.62	34.89
Sensitivity	0.93	0.68
Specificity	0.47	0.46
Youden's J	0.40	0.14
PPV	0.24	0.19
NPV	0.97	0.89
LR+	1.77	1.27
LR-	0.14	0.46
FPR	0.53	0.54
FNR	0.07	0.32
Misclassification rate	0.46	0.50
ROC-AUC (with optimal threshold)		

Discussion

In this machine learning analysis of a large population-based study, we show that WHtR is the strongest determinant of diabetes and prediabetes in children and adolescents in India. We also show that besides the anthropometric parameters, a deranged lipid profile, mainly raised triglycerides, was a strong marker of prediabetes and diabetes in the Indian population.

The leftward shift in the age distribution of type 2 diabetes in India signifies a significant and concerning trend. Traditionally, type 2 diabetes has been more prevalent in older age groups, but recent years have seen a notable increase in its incidence among younger individuals, including children and adolescents. This shift suggests that a growing number of Indians are being diagnosed with type 2 diabetes at an earlier stage of life, likely because of changes in lifestyle, dietary patterns, reduced physical activity, and the increasing prevalence of obesity.

We used nine machine learning algorithms and explored the association of 28 demographic, anthropometric, and clinical variables with prediabetes and diabetes. The best-performing machine learning model predicted WHtR as the best predictor of diabetes and prediabetes in the Indian population, with an optimal cutoff of 0.62. Furthermore, in the study population, it performs better in predicting diabetes and prediabetes than BMI and waist circumference. Several studies have identified WHtR as a valuable predictor for prediabetes and diabetes risk in children and adolescent populations [31–33]. Additionally, WHtR correlates strongly with insulin resistance and metabolic syndrome, both precursors to prediabetes and diabetes [32]. Moreover, WHtR has higher accuracy in predicting prediabetes than other anthropometric indicators like BMI and waist circumference [33]. However, it is imperative to acknowledge that small sample sizes have limited previous studies, which may have resulted in inadequately confounding factors and/or not been conducted in Indian populations. To address these limitations, our research leveraged a large nationwide dataset and employed various machine-learning algorithms to demonstrate the robustness of the waist-to-height ratio (WHR) as a valuable tool for predicting prediabetes and diabetes. This approach underscores the practicality and efficiency of WHtR in the early identification of individuals at risk, facilitating targeted interventions to combat the increasing prevalence of prediabetes and diabetes among Indian children and adolescents.

Our machine learning analysis also shows the association of elevated levels of LDL-C and triglycerides associated with prediabetes and diabetes in children and adolescents in the Indian population. Earlier studies in these populations have highlighted the high prevalence of a deranged lipid profile in this population [5, 9, 34]. These findings emphasize the crucial role of monitoring and addressing dyslipidemia in diabetes management and prevention in the Indian adolescent and teen population. However, no external validation has been done, which can be a future scope of this study.

Conclusions

In summary, our study represents the first of its kind in India, delving into the efficacy of multiple classification algorithms for predicting prediabetes and diabetes in children and adolescents. The findings derived from our machine learning analysis underscore the significance of WHR as a cost-effective and valuable tool for diabetes and prediabetes screening. The implications of these results are particularly crucial for public health service providers. Accurate predictions regarding the health status of children and young adults enable identifying individuals at higher risk, facilitating the timely provision of essential services to mitigate potential health consequences. This proactive approach holds the potential to avert adverse health outcomes that could otherwise manifest if not diagnosed promptly.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s13410-025-01531-9.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Declarations

Ethical Clearance The present study utilizes a secondary dataset from the Comprehensive National Nutrition Survey (CNNS) conducted in 2016-18. The dataset is available in the public domain for legitimate research purposes, with no identifiable information on the survey participants. Therefore, no additional ethical approval is required.

Consent of Patient Before the survey, informed consent was obtained from the parents/caregivers for adolescents aged 11-17 years, and assent was obtained from the adolescents themselves aged 11-17 years.

Conflict of interests The authors declare that they have no conflict of interests.

Acknowledgment The authors gratefully acknowledged the Population Council (India) for providing the data for this study. Further, the authors are grateful and recognize the donors Aditya and Megha Mittal, MoHFW, UNICEF, and CNNS.

References

1. Lawrence JM, Divers J, Isom S, Saydah S, Imperatore G, Pihoker C, et al. Trends in prevalence of type 1 and type 2 diabetes in children and adolescents in the US, 2001–2017. JAMA [Internet]. 2021 [cited 2023 Apr 19];326(8):717–27. Available from: https://jamanetwork.com/journals/jama/fullarticle/2783420
2. Pradeepa R, Mohan V. Epidemiology of type 2 diabetes in India. Indian J Ophthalmol [Internet]. 2021 [cited 2022 Mar 23];69(11):2932–8. Available from: https://journals.lww.com/ijo/Fulltext/2021/11000/Epidemiology_of_type_2_diabetes_in_India.6.aspx
3. Prasad AN. Type 2 diabetes mellitus in young: need for early screening. Indian Pediatr. 2011;48(9):683–8.
4. Webber S. International Diabetes Federation. Vol. 102, Diabetes Research and Clinical Practice. 2013. 147–148 p.
5. Kirti K, Singh SK. Quantifying the burden of lipid anomalies among adolescents in India. BMC Cardiovasc Disord [Internet]. 2022 [cited 2022 Sep 16];22(1):1–10. https://doi.org/10.1186/s12872-022-02819-y
6. Singh SK, Chauhan K, Puri P. Chronic non-communicable disease burden among reproductive-age women in India: evidence from recent demographic and health survey. BMC Womens Health [Internet]. 2023;23(1):1–15. Available from: https://doi.org/10.1186/s12905-023-02171-z
7. Chathurvedi D, Khadgawat R, Kulshrestha B, Gupta N, Joseph AA, Diwedi S, et al. Type 2 diabetes increases risk for obesity among subsequent generations. Diabetes Technol Ther. 2009;11(6):393–8.
8. Bhalwar R. Metabolic syndrome: the Indian public health perspective. Med Journal, Armed Forces India [Internet]. 2020 [cited 2023 Apr 9];76(1):8. Available from: /pmc/articles/PMC6994803/
9. Kumar P, Srivastava S, Mishra PS, Mooss ETK. Prevalence of prediabetes/type 2 diabetes among adolescents (10–19 years) and its association with different measures of overweight/obesity in India: a gendered perspective. BMC Endocr Disord. 2021;21(1):1–12.
10. Gaidhane S, Mittal W, Khatib N, Zahiruddin Q, Muntode P, Gaidhane A. Risk factor of type 2 diabetes mellitus among adolescents from rural area of India. J Fam Med Prim Care. 2017;6(3):600.
11. Dabas A, Aravind T, Yadav S, Mantan M, Kaushik S. Are Indian obese children and adolescents at increased risk for vitamin D deficiency? Indian J Med Sci. 2021;73(3):323–6.
12. Ramesh S, Abraham RA, Sarna A, Khan N, Ramakrishnan L. Prevalence of metabolic syndrome among adolescents in India: a population-based study Population Council Population Council Akash Porwal Population Council Rajib Acharya Population Council Praween K. Agrawal IPE Global Limited Sana Ashraf Population Council. 2021 [cited 2022 Mar 24]; Available from: https://doi.org/10.21203/rs.3.rs-902711/v1
13. Silva VC, Gorgulho B, Marchioni DM, Araujo TA de, Santos I de S, Lotufo PA, et al. Clustering analysis and machine learning algorithms in the prediction of dietary patterns: cross-sectional results of the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil). J Hum Nutr Diet [Internet]. 2022 [cited 2023 Jan 4];35(5):883–94. https://doi.org/10.1111/jhn.12992
14. Lai H, Huang H, Keshavjee K, Guergachi A, Gao X. Predictive models for diabetes mellitus using machine learning techniques. BMC Endocr Disord. 2019;19(1):1–9.
15. Webb GI. Naïve Bayes. Encycl Mach Learn [Internet]. 2011 [cited 2024 May 17];713–4. https://doi.org/10.1007/978-0-387-30164-8_576
16. Yang S, Kim JK. Nearest neighbor imputation for general parameter estimation in survey sampling. 2017.
17. Santhanam T, Padmavathi MS. Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis. Procedia Comput Sci. 2015 1;47(C):76–83.
18. Uddin S, Haque I, Lu H, Moni MA, Gide E. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Sci Rep. 2022. https://doi.org/10.1038/s41598-022-10358-x.
19. Mahboob Alam T, Iqbal MA, Ali Y, Wahab A, Ijaz S, Imtiaz Baig T, et al. A model for early prediction of diabetes. Informatics Med Unlocked [Internet]. 2019;16(July):100204. Available from: https://doi.org/10.1016/j.imu.2019.100204
20. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry [Internet]. 2015 [cited 2024 May 17];27(2):130. Available from: /pmc/articles/PMC4466856/
21. Sarica A, Cerasa A, Quattrone A. Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. Front Aging Neurosci. 2017. 6;9(OCT):284242.
22. Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. A comprehensive survey on support vector machine classification: applications, challenges and trends. Neurocomputing. 2020;408(30):189–215.
23. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot [Internet]. 2013 [cited 2024 May 17];7(DEC). Available from: /pmc/articles/PMC3885826/
24. Wang R. Adaboost for feature selection, classification and its relation with SVM, a review. Phys Procedia. 2012;25(Jan 1):800–7.
25. World Health Organization. Adolescent health [Internet]. 2023 [cited 2022 Aug 3]. Available from: https://www.who.int/southeastasia/health-topics/adolescent-health
26. Sinha R, Fisch G, Teague B, Tamborlane WV, Banyas B, Allen K, et al. Prevalence of impaired glucose tolerance among children and adolescents with marked obesity. N Engl J Med. 2002;346(11):802–10.
27. IDF, Diabetes Atlas, 9th edition 2019. Diabetes Atlas 2019 [Internet]. [cited 2022 Aug 22]. Available from: https://www.idf.org/news/240:diabetes-now-affects-one-in-10-adults-worldwide.html
28. Kirti K, Singh SK. Obesogenic diet and metabolic syndrome among adolescents in India: data-driven cluster analysis. BMC Cardiovasc Disord. 2023;23(1):393.
29. Ministry of Health and Family Welfare (MoHFW), Government of India U, Council and P. Comprehensive National Nutrition Survey (CNNS) National Report. New Delhi; 2019.
30. Cleeman JI. Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III). JAMA [Internet]. 2001 [cited 2025 Feb 10];285(19):2486–97. Available from: https://pubmed.ncbi.nlm.nih.gov/11368702/
31. Chen N, Hu LK, Sun Y, Dong J, Chu X, Lu YK, et al. Associations of waist-to-height ratio with the incidence of type 2 diabetes and mediation analysis: two independent cohort studies. Obes Res Clin Pract [Internet]. 2023 [cited 2023 Nov 15];17(1):9–15. Available from: https://pubmed.ncbi.nlm.nih.gov/36586764/
32. Yoo EG. Waist-to-height ratio as a screening tool for obesity and cardiometabolic risk. Korean J Pediatr [Internet]. 2016 [cited 2023 Nov 15];59(11):425–31. Available from: https://pubmed.ncbi.nlm.nih.gov/27895689/
33. Mondal S, Gargari P, Bose C, Chowdhury S, Mukhopadhyay S. Prevalence and predictors of prediabetes in adolescents and young adults with turner syndrome: a cross-sectional study from eastern

India. Indian J Endocrinol Metab [Internet]. 2023 [cited 2024 Oct 7];27(4):335–45. Available from: https://pubmed.ncbi.nlm.nih.gov/37867982/

34. Al Amiri E, Abdullatif M, Abdulle A, Al Bitar N, Afandi EZ, Parish M, et al. The prevalence, risk factors, and screening measure for prediabetes and diabetes among Emirati overweight/obese children and adolescents. BMC Public Health [Internet]. 2015 [cited 2023 Nov 15];15(1). Available from: https://pubmed.ncbi.nlm.nih.gov/26704130/