

Prediction and Identification of Key Factors Associated with High Soft Drink Consumption Among University Students Using Machine Learning

Arzo Ahmed

arzoahmed1993@gmail.com

Mawlana Bhashani Science and Technology University

Md. Mizanur Rahman

Mawlana Bhashani Science and Technology University

Md. Deluar Jahan Moley

Mawlana Bhashani Science and Technology University

Munni Akter

Mawlana Bhashani Science and Technology University

Research Article

Keywords: Soft drinks, sugar-sweetened beverages, machine learning, feature selection, university students, Bangladesh

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10785246/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Prediction and Identification of Key Factors Associated with High Soft Drink Consumption Among University Students Using Machine Learning

Arzo Ahmed^{1,2}, Md. Mizanur Rahman^{2,3*}, Md. Deluar Jahan Moly^{3,1} and Munni Akter^{4,3}
¹arzoahmed1993@gmail.com, ²mizanur9923@gmail.com, ³deluarmoloy@gmail.com and
⁴munnist13004@gmail.com

¹Assistant Professor, Department of Statistics, Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902, Bangladesh

²Professor, Professor, Department of Statistics, Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902, Bangladesh

³Department of Statistics, Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902, Bangladesh

Corresponding Author: Arzo Ahmed

Department of Statistics, Mawlana Bhashani Science and Technology University,
Santosh, Tangail-1902, Bangladesh

E-mail address: arzoahmed1993@gmail.com

ORCID ID: 0000-0003-3445-4855

Statements and Declarations

Funding: No funding was received for conducting this study.

Conflict of interest: The authors declare that they have no conflict of interest.

Consent to participate: Informed consent was obtained from all individual participants included in the study.

Consent for publication: Not applicable.

Clinical trial number: Not applicable.

Authors' contributions: Md. Mizanur Rahman and Arzo Ahmed were responsible for the conception and design of the underlying study. Data collection was led by Taksina Kabir and Munni Akter. Conception of the paper, material and data preparation were performed by Md. Mizanur Rahman, Taksina Kabir and Munni Akter. Analyses were performed by Md. Mizanur Rahman. The first draft of the manuscript was written by Md. Mizanur Rahman. Arzo Ahmed, Taksina Kabir and Munni Akter reviewed the manuscript.

Acknowledgment: We are grateful to the participants, as well as thankful to the editors and anonymous reviewers.

Data Availability: The datasets generated during the current study are available from the corresponding author upon reasonable request.

Prediction and Identification of Key Factors Associated with High Soft Drink Consumption Among University Students Using Machine Learning

Abstract

Although soft drinks are widely consumed among young people, excessive intake has been associated with several adverse health outcomes. This study aimed to assess soft-drink intake patterns and associated factors among university students, classify consumers using machine-learning algorithms, compare model performance, and identify influential predictors. A cross-sectional study was conducted among 450 students using a structured questionnaire administered through face-to-face interviews. Univariate and bivariate analyses, logistic regression, and seven machine-learning algorithms were applied. Approximately one-third of students (33.6%) were classified as high consumers based on consumption of ≥ 1 L of soft drinks per week. Demographic, behavioral, thematic, and attitudinal variables were significantly associated with intake level. Smoking status and academic year were the most important predictors identified through Random Forest variable-importance analysis. Among the evaluated models, Random Forest demonstrated the highest accuracy (77.04%) and AUC (0.8460), with sensitivity of 95.40%, specificity of 43.75%, and precision of 75.45%. These findings highlight that high soft-drink intake is common among university students and demonstrate the potential value of machine-learning approaches for identifying predictors of consumption behavior. Targeted awareness and prevention strategies should promote healthier beverage choices.

Keywords: Soft drinks; sugar-sweetened beverages; machine learning; feature selection; university students; Bangladesh.

Introduction

Soft drinks are widely consumed non-alcoholic beverages that generally contain refined sugar or other sweeteners, edible acids, natural or artificial flavors, and occasionally fruit juice; they are commonly, but not always, carbonated. Increasing consumption of sugar-sweetened beverages among young populations has become a growing public health concern because excessive intake may contribute to increased energy consumption and adverse health consequences. The World Health Organization (WHO) recommends limiting free-sugar intake to less than 10% of total energy intake to reduce the risk of diet-related health problems (World Health Organization 2003). Evidence from a Norwegian survey of eighth-grade students showed that added sugar contributed approximately 18% of total energy intake, with soft drinks accounting for nearly 30% of refined sugar consumption (Overby et al. 2004). High-sugar beverages may impair dietary compensation, resulting in inadequate reduction of subsequent food intake compared with foods containing equivalent energy content (Rolls 1997; Ludwig 2002; Ebbeling & Ludwig 2001; Ludwig et al. 2001). Furthermore, a randomized trial demonstrated that consumption of sugar-sweetened soft drinks increased energy intake and body weight compared with energy-free beverages containing artificial sweeteners (Raben et al. 2002). These findings indicate that liquid calories from soft drinks may not be effectively compensated for through reduced food intake (Mattes 1996; Tordoff & Alleva 1990; Harnack et al. 1999; Ludwig et al. 2001).

In addition to their role in excessive energy use, frequent soft-drink consumption has been associated with several adverse health outcomes. In low-income populations and countries undergoing economic transition, extensive marketing of soft drinks and energy-dense foods may contribute to dietary changes by replacing traditional nutrient-rich foods (World Health Organization 2003). Higher consumption of sugary beverages has been associated with increased risks of obesity, metabolic disorders, and other non-communicable diseases (Trumbo & Rivers 2014; Basu et al. 2013; Popkin 2001; Harrington 2008; Vartanian et al. 2007). In addition, the high sugar and acidic content of soft drinks may contribute to dental erosion and dental caries (Heller et al. 2001; Moynihan & Petersen 2004; Johansson et al. 2001; Jones & Nunn 1995). Studies have also reported associations between carbonated soft-drink consumption and other health-risk behaviors among adolescents and young adults in low- and middle-income countries (Pengpid & Peltzer 2020). Moreover, increased soft-drink intake may displace nutrient-rich beverages such as milk, potentially reducing calcium and magnesium intake and increasing the risk of poor bone health and fractures (Rodríguez-Artalejo et al. 2003; Harnack et al. 1999).

It is crucial to comprehend the variables linked to soft drink use in order to create successful preventative measures. Studies conducted around the world reported an increase in soft drink intake among young adults (Li et al. 2014; French et al. 2003; Johansson et al. 1996). Previous research suggests that consumption patterns are influenced by multiple demographic, behavioral, socioeconomic, and attitudinal factors. Demographic characteristics, including age, sex, educational status, and socioeconomic conditions, may influence beverage choices. Several studies have reported higher soft-drink consumption among males compared with females (Bipasha et al. 2017; West et al. 2006; Rehm et al. 2008; Forshee & Storey 2003; Grimm et al. 2004; Vereecken et al. 2005). Behavioral factors, including smoking, alcohol consumption, and other lifestyle behaviors, have also been associated with increased soft-drink intake (Bere et al. 2008; Srinivasan & Elangovan 2000; Hwang et al. 2020; Maupome-Carvantes et al. 1995; Chen et al. 2020; Storey et al. 2006; Bruckauf & Walsh 2018; Narain et al. 2017; Han & Powell 2013; Pengpid & Peltzer 2020). Furthermore, individual perceptions, such as beliefs regarding health effects, taste preferences, and awareness of harmful components, may influence consumption decisions. However, associations between these factors and soft-drink consumption are not consistent across populations. For example, Sinha et al. (2021) reported differences in consumption patterns but found no significant influence of gender, age, marital status, or educational qualification, suggesting that contextual differences may affect consumption behaviors.

Previous studies among adolescents and university students have demonstrated considerable variation in soft-drink consumption patterns across countries. Studies from different regions have reported increasing consumption among young people, indicating that soft drinks have become an important component of youth dietary behavior (Li et al. 2014; French et al. 2003; Johansson et al. 1996). Differences in consumption patterns may reflect variations in cultural practices, socioeconomic conditions, availability of beverages, health awareness, and lifestyle behaviors. Among young adults, university students represent a particularly important population because increased independence, academic stress, social influence, and changing dietary habits may contribute to unhealthy beverage choices.

In Bangladesh, soft-drink intake has become increasingly common among young people (Mistry & Puthussery 2015). Among university students, Bipasha et al. (2017) reported that 95.4% consumed sugar-sweetened beverages, with approximately half consuming soft drinks on at least two days per week. Similar concerns have been highlighted in recent research among Bangladeshi university students (Rahman et al. 2023). Although Rahman et al. (2023) also investigated soft-drink consumption among students at MBSTU, their study primarily applied logistic regression as the machine-learning approach and focused on perception-related factors. In contrast, the present study uses more recent data and compares seven machine-learning algorithms KNN, LDA, CART, logistic regression, naïve Bayes, SVM, and random forest across a broader set of demographic, behavioural, and perception-related predictors to identify factors associated with high soft-drink intake. Khan and Uddin (2020) reported that 17% of adolescents consumed carbonated soft drinks more than three times daily. Furthermore, frequent soft-drink intake has been associated with overweight and obesity among Bangladeshi youth (Al Muktadir et al. 2019). However, evidence regarding the factors influencing high soft-drink intake among Bangladeshi university students remains limited.

Although previous studies have explored the prevalence and associated factors of soft-drink intake, most have relied mainly on conventional statistical approaches. Limited research has integrated predictive modelling approaches with traditional statistical methods to classify consumers, evaluate predictive performance, and identify the relative importance of factors associated with high intake. Understanding complex interactions among demographic, behavioral, and attitudinal characteristics requires analytical approaches capable of detecting nonlinear relationships and multiple predictor patterns.

Machine-learning techniques may provide additional insights by identifying complex patterns within multidimensional datasets. Compared with traditional statistical approaches that primarily focus on estimating associations, machine-learning methods can capture complex nonlinear relationships, improve classification performance, and identify influential predictors within multidimensional datasets. Therefore, applying machine-learning methods may enhance understanding of high soft-drink intake patterns among university students and support the development of targeted prevention strategies.

Therefore, this study aimed to examine soft-drink intake patterns among university students in Bangladesh, identify factors associated with high intake, classify consumers according to intake level using machine-learning algorithms, compare predictive performance among different models, and determine the most influential predictors of high soft-drink intake.

2 Materials and Methods

2.1 Study Design and Setting

A cross-sectional study was conducted at Mawlana Bhashani Science and Technology University (MBSTU), Santosh, Tangail, Bangladesh. The study population consisted of undergraduate

students enrolled in different academic departments of the university. The study was designed to assess soft-drink intake patterns, identify factors associated with high intake, and develop predictive classification models using machine-learning approaches.

2.2 Data Collection

Data were collected during the first week of December 2024 using a pre-designed, structured questionnaire containing closed-ended questions. The questionnaire collected information on demographic characteristics, habitual behaviors, and respondents' perceptions, beliefs, and concerns regarding soft drinks. A total of 450 undergraduate students from first to fifth academic year across five faculties participated in the study. Data were collected through face-to-face interviews, and participants completed the questionnaire under interviewer supervision.

2.3 Sampling Technique

A convenience sampling technique was applied for participant recruitment. Students who were available during the data-collection period and agreed to participate voluntarily were included in the study. Participants were recruited from different academic years and faculties to obtain variation in demographic and behavioral characteristics.

2.4 Study Variables

The primary outcome variable was soft-drink intake level. Participants were categorized according to the threshold proposed by Scully et al. (2017), where high soft-drink intake was defined as consumption of ≥ 1 litre per week and normal intake as < 1 litre per week. This binary classification was selected because it provides an established and reproducible approach for identifying higher levels of soft-drink consumption and enables comparison with previous research. Although the threshold was originally developed among Australian secondary-school students, it was applied in the present study as a pragmatic classification criterion due to the absence of a validated soft-drink intake cut-off for Bangladeshi university students. Future studies should consider developing and validating population-specific thresholds for Bangladeshi young adults.

Twelve independent variables were included in the analysis. These variables were grouped into four categories: demographic variables, habitual variables, attitudinal variables, and thematic variables. Demographic variables included age (D1), academic year (D2), and marital status (D3). Habitual variables included smoking status (H1) and alcohol consumption status (H2). Attitudinal variables included beliefs that artificial flavors and colors used in soft drinks are the same (C1), harmful preservatives are used in soft drinks (C2), soft drinks contain a high amount of fat (C3), and soft drinks are very tasty (C4). Thematic variables included beliefs that fresh juice and soft drinks are the same (T1), both fresh juice and soft drinks are healthy (T2), and perceived health problems associated with soft-drink intake (T3).

2.5 Machine-Learning Model Development

Seven supervised machine-learning algorithms were applied to classify participants into high and normal soft-drink intake groups. The selected algorithms represented different classification

approaches, including distance-based, linear, probabilistic, tree-based, and ensemble learning methods. The models included k-nearest neighbors (KNN), linear discriminant analysis (LDA), classification and regression tree (CART), logistic regression (LR), naïve Bayes (NB), support vector machine (SVM), and random forest (RF).

K-nearest neighbors (KNN) is a non-parametric classification algorithm that assigns observations based on the class distribution among neighboring observations in feature space (Cover & Hart, 1967). Linear discriminant analysis (LDA) is a supervised classification method that identifies linear combinations of predictors to maximize separation between predefined classes (Fisher, 1936). Classification and regression tree (CART) recursively partitions predictor variables based on impurity reduction criteria to generate decision rules for classification (Breiman et al., 1984). Logistic regression (LR) estimates the probability of binary outcomes based on predictor variables using the logistic function (Hosmer et al., 2013). Naïve Bayes (NB) is a probabilistic classification method based on Bayes' theorem and assumes conditional independence between predictors (Lewis, 1998). Support vector machine (SVM) identifies an optimal decision boundary between classes by maximizing the margin between observations of different classes (Cortes & Vapnik, 1995). Random forest (RF) is an ensemble learning method that combines multiple decision trees generated using bootstrap samples and random subsets of predictors to improve predictive stability and reduce overfitting (Breiman, 2001).

2.5.1 Model Performance Metrics

Confusion matrix:

		Actual	
		Positive	Negative
Predictive	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

The following performance metrics were used:

Accuracy: Accuracy is the proportion of all observations that are correctly classified, calculated from the total number of true-positive and true-negative predictions relative to all predictions.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Sensitivity (Recall or True-Positive Rate): Sensitivity is the proportion of actual positive cases that are correctly identified by the model.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Specificity (True-Negative Rate): Specificity is the proportion of actual negative cases that are correctly identified by the model.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Precision (Positive Predictive Value): Precision is the proportion of predicted positive cases that are true positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Area Under the Curve (AUC): The receiver operating characteristic (ROC) curve plots sensitivity against the false-positive rate (1 – specificity) across classification thresholds. The AUC summarizes a model’s ability to discriminate between the two outcome classes. Values closer to 1 indicate better discrimination, whereas a value of 0.5 indicates performance no better than chance.

2.6 Statistical Analysis and Software

Descriptive statistics were used to summarize participant characteristics and soft-drink intake patterns. Associations between independent variables and soft-drink intake categories were examined using appropriate statistical tests. Logistic regression analysis was performed to estimate the effects of factors associated with high soft-drink intake. Machine-learning techniques were used to classify, evaluate and compare the predictive performance of different models, and determine the key variables that significantly influence consumer classification.

All statistical analyses and machine-learning procedures were conducted using R version 4.2.3, IBM SPSS Statistics version 25, and Microsoft Office 2019.

2.7 Ethics approval and consent to participate

Ethical approval for this study was obtained from the Department of Statistics, Mawlana Bhashani Science and Technology University (MBSTU), Bangladesh. The study was conducted in accordance with the ethical principles and guidelines of the approving institution. Participation was voluntary, and written informed consent was obtained from all participants before data collection. The confidentiality and anonymity of the participants were maintained throughout the study.

3 Results and Discussion

3.1 Univariate and Bivariate Analyses

Among the 450 participants, 51.3% were male and 48.7% were female. Overall, 33.6% were classified as high soft-drink consumers (≥ 1 L/week), while 66.4% were classified as normal consumers.

Table 1 presents the chi-square tests of independence used to examine associations between the study variables and soft-drink consumption level (normal vs. high). Demographic, habitual, thematic, and attitudinal variables showed statistically significant associations with consumption level.

Age (D1) was significantly associated with consumption level ($P = 0.036$); the highest proportion of high consumers was observed among participants aged 27–30 years (52.8%). Academic year (D2) was also significantly associated with consumption ($P = 0.018$), with second-year students showing the highest proportion of high consumption (48.8%). Marital status (D3) was significantly associated with consumption ($P = 0.029$); 46.4% of married participants were high consumers compared with 31.7% of unmarried participants.

Smoking status (H1) and alcohol consumption status (H2) were strongly associated with soft-drink consumption level ($P < 0.001$). Among smokers, 57.5% were high consumers, compared with 24.1% of non-smokers. Among participants reporting alcohol consumption, 69.6% were high consumers, compared with 28.4% among those reporting no alcohol consumption.

All thematic variables (T1–T3) and attitudinal variables (C1–C4) showed statistically significant associations with soft-drink consumption level ($P < 0.05$). For T1 and T2, high consumption was more common among participants who agreed with the statements, whereas for T3 it was more common among those who reported no health problem related to soft-drink consumption. For C1–C3, high consumption was more common among participants who disagreed with the statements, while for C4 it was more common among those who agreed that soft drinks are very tasty.

Table 1: Distribution of respondents' characteristics and perceptions by soft-drink consumption level.

Variables	Category	Consumption Scale		χ^2	P
		Normal Consumer (%)	High Consumer (%)		
D1	19-22	129 (67.2)	63 (32.8)	6.623	0.036
	23-26	153 (68.9)	69 (31.1)		
	27-30	17 (47.2)	19 (52.8)		
D2	1 st Year	56 (72.7)	21 (27.3)	11.924	0.018
	2 nd Year	42 (51.2)	40 (48.8)		
	3 rd Year	78 (72.9)	29 (27.1)		
	4 th Year	66 (67.3)	32 (32.7)		
	5 th Year	57 (66.3)	29 (33.7)		
D3	Married	30 (53.6)	26 (46.4)	4.754	0.029
	Unmarried	269 (68.3)	125 (31.7)		
H1	Smoker	54 (42.5)	73 (57.5)	45.424	<0.001
	Non -Smoker	245 (75.9)	78 (24.1)		
H2	Yes	17 (30.4)	39 (69.6)	37.359	<0.001
	No	282 (71.6)	112 (28.4)		
T1	Agree	60 (56.6)	46 (43.4)	6.023	0.014
	Disagree	239 (69.5)	105 (30.5)		
T2	Agree	44 (50.6)	43 (49.4)	12.183	<0.001
	Disagree	255 (70.2)	108 (29.8)		
T3	Yes	249 (69.9)	107 (30.1)	9.360	0.002
	No	50 (53.2)	44 (46.8)		
C1	Agree	237 (69.5)	104 (30.5)	5.901	0.015
	Disagree	62 (56.9)	47 (43.1)		
C2	Agree	257 (70.0)	110 (30.0)	11.456	.001
	Disagree	42 (50.6)	41 (49.4)		
C3	Agree	226 (69.8)	98 (30.2)	5.682	0.017
	Disagree	73 (57.9)	53 (42.1)		

C4	Agree	211 (63.0)	124 (37.0)	7.036	0.008
	Disagree	88 (76.5)	27 (23.5)		

Table 2 presents simple and multivariable logistic regression analyses of factors associated with high soft-drink consumption. The multivariable model had an AIC of 491.4. The non-significant Hosmer–Lemeshow goodness-of-fit test ($P = 0.828$) indicates no evidence of poor model fit. At the coefficient level, 11 adjusted comparisons across nine variables were statistically significant ($P < 0.05$): D1, D2, D3, C2, C4, H1, H2, T1, and T3. Several simple-regression contrasts were also significant, although not every category contrast was statistically significant.

Table 2: Simple and multivariable logistic regression estimates for factors associated with high soft-drink consumption.

Variables	Categories	Simple Logistic Regression		Multivariable Logistic Regression	
		Sig.	OR	Sig.	OR
D1	19-22 Vs 27-30®	0.02*	0.44	0.02*	0.19
	23-26 Vs 27-30®	0.01*	0.40	0.01*	0.24
D2	1 st year Vs MS®	0.37	0.74	0.06	3.66
	2 nd year Vs MS®	0.05*	1.87	0.00*	6.83
	3 rd year Vs MS®	0.32	0.73	0.07	2.41
	4 th year Vs MS®	0.88	0.95	0.04*	2.44
D3	Married Vs Unmarried®	0.03*	1.87	0.00*	3.18
C1	Agree Vs Disagree®	0.02*	0.58	0.30	0.71
C2	Agree Vs Disagree®	0.00*	0.44	0.02*	0.47
C3	Agree Vs Disagree®	0.02*	0.60	0.91	0.97
C4	Agree Vs Disagree®	0.01*	1.92	0.00*	2.70
H1	Smoker Vs Non-smoke r®	0.00*	4.25	0.00*	3.31
H2	Yes Vs No®	0.00*	5.78	0.01*	2.76
T1	Agree Vs Disagree®	0.02*	1.75	0.04*	1.75
T2	Agree Vs Disagree®	0.00*	2.31	0.13	1.60
T3	Yes Vs No®	0.00*	0.49	0.04*	0.53
AIC		491.4			
Hosmer and Lemeshow Test (Goodness of Fit)		0.828***			

® indicates the reference category; * indicates $P < 0.05$; OR = odds ratio.

3.2 Machine-Learning Classification

Seven machine-learning algorithms were used to classify participants in the test dataset as high or normal soft-drink consumers. The data were split into 70% for model training and 30% for testing. Predictive performance was evaluated using accuracy, 95% confidence intervals, sensitivity, specificity, precision, and AUC. The results are presented in Table 3 and Figure 1.

Table 3: Predictive performance of seven machine-learning algorithms.

Performance	LR	RF	NB	LDA	CART	SVM	KNN
Accuracy	0.7259	0.7704	0.7333	0.7185	0.7185	0.7407	0.7111

95% CI	(0.6425, 0.7991)	(0.6902, 0.8383)	(0.6504, 0.8057)	(0.6347, 0.7925)	(0.6347, 0.7925)	(0.6583, 0.8123)	(0.6269, 0.7858)
Sensitivity	0.8736	0.9540	0.8391	0.8621	0.8161	0.9770	0.9770
Specificity	0.4583	0.4375	0.5417	0.4583	0.5417	0.3125	0.2292
Precision (PPV)	0.7451	0.7545	0.7684	0.7426	0.7634	0.7203	0.6967
AUC	0.7896	0.8460	0.7853	0.7944	0.7750	0.8020	0.7470

Random Forest achieved the highest overall accuracy (0.7704) and AUC (0.8460). Its sensitivity was high (0.9540), although specificity was modest (0.4375). SVM and KNN had the highest sensitivity (0.9770 each) but lower specificity (0.3125 and 0.2292, respectively). Naïve Bayes had the highest precision (0.7684), while Naïve Bayes and CART had the highest specificity (0.5417 each). On balance, Random Forest was selected as the best-performing model because it achieved the highest accuracy and AUC.

The ROC curves in Figure 1 further illustrate discrimination across the seven models. Random Forest had the highest AUC (0.8460), followed by SVM (0.8020), LDA (0.7944), logistic regression (0.7896), Naïve Bayes (0.7853), CART (0.7750), and KNN (0.7470). Thus, Random Forest showed the strongest overall discrimination between high and normal soft-drink consumers in this dataset, while KNN had the lowest AUC.

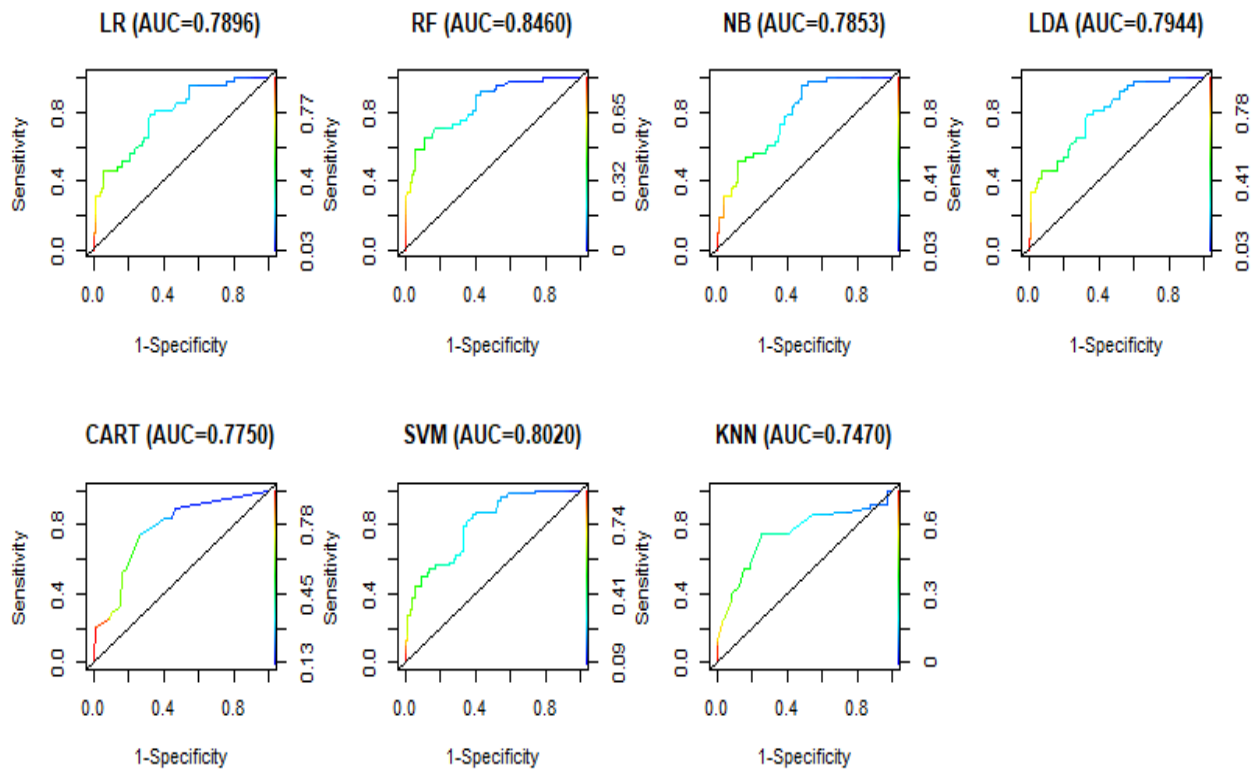


Figure 1: Receiver operating characteristic (ROC) curves and AUC values for the seven machine learning models.

Figure 2 presents the distribution of minimal depth and mean minimal depth for predictors in the Random Forest model. Lower mean minimal depth indicates greater variable importance. Smoking status (H1; mean minimal depth = 2.02) and academic year (D2; 2.07) had the lowest mean minimal depths and were therefore the two most important predictors. In contrast, C1 had the highest mean minimal depth (2.88), indicating comparatively lower importance.

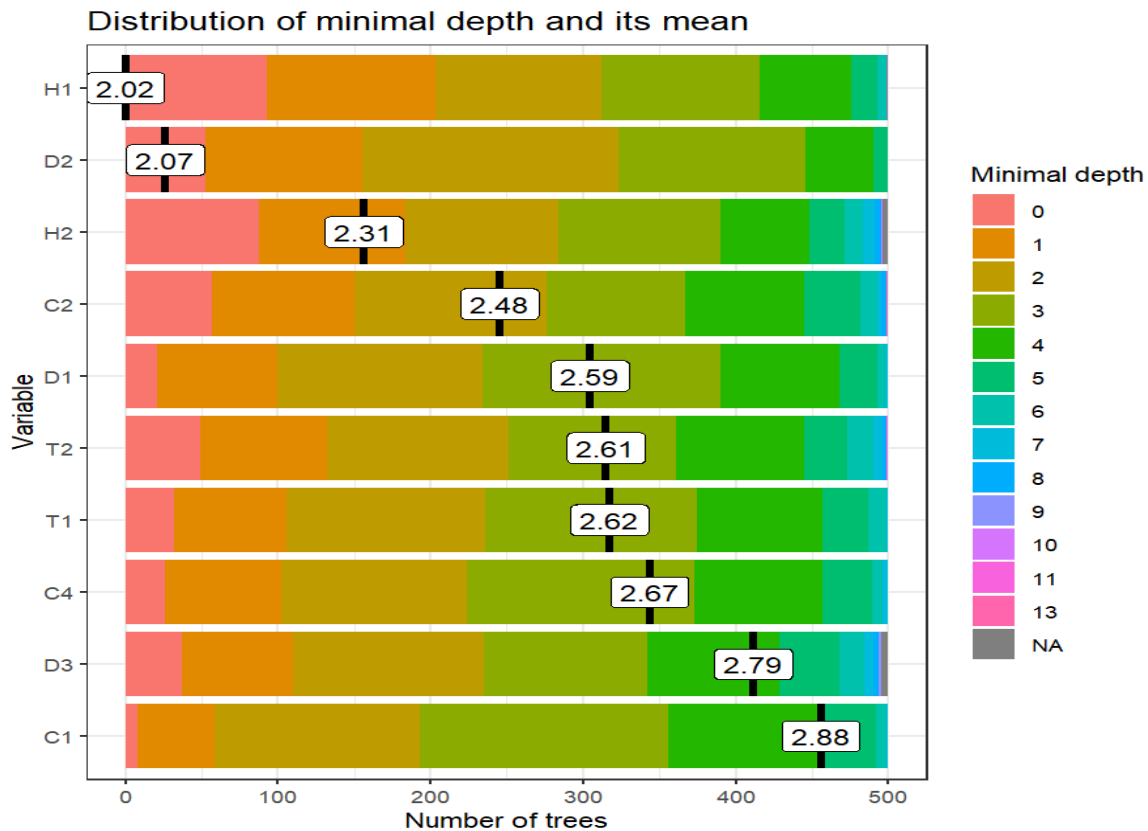


Figure 2: Random Forest variable importance based on minimal depth.

4 Discussion

Our study identified a considerable prevalence of high soft-drink consumption among university students in Bangladesh, with 33.6% classified as high consumers (≥ 1 L/week), which is consistent with findings from a previous study by Rahman et al. (2023). These findings suggest that intake of sugar-sweetened beverages is an important nutritional issue among young adults. High soft-drink consumption was significantly associated with a number of demographic, behavioral and attitudinal factors. Strong associations were observed between high soft-drink intake and smoking and alcohol consumption, suggesting that soft-drink consumption may be part of a broader pattern of unhealthy lifestyle behaviors rather than an isolated dietary choice. Adolescents and young adults have been found to exhibit similar behavioral clustering patterns (Bere et al., 2008; Pengpid & Peltzer, 2020). The academic year was also an important predictor with second and fourth year students having greater odds of high consumption. This may indicate greater independence, academic pressure, peer influence and lifestyle transitions as students' progress through university students (West et al., 2006).

Traditional statistical methods were complemented by machine-learning analysis, which identified complex trends related to consumption behavior. The random forest had the best prediction performance with accuracy of 77.04% and AUC of 0.8460. Smoking status and academic year were the strongest predictors, indicating that they may be useful in identifying students at increased risk for excessive drinking. However, larger and more diverse populations for external validation are needed before implementation in real-world settings.

These findings support the creation of specific university-based interventions for students with multiple risk factors. Strategies that combine health education, the availability of healthier beverages, and supportive campus food environments may help to decrease the consumption of excessive soft drinks. Considering the well-known association between sugar-sweetened beverages and obesity, metabolic disorders, and non-communicable diseases, reducing their consumption among young adults may help prevent long-term disease (Basu et al., 2013; Trumbo & Rivers, 2014; Vartanian et al., 2007).

To summarize, this study highlights the benefits of integrating epidemiological analysis with machine-learning methods to better understand soft-drink consumption patterns. Future longitudinal and multicenter studies are needed to establish causal relationships and improve the generalizability of predictive models.

This study has several limitations that should be considered when interpreting the findings. First, the cross-sectional design prevents establishing temporal relationships or causal associations between soft-drink consumption and the identified demographic, behavioral, and attitudinal factors. Second, participants were recruited using convenience sampling from a single university, which may limit the representativeness of the sample and reduce the generalizability of findings to students from other universities or regions of Bangladesh. Third, soft-drink consumption and related behaviors were self-reported, which may introduce recall bias and social desirability bias. Fourth, the study included a limited number of predictor variables; other potentially important factors, such as dietary patterns, physical activity, socioeconomic status, family influence, availability of beverages, and marketing exposure, were not assessed. Fifth, although machine-learning models showed useful predictive performance, the dataset size was relatively modest and external validation using an independent sample is required before applying these models in broader populations. Future multicenter longitudinal studies with larger samples and additional behavioral and environmental variables are recommended to improve prediction accuracy and provide stronger evidence for intervention planning.

5 Conclusion

Among the 450 university students, 33.6% met the criterion for high soft-drink consumption (≥ 1 L/week). Consumption level was significantly associated with demographic, habitual, thematic, and attitudinal factors. Random Forest identified smoking status and academic year as the two most important predictors and achieved the highest accuracy and AUC among the seven machine-learning models, although its specificity remained modest. These findings support targeted health-promotion interventions to reduce excessive soft-drink intake and encourage healthier beverage choices among university students. Because this study used a cross-sectional design, convenience

sampling, and data from a single university, causal inferences cannot be made and generalizability to other student populations should be interpreted cautiously.

References

- Al Muktadir, M. H., Islam, M. A., Amin, M. N., Ghosh, S., Siddiqui, S. A., Debnath, D., ... & Sultana, F. (2019). Nutrition transition–Pattern IV: Leads Bangladeshi youth to the increasing prevalence of overweight and obesity. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 13(3), 1943-1947. <https://doi.org/10.1016/j.dsx.2019.04.034>
- Basu, S., McKee, M., Galea, G., & Stuckler, D. (2013). Relationship of soft drink consumption to global overweight, obesity and diabetes: A cross-national analysis of 75 countries. *American Journal of Public Health*, 103(11), 2071–2077. <https://doi.org/10.2105/AJPH.2012.300974>
- Bere E. Glomnes E. S. Te Velde S. J. & Klepp K. I. (2008). Determinants of adolescents' soft drink consumption. *Public health nutrition* 11(1) 49-56. <https://doi.org/10.1017/S1368980007000122>
- Bipasha, M., Raisa, T., & Goon, S. (2017). Sugar sweetened beverages consumption among university students of Bangladesh. *Int. J. Public Health Sci*, 6, 157-163. <https://doi.org/10.11591/ijphs.v6i2.6635>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth International Group. ISBN:978-0412048418
- Bruckauf Z. & Walsh S. D. (2018). Adolescents' multiple and individual risk behaviors: Examining the link with excessive sugar consumption across 26 industrialized countries. *Social Science & Medicine* 216 133-141. <https://doi.org/10.1016/j.socscimed.2018.08.029>
- Chen L. Liu R. Zhao Y. & Shi Z. (2020). High consumption of soft drinks is associated with an increased risk of fracture: a 7-year follow-up study. *Nutrients* 12(2) 530. <https://doi.org/10.3390/nu12020530>
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297. <https://doi.org/10.1007/BF00994018>
- Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21-27. <https://doi.org/10.1109/TIT.1967.1053964>
- Ebbeling, C. B., & Ludwig, D. S. (2001). Treating obesity in youth: should dietary glycemic load be a consideration?. *Advances in pediatrics*, 48, 179-212. PMID: 11480757
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 179-188. <https://doi.org/10.1111/j.1469-1809.1936.tb02137.x>

- Forshee R. A. & Storey M. L. (2003). Total beverage consumption and beverage choices among children and adolescents. *International journal of food sciences and nutrition* 54(4) 297-307. <https://doi.org/10.1080/09637480120092143>
- French S. A. Lin B. H. & Guthrie J. F. (2003). National trends in soft drink consumption among children and adolescents age 6 to 17 years: prevalence amounts and sources 1977/1978 to 1994/1998. *Journal of the American Dietetic Association* 103(10) 1326-1331. [https://doi.org/10.1016/S0002-8223\(03\)01076-9](https://doi.org/10.1016/S0002-8223(03)01076-9)
- Grimm G. C. Harnack L. & Story M. (2004). Factors associated with soft drink consumption in school-aged children. *Journal of the American Dietetic Association* 104(8) 1244-1249. <https://doi.org/10.1016/j.jada.2004.05.206>
- Han E. & Powell L. M. (2013). Consumption patterns of sugar-sweetened beverages in the United States. *Journal of the Academy of Nutrition and Dietetics* 113(1) 43-53. <https://doi.org/10.1016/j.jand.2012.09.016>
- Harnack L. Stang J. & Story M. (1999). Soft drink consumption among US children and adolescents: nutritional consequences. *Journal of the American Dietetic Association* 99(4) 436-441. [https://doi.org/10.1016/S0002-8223\(99\)00106-6](https://doi.org/10.1016/S0002-8223(99)00106-6)
- Harrington S. (2008). The role of sugar-sweetened beverage consumption in adolescent obesity: a review of the literature. *The Journal of School Nursing* 24(1) 3-12. <https://doi.org/10.1177/10598405080240010201>
- Heller K. E. Burt B. A. & Eklund S. A. (2001). Sugared soda consumption and dental caries in the United States. *Journal of Dental Research* 80(10) 1949-1953. <https://doi.org/10.1177/00220345010800101701>
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. Wiley.
- Hwang S. B. Park S. Jin G. R. Jung J. H. Park H. J. Lee S. H. ... & Lee B. H. (2020). Trends in beverage consumption and related demographic factors and obesity among Korean children and adolescents. *Nutrients* 12(9) 2651. <https://doi.org/10.3390/nu12092651>
- Johansson A. K. Johansson A. Birkhed D. Omar R. Baghdadi S. & Carlsson G. E. (1996). Dental erosion soft-drink intake and oral health in young Saudi men and the development of a system for assessing erosive anterior tooth wear. *Acta Odontologica Scandinavica* 54(6) 369-378. <https://doi.org/10.3109/00016359609003554>
- Johansson A. K. Sorvari R. Birkhed D. & Meurman J. H. (2001). Dental erosion in deciduous teeth—an in vivo and in vitro study. *Journal of dentistry* 29(5) 333-340. [https://doi.org/10.1016/S0300-5712\(01\)00029-X](https://doi.org/10.1016/S0300-5712(01)00029-X)
- Jones S. G. & Nunn J. H. (1995). The dental health of 3-year-old children in east Cumbria 1993. *Community dental health* 12(3) 161-166. PMID: 7584584

- Khan, A., & Uddin, R. (2020). Is consumption of fast-food and carbonated soft drink associated with anxiety-induced sleep disturbance among adolescents? A population-based study. *Clinical nutrition ESPEN*, 36, 162-165. <https://doi.org/10.1016/j.clnesp.2020.01.011>
- Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. *Proceedings of ECML-98*, 4(1), 4-15. <https://doi.org/10.1007/BFb0026666>
- Li D. Yu D. & Zhao L. (2014). Trend of sugar-sweetened beverage consumption and intake of added sugar in China nine provinces among adults. *Wei Sheng yan jiu= Journal of Hygiene Research* 43(1) 70-72. PMID: 24564114
- Ludwig D. S. (2002). The glycemic index: physiological mechanisms relating to obesity diabetes and cardiovascular disease. *Jama* 287(18) 2414-2423. <https://doi.org/10.1001/jama.287.18.2414>
- Ludwig D. S. Peterson K. E. & Gortmaker S. L. (2001). Relation between consumption of sugar-sweetened drinks and childhood obesity: a prospective observational analysis. *The lancet* 357(9255) 505-508. [https://doi.org/10.1016/S0140-6736\(00\)04041-1](https://doi.org/10.1016/S0140-6736(00)04041-1)
- Mattes R. D. (1996). Dietary compensation by humans for supplemental energy provided as ethanol or carbohydrate in fluids. *Physiology & behavior* 59(1) 179-187. [https://doi.org/10.1016/0031-9384\(95\)02007-1](https://doi.org/10.1016/0031-9384(95)02007-1)
- Maupome-Carvantes G. Sanchez-Reyes V. Laguna-Ortega S. & Andrade-Delgado L. D. C. (1995). Soft drink consumption patterns in a Mexican population. *Salud Publica de Mexico* 37(4) 323-328. PMID: 7502154
- Mistry S. K. & Puthussery S. (2015). Risk factors of overweight and obesity in childhood and adolescence in South Asian countries: a systematic review of the evidence. *Public health* 129(3) 200-209. <https://doi.org/10.1016/j.puhe.2014.12.004>
- Moynihan P. & Petersen P. E. (2004). Diet nutrition and the prevention of dental diseases. *Public health nutrition* 7(1a) 201-226. <https://doi.org/10.1079/PHN2003589>
- Narain A. Kwok C. S. & Mamas M. A. (2017). Soft drink intake and the risk of metabolic syndrome: A systematic review and meta-analysis. *International journal of clinical practice* 71(2) e12927. <https://doi.org/10.1111/ijcp.12927>
- Øverby N. C. Lillegaard I. T. Johansson L. & Andersen L. F. (2004). High intake of added sugar among Norwegian children and adolescents. *Public health nutrition* 7(2) 285-293. <https://doi.org/10.1079/PHN2003515>
- Pengpid, S., & Peltzer, K. (2020). High carbonated soft drink intake is associated with health risk behavior and poor mental health among school-going adolescents in six Southeast Asian countries. *International Journal of Environmental Research and Public Health*, 17(1), 132. <https://doi.org/10.3390/ijerph17010132>
- Popkin B. M. (2001). The nutrition transition and obesity in the developing world. *The Journal of nutrition* 131(3) 871S-873S. <https://doi.org/10.1093/jn/131.3.871S>

- Raben A. Vasilaras T. H. Møller A. C. & Astrup A. (2002). Sucrose compared with artificial sweeteners: different effects on ad libitum food intake and body weight after 10 wk of supplementation in overweight subjects. *The American journal of clinical nutrition* 76(4) 721-729. <https://doi.org/10.1093/ajcn/76.4.721>
- Rahman, M. M., Ahmed, A., Islam, M. R., & Akter, M. (2023). High consumption of soft drinks and perceptions of consumption among university students: A multivariate analysis with a machine learning (ML) algorithm. *Journal of Public Health*, 1-9. <https://doi.org/10.1007/s10389-023-02043-8>
- Rehm C. D. Matte T. D. Van Wye G. Young C. & Frieden T. R. (2008). Demographic and behavioral factors associated with daily sugar-sweetened soda consumption in New York City adults. *Journal of Urban Health* 85 375-385. <https://doi.org/10.1007/s11524-008-9269-8>
- Rodríguez-Artalejo, F., García, E. L., Gorgojo, L., Garcés, C., Royo, M. A., Moreno, J. M. M., ... & De Oya, M. (2003). Consumption of bakery products, sweetened soft drinks and yogurt among children aged 6–7 years: association with nutrient intake and overall diet quality. *British Journal of Nutrition*, 89(3), 419-428. <https://doi.org/10.1079/BJN2002787>
- Rolls B. J. (1997). Fat and Sugar Substitutes and the Control of Food Intake a. *Annals of the New York Academy of Sciences* 819(1) 180-193. <https://doi.org/10.1111/j.1749-6632.1997.tb51808.x>
- Scully, M., Morley, B., Niven, P., Crawford, D., Pratt, I. S., & Wakefield, M. (2017). Factors associated with high consumption of soft drinks among Australian secondary-school students. *Public Health Nutrition*, 20(13), 2340-2348. <https://doi.org/10.1017/S1368980017000118>
- Sinha, S., Sinha, D., & Mittal, N. (2021). Consumer characteristics and consumption patterns of soft drinks. 2021 9th International Conference on Reliability, Infocom Technologies and Optimization, 1–6. <https://doi.org/10.1109/ICRITO51393.2021.9596207>
- Srinivasan, N., Elangovan, S., & Chinnaiyan, P. (2000). Consumer perception towards processed fruit and vegetables products. *Indian Economic Panorama*, 10(3), 11-12.
- Storey M. L. Forshee R. A. & Anderson P. A. (2006). Beverage consumption in the US population. *Journal of the American Dietetic Association* 106(12) 1992-2000. <https://doi.org/10.1016/j.jada.2006.09.009>
- Tordoff M. G. & Alleva A. M. (1990). Effect of drinking soda sweetened with aspartame or high-fructose corn syrup on food intake and body weight. *The American journal of clinical nutrition* 51(6) 963-969. <https://doi.org/10.1093/ajcn/51.6.963>
- Trumbo P. R. & Rivers C. R. (2014). Systematic review of the evidence for an association between sugar-sweetened beverage consumption and risk of obesity. *Nutrition reviews* 72(9) 566-574. <https://doi.org/10.1111/nure.12128>

- Vartanian L. R. Schwartz M. B. & Brownell K. D. (2007). Effects of soft drink consumption on nutrition and health: a systematic review and meta-analysis. *American journal of public health* 97(4) 667-675. <https://doi.org/10.2105%2FAJPH.2005.083782>
- Vereecken C. A. Inchley J. Subramanian S. V. Hublet A. & Maes L. (2005). The relative influence of individual and contextual socio-economic status on consumption of fruit and soft drinks among adolescents in Europe. *The European Journal of Public Health* 15(3) 224-232. <https://doi.org/10.1093/eurpub/cki005>
- West D. S. Bursac Z. Quimby D. Prewitt T. E. Spatz T. Nash C. ... & Eddings K. (2006). Self-reported sugar-sweetened beverage intake among college students. *Obesity* 14(10) 1825-1831. <https://doi.org/10.1038/oby.2006.210>
- World Health Organization. (2003). Diet nutrition and the prevention of chronic diseases: report of a joint WHO/FAO expert consultation (Vol. 916). World Health Organization.