Original Paper

Machine Learning–Based Prediction of Substance Use in Adolescents in Three Independent Worldwide Cohorts: Algorithm Development and Validation Study

Soeun Kim[1,2*], MS; Hyejun Kim[1,3*], BS; Seokjun Kim[1,4*], MD; Hojae Lee[1], MS; Ahmed Hammoodi[5], BBA; Yujin Choi[1,6], BS; Hyeon Jin Kim[1,2], PhD; Lee Smith[7], PhD; Min Seo Kim[8], MD; Guillaume Fond[9], MD, PhD; Laurent Boyer[9], MD, PhD; Sung Wook Baik[10], PhD; Hayeon Lee[1,11], PhD; Jaeyu Park[1,2], PhD; Rosie Kwon[1], MS; Selin Woo[1,4], PhD; Dong Keon Yon[1,2,4,12], MD, PhD

[1]Center for Digital Health, Medical Science Research Institute, Kyung Hee University Medical Center, Seoul, Republic of Korea
[2]Department of Precision Medicine, Kyung Hee University College of Medicine, Seoul, Republic of Korea
[3]Department of Applied Information Engineering, Yonsei University, Seoul, Republic of Korea
[4]Department of Medicine, Kyung Hee University College of Medicine, Seoul, Republic of Korea
[5]Department of Business Administration, Kyung Hee University School of Management, Seoul, Republic of Korea
[6]Department of Korean Medicine, Kyung Hee University College of Korean Medicine, Seoul, Republic of Korea
[7]Centre for Health, Performance and Wellbeing, Anglia Ruskin University, Cambridge, United Kingdom
[8]Cardiovascular Disease Initiative, Broad Institute of MIT and Harvard, Cambridge, MA, United States
[9]Research Centre on Health Services and Quality of Life, Assistance Publique-Hopitaux de Marseille, Aix Marseille University, Marseille, France
[10]Department of Software, Sejong University College of Electronics and Information Engineering, Seoul, Republic of Korea
[11]Department of Electronics and Information Convergence Engineering, Kyung Hee University, Seoul, Republic of Korea
[12]Department of Pediatrics, Kyung Hee University Medical Center, Kyung Hee University College of Medicine, Seoul, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Dong Keon Yon, MD, PhD
Department of Pediatrics
Kyung Hee University Medical Center
Kyung Hee University College of Medicine
23 Kyungheedae-ro, Dongdaemun-gu
Seoul, 02447
Republic of Korea
Phone: 82 269352476
Fax: 82 5044780201
Email: yonkkang@gmail.com

Abstract

Background: To address gaps in global understanding of cultural and social variations, this study used a high-performance machine learning (ML) model to predict adolescent substance use across three national datasets.

Objective: This study aims to develop a generalizable predictive model for adolescent substance use using multinational datasets and ML.

Methods: The study used the Korea Youth Risk Behavior Web-Based Survey (KYRBS) from South Korea (n=1,098,641) to train ML models. For external validation, we used the Youth Risk Behavior Survey (YRBS) from the United States (n=2,511,916) and Norwegian nationwide Ungdata surveys (Ungdata) from Norway (n=700,660). After developing various ML models, we evaluated the final model's performance using multiple metrics. We also assessed feature importance using traditional methods and further analyzed variable contributions through SHapley Additive exPlanation values.

Results: The study used nationwide adolescent datasets for ML model development and validation, analyzing data from 1,098,641 KYRBS adolescents, 2,511,916 YRBS participants, and 700,660 from Ungdata. The XGBoost model was the top performer on the KYRBS, achieving an area under receiver operating characteristic curve (AUROC) score of 80.61% (95% CI 79.63-81.59) and precision of 30.42 (95% CI 28.65-32.16) with detailed analysis on sensitivity of 31.30 (95% CI 29.47-33.20), specificity of 99.16 (95% CI 99.12-99.20), accuracy of 98.36 (95% CI 98.31-98.42), balanced accuracy of 65.23 (95% CI 64.31-66.17), F_1-score of 30.85 (95% CI 29.25-32.51), and area under precision-recall curve of 32.14 (95% CI 30.34-33.95). The model achieved an AUROC score of 79.30% and a precision of 68.37% on the YRBS dataset, while in external validation using the Ungdata dataset, it recorded an AUROC score of 76.39% and a precision of 12.74%. Feature importance and SHapley Additive exPlanation value analyses identified smoking status, BMI, suicidal ideation, alcohol consumption, and feelings of sadness and despair as key contributors to the risk of substance use, with smoking status emerging as the most influential factor.

Conclusions: Based on multinational datasets from South Korea, the United States, and Norway, this study shows the potential of ML models, particularly the XGBoost model, in predicting adolescent substance use. These findings provide a solid basis for future research exploring additional influencing factors or developing targeted intervention strategies.

(J Med Internet Res 2025;27:e62805) doi: 10.2196/62805

KEYWORDS

adolescents; machine learning; substance; prediction; XGBoost; random forest; ML; substance use; adolescents; adolescent; South Korea; United States; Norway; web-based survey; survey; risk behavior; smoking; alcohol; intervention; interventions

Introduction

Substance use among adolescents remains a global concern, often leading to both immediate and long-term health challenges, such as mental health disorders and addiction [1]. When initiated at an early age, these behaviors can escalate to more serious health conditions, including chronic substance dependence and comorbid mental health issues [2]. As globalization and cultural integration continue to expand, substance use patterns vary significantly across regions, making it paramount to understand these patterns across a diverse cultural landscape [3]. Conventional statistical methods have long been used to examine the predictors and outcomes of adolescent substance use [4]. However, these approaches often fall short in capturing complex, nonlinear relationships between variables. Therefore, with recent advancements, machine learning (ML) has introduced powerful tools capable of identifying complex patterns and relationships, offering a deeper understanding of adolescent substance use [5].

Existing studies provide insights into the epidemiology and sociocultural factors associated with adolescent substance use in various contexts, but few have used ML techniques with multinational datasets [4,6]. Such approaches have the potential to yield more precise and globally relevant insights [5,7]. By identifying key predictors of adolescent substance use that remain consistent across diverse cultural contexts, this study aims to develop a prediction model adaptable to global public health initiatives. To validate its generalizability, the model was tested using datasets from two additional countries, highlighting its adaptability to diverse sociocultural environments [8].

This study developed an ML-based prediction model for adolescent substance use, using comprehensive datasets from South Korea, and extended our validation process by using datasets from the United States and Norway [9]. This validation process and refinement process ensured the model's accuracy and applicability across diverse cultural and national contexts. By integrating these global datasets, we developed a predictive model that reflects our collaborative international research. Our novel approach equips stakeholders with a sophisticated tool informed by global data. This helps address and preempt adolescent substance use effectively across different national contexts.

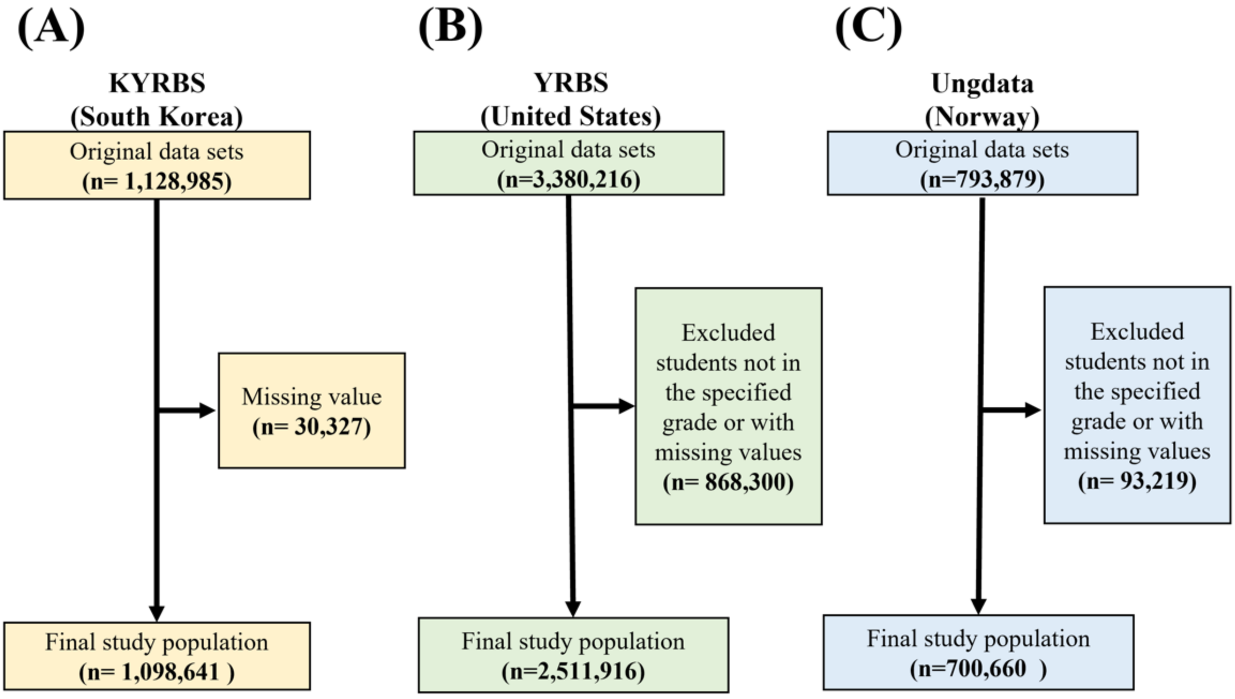
Methods

Study Design and Participants

Adolescents enrolled in middle and high school who completed their respective surveys were included. In the context of educational systems, adolescents are more appropriately categorized by grade level rather than age. To ensure consistency across the three countries, participants were limited to students from middle school (7th grade) to high school (12th grade). This study was primarily designed to develop an ML model for substance use prediction among adolescents using three distinct nationwide datasets: Korea Youth Risk Behavior Web-Based Survey (KYRBS) from South Korea [10], Youth Risk Behavior Survey (YRBS) from the United States [6,11], and Norwegian nationwide Ungdata surveys (Ungdata) from Norway [12]. KYRBS was initially used to train the ML model, followed by the external validation process using the YRBS and Ungdata.

The discovery dataset, KYRBS, was conducted annually by the Korean Disease Control and Prevention Agency (KDCA) from 2008 to 2022, to assess health behaviors among Korean middle and high school students. It began with 1,128,985 participants and ended up with 1,098,641 after exclusions from the missing values in BMI [13], primarily representing a demographic largely comprised of East Asians (Figure 1A). The KYRBS dataset can be accessed through the official website of the KDCA.

Figure 1. Study population. KYRBS: Korea Youth Risk Behavior Web-Based Survey; Ungdata: Norwegian nationwide Ungdata surveys; YRBS: Youth Risk Behavior Survey.



The YRBS, which was conducted by the US Centers for Disease Control and Prevention (CDC), along with state and local education and health agencies, began with 3,380,216 participants, was narrowed down to 2,511,916 after excluding students not enrolled in middle or high school and those with missing data (Figure 1B**)**. YRBS encompasses a diverse racial spectrum of American adolescents. YRBS data are downloaded and available through the US CDC's official website.

Similarly, from the initial count of 793,879 in the Ungdata, a cross-sectional survey conducted by the social research institute, Norwegian Social Research Institute (NOVA), at Oslo Metropolitan University consists of a questionnaire for school pupils throughout, only 700,660 participants were selected after data processing (Figure 1C). Ungdata is offered to all local and county councils in Norway, who administer the questionnaire in collaboration with NOVA and regional centers for substance use rehabilitation. The dataset can be accessed via the official Ungdata website. During the data processing phase, individual missing values in the KYRBS and YRBS datasets were imputed using a random forest regression–based imputation method [14]. For the extravalidation cohorts, YRBS and Ungdata, variables that were entirely absent were imputed using the median values derived from the discovery dataset, KYRBS.

In order to evaluate the generalizability of our model across diverse cultural groups, we used a phased validation approach. The validation process used the YRBS dataset, which includes participants from diverse ethnic backgrounds, as well as the Ung dataset from Norway, representing a markedly different cultural context. By using validation datasets encompassing a wide range of cultural groups, we conducted a rigorous assessment of the model's robustness and versatility across diverse populations [15].

Our primary outcome, substance use, was derived from the question "Have you ever consumed illicit substances at least once in your lifetime." We distinguished smoking and alcohol from other substances in our analysis, recognizing their unique consumption patterns, sociocultural implications, and health effects [16]. Substances other than smoking and alcohol are distinguished primarily due to concerns regarding their potential for misuse and health risks [17]. This decision was made to ensure that our model captures nuances specific to each substance, thereby enhancing the specificity and relevance of our predictions. Integral covariates under consideration spanned across factors including grade, sex, region, BMI, academic achievement, household income, smoking status, alcoholic consumption, stress status, sadness and despair, suicidal thinking, and suicide attempts [13].

Model Development and Validation

Using the KYRBS, we developed a predictive model to extrapolate the behavioral patterns of Korean adolescents regarding substance use. Due to rigorous substance regulations and law enforcement measures in South Korea, accessibility and consumption of substances are notably limited [18]. Consequently, the number of instances representing substance use within the KYRBS was sparse (n=12,803, 1.17%).

Given the intricate nature of the data, we used a comprehensive approach to model development, leveraging 10-fold cross-validation to divide the KYRBS dataset into training and testing subsets. Various tree-based and statistical models, including LightGBM, CatBoost, AdaBoost, random forest, and XGBoost, were evaluated in comparison with logistic regression, a widely used baseline model, to determine the most effective algorithm for predicting substance use [19,20].

After model development using the KYRBS, external validation was conducted with datasets from diverse cultural contexts (YRBS and Ungdata) to evaluate the generalizability of the predictive model. This step ensured the applicability of the modeling approach across heterogeneous adolescent populations [21].

To further strengthen the validity of our results, hyperparameter tuning was performed for each algorithm using GridSearchCV, focusing on maximizing performance metrics such as the area under the receiver operating characteristic curve (AUROC) and precision. The hyperparameter values for the selected model are detailed in Table S1 in Multimedia Appendix 1. The sensitivity and specificity of the model were determined based on a classification threshold of 0.5. Various metrics, including AUROC score, accuracy, sensitivity, specificity, balanced accuracy, precision, F_1-score (ie, the harmonic mean of the precision and recall), and the area under precision-recall curve (AUPRC) were used to evaluate the model's performance across datasets [13].

Performance Assessment

The tools and techniques we used for assessment were consistent. Our evaluation metrics comprised AUROC, accuracy, sensitivity, specificity, balanced accuracy, precision, F_1-score, and AUPRC. These metrics were collectively considered to comprehensively evaluate and compare model performance. To provide a visual representation of the model efficacy, we used visualization techniques, notably the ROC curve [19,20,22-24].

SHapley Additive ExPlanation Value Analysis

To interpret and better understand the model's predictions, we calculated the SHapley Additive explanation (SHAP) values based on the summary model [25]. SHAP is a widely used method for providing local explanations of ML model predictions. Proposed by Lundberg and Lee, SHAP offers a unified framework for explaining the output of any ML model [25]. This technique visualizes the contribution of each feature to the model's prediction for a specific instance, illustrating how each feature shifts the model's output from the base value.

Software and Libraries

All computations, model training, validation, and evaluation processes were executed using Python (version 3.12.4; Python Software Foundation). Key libraries from our toolbox included Scikit-learn (version 1.5.2; Scikit-learn development team), NumPy (version 1.26.4; Python Software Foundation), and Pandas (version 2.2.0; Python Software Foundation) for ML tasks and data wrangling. Visualization was facilitated using Matplotlib (version 3.8.4; Python Software Foundation) and Seaborn (version 0.13.2; Python Software Foundation).

Ethical Considerations

The study protocol was approved by the institutional review board of the KDCA (2014-06EXP-02-P-A), US CDC (#1969.0), and NOVA (18778329) [12], and all participants provided written informed consent. This research followed the guidelines outlined in the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) statement [26].

Results

Demographic Characteristics

This research used a detailed exploration using nationwide adolescent datasets from South Korea, aiming to design and validate an ML model to predict substance use tendencies among adolescents. The primary demographic consisted of middle (7th grade) to high school (12th grade) students (Figure 1).

We collected data from the KYBS, the YRBS, and the Ungdata, and subsequently standardized the covariates for the ML predictive modeling process. Within the primary cohort from the KYRBS to develop the prediction model, the sex distribution was as follows: male (n=566,437, 51.56%) and female (n=532,204, 48.44%). For the initial external validation cohorts of an external validation process, the YRBS features a sex distribution of: male (n=1,233,846, 49.12%) and female (n=1,278,070, 50.88%). In the second validation step, the Ungdata has the following sex distribution: male (n=345,428, 49.30%) and female (n=355,232, 50.70%; Table 1).

Table 1. Demographic characteristics of KYRBS[a] from South Korea (2005-2022), YRBS[b] from the United States (1998-2022), and Ungdata from Norway (2014-2021).

	South Korea (KYRBS; n=1,098,641), n (%)	United States (YRBS; n=2,511,916), n (%)	Norway (Ungdata; n=700,660), n (%)
Region			
Urban	509,058 (46.34)	N/A[c]	N/A
Rural	589,583 (53.66)	N/A	N/A
Grade			
7th grade	190,112 (17.30)	448,865 (17.87)	144,229 (20.58)
8th grade	190,166 (17.31)	402,753 (16.03)	141,806 (20.24)
9th grade	189,842 (17.28)	476,360 (18.96)	143,938 (20.54)
10th grade	182,908 (16.65)	448,865 (17.87)	127,053 (18.13)
11th grade	181,428 (16.51)	402,753 (16.03)	89,637 (12.79)
12th grade	164,185 (14.94)	332,320 (13.23)	53,997 (7.71)
Sex			
Male	566,437 (51.56)	1,233,846 (49.12)	345,428 (49.30)
Female	532,204 (48.44)	1,278,070 (50.88)	355,232 (50.70)
BMI[d]			
Unknown	2940 (0.27)	448,630 (17.86)	N/A
Underweight	88,787 (8.08)	62,540 (2.49)	N/A
Normal	830,683 (75.61)	1,428,266 (56.86)	N/A
Overweight	89,971 (8.19)	308,141 (12.27)	N/A
Obese	86,260 (7.85)	264,339 (10.52)	N/A
Academic achievement			
Low (0-19 percentile)	114,832 (10.45)	N/A	N/A
Lower-middle (20-39 percentile)	255,845 (23.29)	N/A	N/A
Middle (40-59 percentile)	314,304 (28.61)	N/A	N/A
Upper-middle (60-79 percentile)	278,359 (25.34)	N/A	N/A
High (80-100 percentile)	135,301 (12.32)	N/A	N/A
Household income			
Low (0-19 percentile)	43,920 (4.00)	N/A	7968 (1.14)
Lower-middle (20-39 percentile)	159,283 (14.50)	N/A	27,885 (3.98)
Middle (40-59 percentile)	516,595 (47.02)	N/A	121,858 (17.39)
Upper-middle (60-79 percentile)	290,304 (26.42)	N/A	236,136 (33.70)
High (80-100 percentile)	88,539 (8.06)	N/A	306,813 (43.79)
Smoking status			
Nonsmoker	869,907 (79.18)	1,179,077 (46.94)	562,118 (80.23)
Smoker	228,734 (20.82)	1,332,839 (53.06)	138,542 (19.77)
Alcoholic consumption			
Nondrinker	581,345 (52.91)	1,335,962 (53.18)	306,210 (43.70)
More than one time	517,296 (47.09)	1,175,954 (46.82)	394,450 (56.30)
Stress status[e]			
Low	30,456 (2.77)	N/A	N/A
Mild	159,720 (14.54)	N/A	N/A

	South Korea (KYRBS; n=1,098,641), n (%)	United States (YRBS; n=2,511,916), n (%)	Norway (Ungdata; n=700,660), n (%)
Moderate	455,026 (41.42)	N/A	N/A
High	324,847 (29.57)	N/A	N/A
Severe	128,592 (11.70)	N/A	N/A
Sadness and despair in the past year			
Unknown	N/A	1,079,817 (42.99)	N/A
No	749,479 (68.22)	1,005,251 (40.02)	522,702 (74.60)
Yes	349,162 (31.78)	426,848 (16.99)	177,958 (25.40)
Suicidal thinking in the past year			
No	914,543 (83.24)	1,506,410 (59.97)	N/A
Yes	184,098 (16.76)	1,005,506 (40.03)	N/A
Suicide attempts in the past year			
No	1,057,142 (96.22)	1,899,009 (75.60)	N/A
Yes	41,499 (3.78)	612,907 (24.40)	N/A
Substance use			
No	1,085,838 (98.83)	1,310,191 (52.16)	685,866 (97.89)
Yes	12,803 (1.17)	1,201,725 (47.84)	14,794 (2.11)

[a] KYRBS: Korea Youth Risk Behavior Web-Based Survey.

[b] YRBS: Youth Risk Behavior Survey.

[c] N/A: not applicable.

[d] BMI was divided into four groups according to the 2017 Korean National Growth Charts: underweight (0-4 percentile), normal (5-84 percentile), overweight (85-94 percentile), and obese (95-100 percentile).

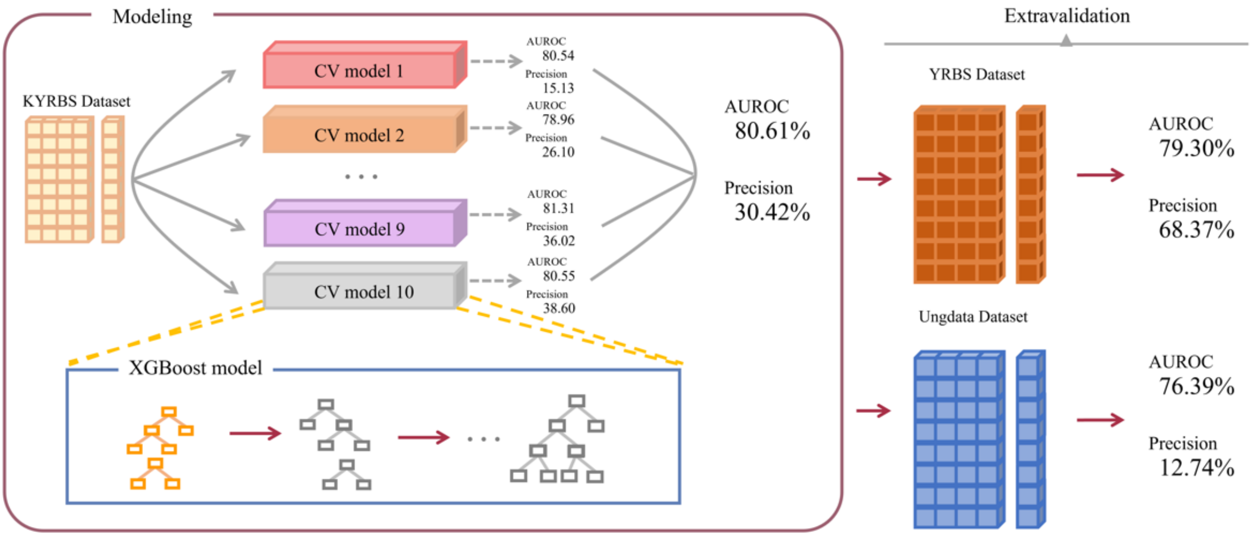
[e] Stress was defined by the receipt of mental health counseling owing to stress.

We compared the distributions of key variables across the three cohorts (KYRBS, YRBS, and Ungdata), as presented in Table 1. Notable similarities and differences were identified. For instance, the proportion of smokers differed significantly between KYRBS (228,734, 20.82%) and YRBS (n=1,332,839, 53.06%), while the Ungdata cohort exhibited a much lower smoking prevalence (n=138,542, 19.77%). Similarly, alcohol consumption rates were higher in the Ungdata cohort (n=394,450, 56.30%) compared to KYRBS (n=517,296, 47.09%) and YRBS (n=1,175,954, 46.82%). These differences likely reflect cultural and policy variations influencing substance accessibility and behavioral norms in each country. Furthermore, visual comparisons of key baseline characteristics, such as smoking and alcohol consumption, are provided in Figure S1 in Multimedia Appendix 1 to offer additional insights.

ML Model Results

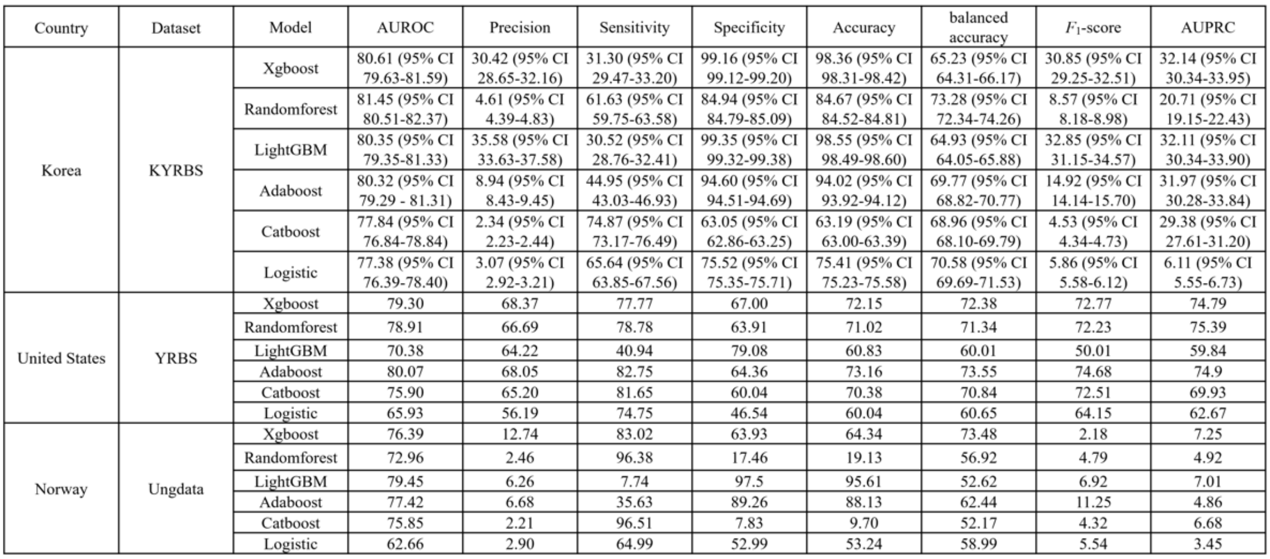
Extensive model evaluations, considering both AUROC and precision, revealed that the XGBoost model was the optimal model for predicting substance use among adolescents (Figures 2 and 3). The primary model, sourced from the KYRBS and assessed disclosed that the XGBoost model notched an AUROC score of 80.61% (95% CI 79.63-81.59) and precision of 30.42 (95% CI 28.65-32.16) with detailed analysis on a sensitivity of 31.30 (95% CI 29.47-33.20), specificity of 99.16 (95% CI 99.12-99.20), accuracy of 98.36 (95% CI 98.31-98.42), balanced accuracy of 65.23 (95% CI 64.31-66.17), F_1-score of 30.85 (95% CI 29.25-32.51), and AUPRC of 32.14 (95% CI 30.34-33.95). Other models exhibited the following AUROC scores: random forest at 81.45 (95% CI 80.51-82.37), LightGBM at 80.35 (95% CI 79.35-81.33), AdaBoost at 80.32 (95% CI 79.29-81.31), CatBoost at 77.84 (95% CI 76.84-78.84), and Logistic 77.38 (95% CI 76.39-78.40). Precision scores were as follows: LightGBM at 35.58 (95% CI 33.63-37.58), AdaBoost at 8.94 (95% CI 8.43-9.45), random forest at 4.61 (95% CI 4.39-4.83), Logistic at 3.07 (95% CI 2.92-3.21), and CatBoost at 2.34 (95% CI 2.23-2.44).

Figure 2. Model architecture. The original KYRBS was partitioned into the original data set for model development, with performance assessed using the AUROC score. Selected high-performing models were further validated. The external validations were generated using YRBS and Ungdata. AUROC: area under the receiver operating characteristic curve; KYRBS: Korea Youth Risk Behavior Web-Based Survey; YRBS: Youth Risk Behavior Survey.

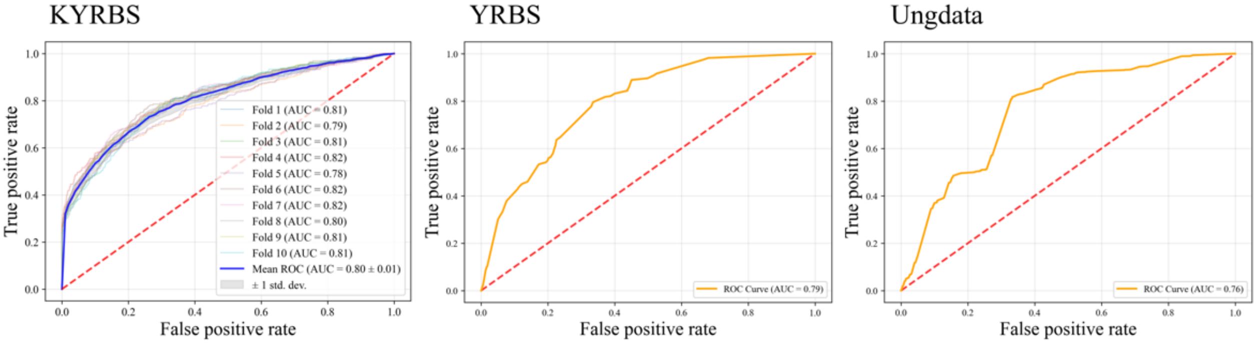


For the initial external validation, the independent YRBS dataset was used. The XGBoost displayed an AUROC score of 79.30%, followed by a precision of 68.37%, sensitivity of 77.77%, specificity of 67%, accuracy of 72.15%, and balanced accuracy of 72.38%, F_1-score of 72.77%, and AUPRC of 74.79%. In the subsequent external validation using the Ung dataset, the XGBoost model yielded an AUROC score of 76.39%, coupled with a precision of 12.74%, sensitivity of 83.02%, specificity of 63.93%, accuracy of 64.34%, balanced accuracy of 73.48%, F_1-score of 2.18, and AUPRC of 7.25 (Figure 3).

Figure 3. The assessment of five different machine learning algorithms using AUROC score and ROC curve for initial model construction with the KYRBS, and external validation with the YRBS and Ungdata. AUROC: area under the receiver operating characteristic curve; KYRBS: Korea Youth Risk Behavior Web-Based Survey; LightGBM: light gradient boosting model; ROC: receiver operating characteristic; XGBoost: extreme Gradient Boosting model; YRBS: Youth Risk Behavior Survey.



Country	Dataset	Model	AUROC	Precision	Sensitivity	Specificity	Accuracy	balanced accuracy	F_1-score	AUPRC
Korea	KYRBS	Xgboost	80.61 (95% CI 79.63-81.59)	30.42 (95% CI 28.65-32.16)	31.30 (95% CI 29.47-33.20)	99.16 (95% CI 99.12-99.20)	98.36 (95% CI 98.31-98.42)	65.23 (95% CI 64.31-66.17)	30.85 (95% CI 29.25-32.51)	32.14 (95% CI 30.34-33.95)
		Randomforest	81.45 (95% CI 80.51-82.37)	4.61 (95% CI 4.39-4.83)	61.63 (95% CI 59.75-63.58)	84.94 (95% CI 84.79-85.09)	84.67 (95% CI 84.52-84.81)	73.28 (95% CI 72.34-74.26)	8.57 (95% CI 8.18-8.98)	20.71 (95% CI 19.15-22.43)
		LightGBM	80.35 (95% CI 79.35-81.33)	35.58 (95% CI 33.63-37.58)	30.52 (95% CI 28.76-32.41)	99.35 (95% CI 99.32-99.38)	98.55 (95% CI 98.49-98.60)	64.93 (95% CI 64.05-65.88)	32.85 (95% CI 31.15-34.57)	32.11 (95% CI 30.34-33.90)
		Adaboost	80.32 (95% CI 79.29 - 81.31)	8.94 (95% CI 8.43-9.45)	44.95 (95% CI 43.03-46.93)	94.60 (95% CI 94.51-94.69)	94.02 (95% CI 93.92-94.12)	69.77 (95% CI 68.82-70.77)	14.92 (95% CI 14.14-15.70)	31.97 (95% CI 30.28-33.84)
		Catboost	77.84 (95% CI 76.84-78.84)	2.34 (95% CI 2.23-2.44)	74.87 (95% CI 73.17-76.49)	63.05 (95% CI 62.86-63.25)	63.19 (95% CI 63.00-63.39)	68.96 (95% CI 68.10-69.79)	4.53 (95% CI 4.34-4.73)	29.38 (95% CI 27.61-31.20)
		Logistic	77.38 (95% CI 76.39-78.40)	3.07 (95% CI 2.92-3.21)	65.64 (95% CI 63.85-67.56)	75.52 (95% CI 75.35-75.71)	75.41 (95% CI 75.23-75.58)	70.58 (95% CI 69.69-71.53)	5.86 (95% CI 5.58-6.12)	6.11 (95% CI 5.55-6.73)
United States	YRBS	Xgboost	79.30	68.37	77.77	67.00	72.15	72.38	72.77	74.79
		Randomforest	78.91	66.69	78.78	63.91	71.02	71.34	72.23	75.39
		LightGBM	70.38	64.22	40.94	79.08	60.83	60.01	50.01	59.84
		Adaboost	80.07	68.05	82.75	64.36	73.16	73.55	74.68	74.9
		Catboost	75.90	65.20	81.65	60.04	70.38	70.84	72.51	69.93
		Logistic	65.93	56.19	74.75	46.54	60.04	60.65	64.15	62.67
Norway	Ungdata	Xgboost	76.39	12.74	83.02	63.93	64.34	73.48	2.18	7.25
		Randomforest	72.96	2.46	96.38	17.46	19.13	56.92	4.79	4.92
		LightGBM	79.45	6.26	7.74	97.5	95.61	52.62	6.92	7.01
		Adaboost	77.42	6.68	35.63	89.26	88.13	62.44	11.25	4.86
		Catboost	75.85	2.21	96.51	7.83	9.70	52.17	4.32	6.68
		Logistic	62.66	2.90	64.99	52.99	53.24	58.99	5.54	3.45



Across all evaluations, both internal and external, the XGBoost model consistently exhibited a predominant performance, particularly in terms of the AUROC score and precision, cementing its adoption for the study objective.

Feature Importance

Table 2 illustrates the importance of various features as determined by the XGBoost model in predicting substance use among adolescents. Specifically, smoking status was identified as the most significant predictor, accounting for 16.62% importance. This was closely followed by BMI with 13.45%, and suicidal thinking at 12.58%. Other notable features include suicide attempts (9.88%), grade (9.83%), stress status (8.99%), academic achievement (7.51%), household income (7.45%), sadness and despair (6.84%), alcoholic consumption (3.78%), sex (2.76%), and region (0.32%).

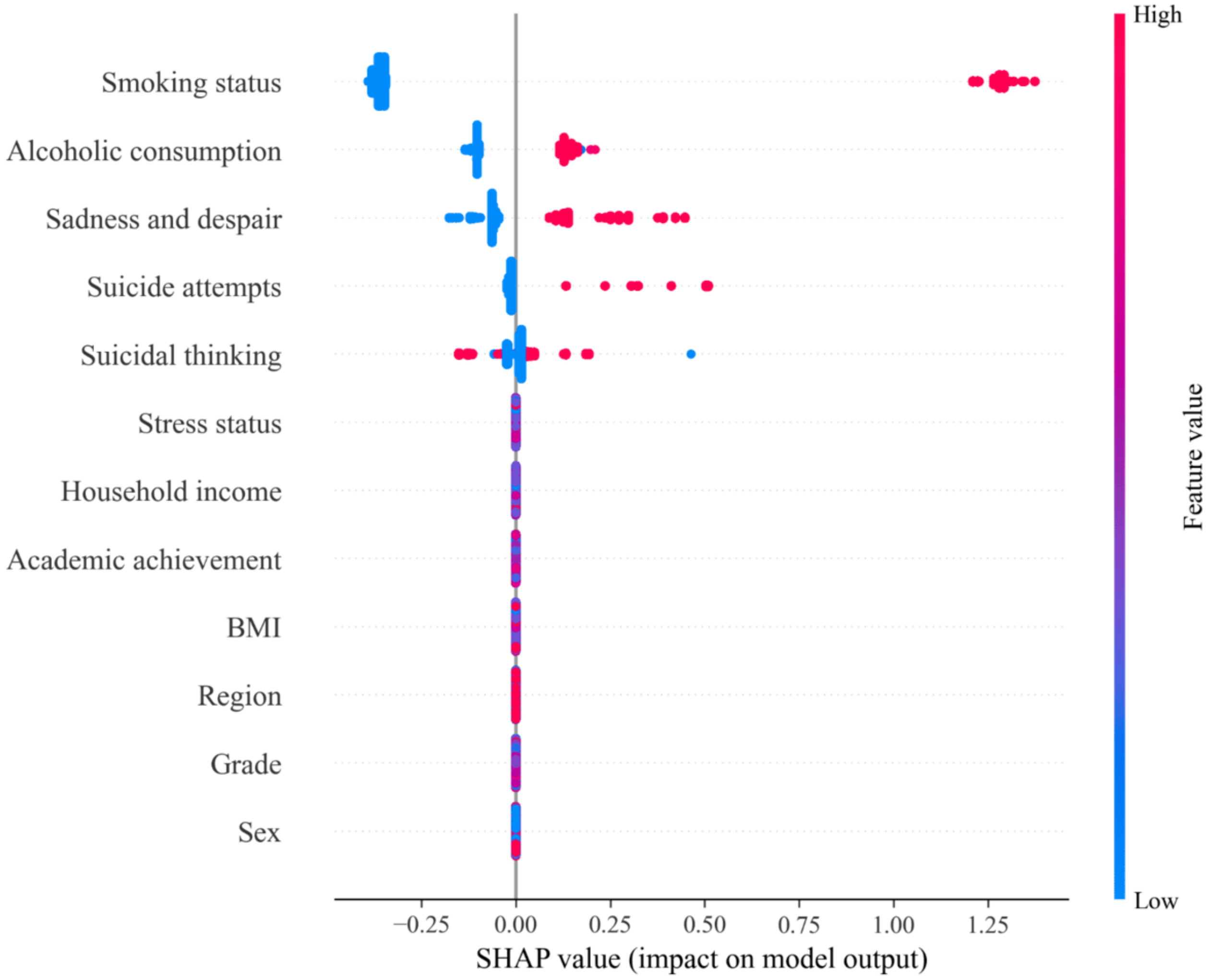
Table 2. Feature importance of the XGBoost model.

Feature	Importance (%)
Smoking status	16.62
BMI	13.45
Suicidal thinking	12.58
Suicide attempts	9.88
Grade	9.83
Stress status	8.99
Academic achievement	7.51
Household income	7.45
Sadness and despair	6.84
Alcoholic consumption	3.78
Sex	2.76
Region	0.32

SHAP Values

Figure 4 presents the SHAP analysis results for the substance use prediction model, illustrating the contribution of each variable to predicting the likelihood of substance use [25]. The analysis identified smoking status as the most influential variable, with higher smoking levels strongly associated with an increased likelihood of substance use. Similarly, alcoholic consumption and sadness and despair were identified as key factors; alcohol consumption consistently increased the likelihood of substance use, while emotional states such as sadness exhibited both positive and negative effects depending on their levels.

Figure 4. SHAP value of the XGBoost model. SHAP: SHapley Additive exPlanation.



Suicidal thinking and suicide attempts also showed distinct and significant impacts, with higher values substantially increasing the likelihood of substance use. In contrast, variables such as BMI, grade, and sex had relatively minimal contributions, indicating that the model is more sensitive to psychological and behavioral factors than to demographic characteristics. These findings offer critical insights into the primary predictors of substance use and their complex interactions.

Code Availability

Based on the results of the ML model, we established a web-based application for policy implementation or health system management to support their decision-making process for cases involving substance use in adolescents [27]. An example of a web interface and the results are shown in Figure S2 in Multimedia Appendix 1. Custom code for the website is available on the web [28].

Discussion

Key Findings

This study stands out as one of the first comprehensive ML-based approaches to predict adolescent substance use on an international scale. One of our critical insights revolves around the influence of cultural diversity on substance use, drawing datasets from South Korea, the United States, and Norway [21]. Moreover, the outcome revealed that the XGBoost model proves to be commendable. It displayed predictive capabilities with an AUROC score of 80.61% (95% CI 79.63-81.59) and a precision of 30.42% (95% CI 28.65-32.16) in the discovery dataset. Our model consistently exhibited robust performance across external validation sets, achieving AUROC scores of 79.30% and 76.39% and precision of 68.37% and 12.74 in each respective dataset. Upon closer inspection, we discerned pivotal features influencing adolescent substance use predictions. Smoking status emerged as the predominant predictor for substance use, followed by BMI and suicidal thinking. Furthermore, the SHAP value analysis confirmed smoking status as a critical variable, with alcoholic consumption and sadness and despair identified as additional influential factors. To apply our findings to real-world scenarios, we devised a cutting-edge web-based platform. We believe that this tool will serve as an insightful methodology for the public to navigate potential substance-related challenges.

Plausible Mechanism

The influence of smoking status and alcohol consumption on adolescent substance use is noteworthy and warrants consideration. Neurobiological evidence indicates that nicotine, especially when introduced during formative years, can alter the brain reward pathways, making other substances more appealing [29]. Similarly, alcohol can modify neurotransmitter levels during these formative years, making the brain more susceptible to effects from other substances [30,31]. Smoking and drinking during adolescence often signal risk-taking behaviors [32], leading to further experimentation with other substances.

Both smoking and alcoholic consumption align closely with societal expectations and peer pressures. Societal dynamics and peer interactions, which may normalize or even perceive smoking and drinking as socially acceptable, could act as influential factors in encouraging adolescents to engage in these behaviors [33,34]. In many cultures, both behaviors are viewed as rites of passage that expose adolescents to other available illicit substances [35].

Psychological factors also play a role. Many adolescents resort to smoking or drinking as coping mechanisms for stress or emotional turmoil [36,37]. These initial coping behaviors may prompt adolescents to seek stronger stimuli for more intense experiences, potentially resulting in substance addiction or misuse [38].

This study further emphasized the relationship between BMI and substance use, as another predictor of adolescent substance use. From a physiological perspective, substances can modulate metabolic rates and appetite. Adolescents engaged in substance use might experience weight changes, either due to the direct effects of the substance or inconsistent dietary habits [39]. Additionally, the role of BMI extends beyond mere physiological changes. Adolescents with "nonstandard" BMIs often face societal challenges such as weight-centric bullying and entrenched societal ideals regarding body standards. The prevailing societal standards of an ideal body may influence some adolescents toward substance use, either as a means to conform to societal expectations or to address the associated mental distress [40].

Another noteworthy observation was the understated importance of academic achievement and stress status. Although stress is traditionally considered influential in adolescent behaviors [41,42], its limited representation in this study may be attributed to data constraints. These variables were absent in our extravalidation dataset, and we resorted to imputing these values using the median from our primary training cohort. This modification might have contributed to its reduced importance in our results.

Clinical and Policy Implications

The results of this study offer significant insights into both clinical and political implications. We underscore the vital role of factors such as smoking status, BMI, and alcoholic consumption in predicting substance use among adolescents. These critical determinants enable clinicians to identify and monitor at-risk adolescents more effectively, assisting in their decision-making process [43]. Following further refinement, this model has potential commercial viability [44], especially when combined with a streamlined self-report questionnaire. The existence of multiple models assessing substance use further attests to the commercial potential of our model [45]. Emphasizing characteristics predictive of substance use is essential, suggesting the need for systems to alert parents about potential risks their children might face. Since parental intervention has proven to be effective in preventing adolescent substance use [46], establishing an early detection system becomes paramount.

Strengths and Limitations

Findings from this study must be interpreted in light of several limitations. The external validation datasets contained numerous missing values, which could have impacted the predictive accuracy of the model. Specifically, we were unable to find corresponding data for variables like BMI (due to the absence of height and weight) and academic achievement in external cohorts such as Ungdata. Moreover, stress status, suicidal thinking, and suicide attempts were also missing or unmatched in datasets like YRBS and Ungdata, leading to gaps in these key areas. To address this gap, we applied imputation methods, including median imputation based on the KYRBS dataset for entirely missing variables [47]. While this approach mitigated the issue of missing data, it may have introduced biases, highlighting the need for more harmonized and comprehensive data collection protocols in future studies to enhance model generalizability and robustness. Additionally, this study used a discovery dataset derived from adolescents in South Korea. This biased discovery dataset could unexpectedly reflect the specific racial and cultural features unique to Korean adolescents. While our model underwent external validation from a diverse cultural and demographical landscape, we also acknowledge that it may reduce sample diversity and potentially cause overfitting issues [48]. Furthermore, this study did not pinpoint a definitive causal link between the significant risk factors and adolescent substance use. In other words, it remains unclear whether substance use influences other factors or if those factors stimulate substance use. Thus, further comprehensive studies are needed to elucidate this intricate cause-and-effect relationship. Another limitation of our study is the potential variability in feature importance rankings across different ML algorithms. Different algorithms may prioritize predictors differently due to their inherent characteristics and methods of handling data. This variability suggests that the identified predictors, such as smoking status, BMI, and alcohol consumption, should not be generalized as the sole determinants of adolescent substance use. Instead, these results should be interpreted with caution, and further studies are needed to validate the findings across diverse models and datasets. Finally, this study identifies factors associated with current and past substance use rather than explicitly predicting future trends. While the model provides valuable insights into risk factors and enables early intervention strategies, its predictive performance may be limited by the imbalanced nature of the dataset and the lack of longitudinal data. Future studies should address these limitations by incorporating balanced datasets and temporal data to enhance predictive accuracy and generalizability.

Despite these limitations, this study offers significant contributions. By using extensive datasets from South Korea, the United States, and Norway, our ML model boasts enhanced prediction accuracy, highlighting its global relevance and robustness [49]. Our strategic phased validation approach, beginning with the YRBS and progressing to the distinct Ungdata from Norway, underlines the model's versatility across diverse sociocultural backgrounds. This phased validation not only ensures consistent model evaluation but also establishes its capability in different cultural contexts [45,50,51]. Moreover, the features incorporated into the model are derived from simple questionnaires. The primary advantage of our model-based platform is its exceptional accessibility, allowing users to gather insights through straightforward surveys. This ease enables swift evaluations and widens its scope of use, equipping both clinicians and individuals with valuable insights conveniently. Our findings provide the relative importance of numerous factors. These results can guide the decision-making process by identifying key areas for the prevention of substance use among adolescents.

Conclusions

This study introduced an ML model using data from three distinct national cohorts to predict adolescent substance use. Among six unique predictive models, the XGBoost model consistently revealed a notable performance (AUROC: KYRBS, 80.61% [discovery]; YRBS, 79.30% [extravalidation]; and Ungdata, 76.39% [extravalidation], and precision: KYRBS, 30.42% [discovery]; YRBS, 68.37% [extravalidation]; and Ungdata, 12.74% [extravalidation]). Feature importance analysis identified smoking status, BMI, and suicidal thinking as significant contributors to the risk of substance use. Further insights into the influence of these variables were derived from SHAP value analysis, which identified smoking status, alcoholic consumption, and sadness and despair as the most impactful factors, in that order. The findings of this study indicate the potential of ML-driven predictive models to swiftly predict the likelihood of substance use among adolescents using a simplistic survey. It is anticipated that with further refinement and development, these models could be broadly used as efficient tools for preventing adolescent substance use.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education (RS-2024-00460379), the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (RS-2024-00509257, Global AI Frontier Lab), and the Information Technology Research Center support program (IITP-2024-RS-2024-00438239) supervised by the IITP. The funders had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Data Availability

The datasets generated during or analyzed during this study are available from the corresponding author upon reasonable request. The study protocol and statistical code are available from DKY. The dataset can be accessed from the Korean Disease Control and Prevention Agency, US Centers for Disease Control and Prevention, and Norwegian Social Research Institute (NOVA) through a data use agreement.

Authors' Contributions

DKY had full access to all of the data in the study and took responsibility for the integrity of the data and the accuracy of the data analysis. All authors approved the final version before submission. The study concept and design were contributed by Soeun Kim, HK, Seokjun Kim, SW, and DKY. Data acquisition, analysis, or interpretation were conducted by HK, SW, and DKY. The manuscript was drafted by HK, SW, and DKY. All authors critically revised the manuscript for important intellectual content. Statistical analysis was performed by HK, SW, and DKY. Study supervision was conducted by SW and DKY. DKY supervised the study and is the guarantor for this study. Soeun Kim, HK, and Seokjun Kim contributed equally as cofirst authors. DKY, SW, and RK contributed equally as cocorresponding authors. The corresponding author attests that all listed authors meet the authorship criteria and that no others meeting the criteria have been omitted.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Additional material.
[DOCX File , 498 KB-Multimedia Appendix 1]

References

1. Davidson LL, Grigorenko EL, Boivin MJ, Rapa E, Stein A. A focus on adolescence to reduce neurological, mental health and substance-use disability. Nature. 2015;527(7578):S161-S166. [doi: 10.1038/nature16030] [Medline: 26580322]
2. Volkow ND, Wargo EM. Association of severity of adolescent substance use disorders and long-term outcomes. JAMA Netw Open. 2022;5(4):e225656. [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.5656] [Medline: 35363272]

3. Williams EC, Fletcher OV, Frost MC, Harris AHS, Washington DL, Hoggatt KJ. Comparison of substance use disorder diagnosis rates from electronic health record data with substance use disorder prevalence rates reported in surveys across sociodemographic groups in the veterans health administration. JAMA Netw Open. 2022;5(6):e2219651. [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.19651] [Medline: 35771574]
4. Chaffee BW, Cheng J, Couch ET, Hoeft KS, Halpern-Felsher B. Adolescents' substance use and physical activity before and during the COVID-19 pandemic. JAMA Pediatr. 2021;175(7):715-722. [FREE Full text] [doi: 10.1001/jamapediatrics.2021.0541] [Medline: 33938922]
5. Kim J, Lee H, Lee J, Rhee SY, Shin JI, Lee SW, et al. Quantification of identifying cognitive impairment using olfactory-stimulated functional near-infrared spectroscopy with machine learning: a post hoc analysis of a diagnostic trial and validation of an external additional trial. Alzheimers Res Ther. 2023;15(1):127. [FREE Full text] [doi: 10.1186/s13195-023-01268-9] [Medline: 37481573]
6. Lee H, Cho JK, Park J, Lee H, Fond G, Boyer L, et al. Machine learning-based prediction of suicidality in adolescents with allergic rhinitis: derivation and validation in 2 independent nationwide cohorts. J Med Internet Res. 2024;26:e51473. [FREE Full text] [doi: 10.2196/51473] [Medline: 38354043]
7. Lam JY, Shimizu C, Tremoulet AH, Bainto E, Roberts SC, Sivilay N, et al. A machine-learning algorithm for diagnosis of multisystem inflammatory syndrome in children and Kawasaki disease in the USA: a retrospective model development and validation study. Lancet Digital Health. 2022;4(10):e717-e726. [FREE Full text] [doi: 10.1016/s2589-7500(22)00149-2]
8. Saux P, Bauvin P, Raverdy V, Teigny J, Verkindt H, Soumphonphakdy T, et al. Development and validation of an interpretable machine learning-based calculator for predicting 5-year weight trajectories after bariatric surgery: a multinational retrospective cohort SOPHIA study. Lancet Digital Health. 2023;5(10):e692-e702. [doi: 10.1016/s2589-7500(23)00135-8]
9. Clift AK, Dodwell D, Lord S, Petrou S, Brady M, Collins GS, et al. Development and internal-external validation of statistical and machine learning models for breast cancer prognostication: cohort study. BMJ. 2023;381:e073800. [FREE Full text] [doi: 10.1136/bmj-2022-073800] [Medline: 37164379]
10. Woo HG, Park S, Yon H, Lee SW, Koyanagi A, Jacob L, et al. National trends in sadness, suicidality, and COVID-19 pandemic-related risk factors among South Korean adolescents from 2005 to 2021. JAMA Netw Open. 2023;6(5):e2314838. [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.14838] [Medline: 37223902]
11. Boakye E, Erhabor J, Obisesan O, Tasdighi E, Mirbolouk M, Osuji N, et al. Comprehensive review of the national surveys that assess E-cigarette use domains among youth and adults in the United States. Lancet Reg Health Am. 2023;23:100528. [FREE Full text] [doi: 10.1016/j.lana.2023.100528] [Medline: 37497394]
12. Leonhardt M, Granrud MD, Bonsaksen T, Lien L. Associations between mental health, lifestyle factors and worries about climate change in Norwegian adolescents. Int J Environ Res Public Health. 2022;19(19):12826. [FREE Full text] [doi: 10.3390/ijerph191912826] [Medline: 36232127]
13. Kwon R, Lee H, Kim MS, Lee J, Yon DK. Machine learning-based prediction of suicidality in adolescents during the COVID-19 pandemic (2020-2021): derivation and validation in two independent nationwide cohorts. Asian J Psychiatr. 2023;88:103704. [doi: 10.1016/j.ajp.2023.103704] [Medline: 37541104]
14. Pelgrims I, Devleesschauwer B, Vandevijvere S, De Clercq EM, Vansteelandt S, Gorasso V, et al. Using random-forest multiple imputation to address bias of self-reported anthropometric measures, hypertension and hypercholesterolemia in the Belgian health interview survey. BMC Med Res Methodol. 2023;23(1):69. [FREE Full text] [doi: 10.1186/s12874-023-01892-x] [Medline: 36966305]
15. Johnston KC, Connors AF, Wagner DP, Haley EC. Predicting outcome in ischemic stroke: external validation of predictive risk models. Stroke. Jan 2003;34(1):200-202. [FREE Full text] [doi: 10.1161/01.str.0000047102.61863.e3] [Medline: 12511774]
16. Nguipdop-Djomo P, Rodrigues LC, Smith PG, Abubakar I, Mangtani P. Drug misuse, tobacco smoking, alcohol and other social determinants of tuberculosis in UK-born adults in England: a community-based case-control study. Sci Rep. 2020;10(1):5639. [FREE Full text] [doi: 10.1038/s41598-020-62667-8] [Medline: 32221405]
17. Compton WM, Flannagan KSJ, Silveira ML, Creamer MR, Kimmel HL, Kanel M, et al. Tobacco, alcohol, cannabis, and other drug use in the US before and during the early phase of the COVID-19 pandemic. JAMA Netw Open. 2023;6(1):e2254566. [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.54566] [Medline: 36719678]
18. Huang J. Drug licensing as evidence of evolution, diffusion and catch-up in East Asia. Nat Biotechnol. 2023;41(2):189-192. [doi: 10.1038/s41587-023-01659-1] [Medline: 36792699]
19. Garriga R, Mas J, Abraha S, Nolan J, Harrison O, Tadros G, et al. Machine learning model to predict mental health crises from electronic health records. Nat Med. 2022;28(6):1240-1248. [FREE Full text] [doi: 10.1038/s41591-022-01811-5] [Medline: 35577964]
20. Montgomery-Csobán T, Kavanagh K, Murray P, Robertson C, Barry SJE, Ukah UV, et al. Machine learning-enabled maternal risk assessment for women with pre-eclampsia (the PIERS-ML model): a modelling study. Lancet Digital Health. 2024;6(4):e238-e250. [doi: 10.1016/s2589-7500(23)00267-4]
21. Kim H, Son Y, Lee H, Kang J, Hammoodi A, Choi Y, et al. Machine learning-based prediction of suicidal thinking in adolescents by derivation and validation in 3 independent worldwide cohorts: algorithm development and validation study. J Med Internet Res. 2024;26:e55913. [FREE Full text] [doi: 10.2196/55913] [Medline: 38758578]

22. Bolo K, Aroca GAA, Pardeshi AA, Chiang M, Burkemper B, Xie X, et al. Automated expert-level scleral spur detection and quantitative biometric analysis on the ANTERION anterior segment OCT system. Br J Ophthalmol. 2024;108(5):702-709. [FREE Full text] [doi: 10.1136/bjo-2022-322328] [Medline: 37798075]
23. Hwang SH, Lee H, Lee JH, Lee M, Koyanagi A, Smith L, et al. Machine learning-based prediction for incident hypertension based on regular health checkup data: derivation and validation in 2 independent nationwide cohorts in South Korea and Japan. J Med Internet Res. 2024;26:e52794. [FREE Full text] [doi: 10.2196/52794] [Medline: 39499554]
24. Sang H, Lee H, Lee M, Park J, Kim S, Woo HG, et al. Prediction model for cardiovascular disease in patients with diabetes using machine learning derived and validated in two independent Korean cohorts. Sci Rep. 2024;14(1):14966. [FREE Full text] [doi: 10.1038/s41598-024-63798-y] [Medline: 38942775]
25. Thorsen-Meyer H, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. Lancet Digital Health. 2020;2(4):e179-e191. [doi: 10.1016/s2589-7500(20)30018-2]
26. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594. [FREE Full text] [doi: 10.1136/bmj.g7594] [Medline: 25569120]
27. Probability for substance use. Streamlit. URL: https://predictsubstance.streamlit.app/ [accessed 2025-01-28]
28. CenterForDH/substance. GitHub. URL: https://github.com/centerfordh/predict_substance/ [accessed 2025-01-28]
29. Le Foll B, Piper ME, Fowler CD, Tonstad S, Bierut L, Lu L, et al. Tobacco and nicotine use. Nat Rev Dis Primers. 2022;8(1):19. [doi: 10.1038/s41572-022-00346-w] [Medline: 35332148]
30. Yip SW, Lichenstein SD, Liang Q, Chaarani B, Dager A, Pearlson G, et al. Brain networks and adolescent alcohol use. JAMA Psychiatry. 2023;80(11):1131-1141. [doi: 10.1001/jamapsychiatry.2023.2949] [Medline: 37647053]
31. Kang J, Kim HJ, Kim T, Lee H, Kim M, Lee SW, et al. Prenatal opioid exposure and subsequent risk of neuropsychiatric disorders in children: nationwide birth cohort study in South Korea. BMJ. 2024;385:e077664. [FREE Full text] [doi: 10.1136/bmj-2023-077664] [Medline: 38658035]
32. Kruckow S, Santini ZI, Hjarnaa L, Becker U, Andersen O, Tolstrup JS. Associations between alcohol intake and hospital contacts due to alcohol and unintentional injuries in 71,025 Danish adolescents—a prospective cohort study. EClinicalMedicine. 2023;64:102187. [FREE Full text] [doi: 10.1016/j.eclinm.2023.102187] [Medline: 37936661]
33. Zhao Y, Di X, Li S, Zeng X, Wang X, Nan Y, et al. Prevalence, frequency, intensity, and location of cigarette use among adolescents in China from 2013-14 to 2019: findings from two repeated cross-sectional studies. Lancet Reg Health West Pac. 2022;27:100549. [FREE Full text] [doi: 10.1016/j.lanwpc.2022.100549] [Medline: 35923777]
34. Kim S, Lee H, Lee J, Lee SW, Kwon R, Kim MS, et al. Short- and long-term neuropsychiatric outcomes in long COVID in South Korea and Japan. Nat Hum Behav. 2024;8(8):1530-1544. [doi: 10.1038/s41562-024-01895-8] [Medline: 38918517]
35. Maeng SJ, Lee DJ, Kang JH. First drinking experiences during adolescence in South Korea: a qualitative study focusing on the internal and external factors. Int J Environ Res Public Health. 2021;18(15):8200. [FREE Full text] [doi: 10.3390/ijerph18158200] [Medline: 34360493]
36. Meienberg A, Mayr M, Vischer A, Zellweger MJ, Burkard T. Smoking prevention in adolescents: a cross-sectional and qualitative evaluation of a newly implemented prevention program in Switzerland. BMJ Open. 2021;11(12):e048319. [doi: 10.1136/bmjopen-2020-048319]
37. Skylstad V, Babirye JN, Kiguli J, Skar AMS, Kühl MJ, Nalugya JS, et al. Are we overlooking alcohol use by younger children? BMJ Paediatr Open. 2022;6(1):e001242. [FREE Full text] [doi: 10.1136/bmjpo-2021-001242] [Medline: 36053657]
38. Cheron J, d'Exaerde ADK. Drug addiction: from bench to bedside. Transl Psychiatry. 2021;11(1):424. [FREE Full text] [doi: 10.1038/s41398-021-01542-0] [Medline: 34385417]
39. Treasure J, Duarte TA, Schmidt U. Eating disorders. Lancet. 2020;395(10227):899-911. [FREE Full text] [doi: 10.1016/s0140-6736(20)30059-3]
40. Bornioli A, Lewis-Smith H, Slater A, Bray I. Body dissatisfaction predicts the onset of depression among adolescent females and males: a prospective study. J Epidemiol Community Health. 2020;75(4):343-348. [doi: 10.1136/jech-2019-213033] [Medline: 33288655]
41. Sylvestre MP, Dinkou GDT, Naja M, Riglea T, Pelekanakis A, Bélanger M, et al. A longitudinal study of change in substance use from before to during the COVID-19 pandemic in young adults. Lancet Reg Health Am. 2022;8:100168. [FREE Full text] [doi: 10.1016/j.lana.2021.100168] [Medline: 35469267]
42. Liu M, Koh KA, Hwang SW, Wadhera RK. Mental health and substance use among homeless adolescents in the US. JAMA. 2022;327(18):1820-1822. [FREE Full text] [doi: 10.1001/jama.2022.4422] [Medline: 35536272]
43. Watson DS, Krutzinna J, Bruce IN, Griffiths CE, McInnes IB, Barnes MR, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ. 2019;364:l886. [FREE Full text] [doi: 10.1136/bmj.l886] [Medline: 30862612]
44. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. Lancet Digital Health. 2021;3(3):e195-e203. [FREE Full text] [doi: 10.1016/s2589-7500(20)30292-2]

45. Lo-Ciganic WH, Donohue JM, Yang Q, Huang JL, Chang CY, Weiss JC, et al. Developing and validating a machine-learning algorithm to predict opioid overdose in medicaid beneficiaries in two US states: a prognostic modelling study. Lancet Digital Health. 2022;4(6):e455-e465. [FREE Full text] [doi: 10.1016/s2589-7500(22)00062-0]
46. Kuntsche S, Kuntsche E. Parent-based interventions for preventing or reducing adolescent substance use—a systematic literature review. Clin Psychol Rev. 2016;45:89-101. [doi: 10.1016/j.cpr.2016.02.004] [Medline: 27111301]
47. Pokorney SD, Holmes DN, Thomas L, Fonarow GC, Kowey PR, Reiffel JA, et al. Association between warfarin control metrics and atrial fibrillation outcomes in the outcomes registry for better informed treatment of atrial fibrillation. JAMA Cardiol. 2019;4(8):756-764. [FREE Full text] [doi: 10.1001/jamacardio.2019.1960] [Medline: 31268487]
48. Blackburn AM, Vestergren S, COVIDiSTRESS II Consortium. Author Correction: COVIDiSTRESS diverse dataset on psychological and behavioural outcomes one year into the COVID-19 pandemic. Sci Data. 2023;10(1):12. [FREE Full text] [doi: 10.1038/s41597-022-01896-0] [Medline: 36599849]
49. Riley RD, Snell KIE, Altman DG, Collins GS. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. Jun 25, 2019;365:l4379. [FREE Full text] [doi: 10.1136/bmj.l4379] [Medline: 31239248]
50. Kaushal A, Altman R, Langlotz C. Geographic distribution of US cohorts used to train deep learning algorithms. JAMA. 2020;324(12):1212-1213. [FREE Full text] [doi: 10.1001/jama.2020.12067] [Medline: 32960230]
51. Jo H, Park J, Lee H, Lee K, Lee H, Son Y, et al. Nationwide trends in sadness, suicidal ideation, and suicide attempts among multicultural and monocultural adolescents in South Korea during the COVID-19 pandemic, 2011-2022. World J Pediatr. 2024;20(12):1249-1269. [doi: 10.1007/s12519-024-00858-3] [Medline: 39614994]

Abbreviations

AUPRC: area under precision-recall curve
AUROC: area under receiver operating characteristic curve
CDC: Centers for Disease Control and Prevention
KDCA: Korean Disease Control and Prevention Agency
KYRBS: Korea Youth Risk Behavior Web-Based Survey
ML: machine learning
NOVA: Norwegian Social Research Institute
SHAP: SHapley Additive exPlanation
TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis
YRBS: Youth Risk Behavior Survey

Edited by F Xie; submitted 31.05.24; peer-reviewed by H Lin, K-S Lee, K Kim; comments to author 19.08.24; revised version received 27.12.24; accepted 16.01.25; published 24.02.25

Please cite as:
Kim S, Kim H, Kim S, Lee H, Hammoodi A, Choi Y, Kim HJ, Smith L, Kim MS, Fond G, Boyer L, Baik SW, Lee H, Park J, Kwon R, Woo S, Yon DK
Machine Learning–Based Prediction of Substance Use in Adolescents in Three Independent Worldwide Cohorts: Algorithm Development and Validation Study
J Med Internet Res 2025;27:e62805
URL: https://www.jmir.org/2025/1/e62805
doi: 10.2196/62805
PMID: