Data-Driven Analysis of Social and Demographic Determinants of Youth Substance Use in Canada

Anjiya Nooruddin
University of Waterloo
Waterloo, Canada
a24adil@uwaterloo.ca

Chavvi Bhatia
University of Waterloo
Waterloo, Canada
cbhatia@uwaterloo.ca

Chirag Seth
University of Waterloo
Waterloo, Canada
cseth@uwaterloo.ca

Hrishita Sharma
University of Waterloo
Waterloo, Canada
h56sharm@uwaterloo.ca

Utkarsh Singh
University of Waterloo
Waterloo, Canada
u25singh@uwaterloo.ca

Abstract

This study investigates the social and demographic determinants of substance use among Canadian students in grades 7–12 using data from the 2018/2019 Canadian Student Tobacco, Alcohol, and Drugs Survey (CSTADS). Leveraging a cleaned dataset and Bayesian logistic regression models, we identify key predictors of tobacco, alcohol, and drug use, with a particular focus on the role of social influence. By controlling for confounding variables such as grade, sex, and household income, our analysis offers data-driven insights into how demographic and social factors jointly shape substance use behaviors in adolescents. The findings aim to inform targeted, evidence-based prevention strategies for Canadian youth.

1 Literature Review

Prior studies have explored various predictors of adolescent substance use, including parental monitoring and trends in usage of drugs in the early age of youths. For instance, [2] examined associations between mental health and cannabis use, reporting significant correlations, but their cross-sectional design limited causal interpretation, and confounders like social influence were not adequately controlled. Also, [5] used the 2014/15 CSTADS dataset to examine differences in tobacco, alcohol, and marijuana use between Indigenous and non-Indigenous students in grades 9–12. They applied descriptive statistics, chi-square tests, and multivariable logistic regression to identify disparities and predictors. The study found significantly higher substance use rates among Indigenous youth, with social influence and perceived risk as strong factors. However, it was limited by a uniform binge drinking threshold, exclusion of on-reserve youth, and lack of deeper analysis into social determinants.

Leatherdale *et al.* [9] analyzed the 2008/2009 YSS data, reporting high substance use prevalence using descriptive statistics, but lacked multivariate analysis for confounders. Sikorski *et al.* [15] used logistic regression on 2014/2015 CSTADS to show higher substance use among Indigenous youth, yet omitted socioeconomic factors and advanced modeling. Many studies use single-level logistic regression models, which do not account for nested structures in school and community environments. Montreuil *et al.* [11] applied logistic regression to 2014/2015 CSTADS, linking e-cigarette use to social networks, but focused narrowly on e-cigarettes without hierarchical models. Zuckermann *et al.* [17] employed multilevel logistic regression on COMPASS and CSTADS data, identifying school-level cannabis trends, though it used pre-2018 data and underexplored peer influence. Lowry and Corsi [10] conducted logistic regression on CTADS 2004–2017, finding youth cannabis use trends, but diluted focus on grades 7–12 and ignored social variables like bullying. Arbour-Nicitopoulos *et al.* [1] used generalized linear models on 2010/2011 YSS to link physical activity with reduced smoking, missing peer and mental health factors. Qian *et al.* [13] applied logistic regression to 2016/2017 CSTADS, associating vaping with tobacco use, but lacked causal inference or Bayesian methods. Elton-Marshall *et al.* [4] utilized mixed-effects models on COMPASS data to study cannabis co-use, yet overlooked post-legalization shifts and peer influence quantification. Patte *et al.* [12] employed multilevel modeling on COMPASS to correlate mental health with substance use, but did not integrate socioeconomic or bullying variables. Finally, Herciu *et al.* [8] used logistic regression on 2012/2013 YSS to examine bullying and smoking, missing recent data and advanced techniques like Lasso regression. These studies collectively fall short in using current 2018/2019 CSTADS data, fully exploring peer influence (e.g., PP_010), and applying sophisticated methods like Bayesian regression for probabilistic and causal insights. Most rely on basic logistic regression or descriptive approaches, neglecting feature selection (e.g., Lasso) or model comparison (e.g., BIC/DIC), limiting predictive accuracy and generalizability. The proposed study addresses these gaps by integrating a broad predictor set—demographic, social, behavioral, and mental health factors—using the latest CSTADS data and a suite of advanced statistical tools (Bayesian, Lasso, multilevel variable regression). By measuring the impact of peer influence and evaluating different modeling approaches, this study

seeks to provide deeper insights into substance use patterns. The goal is to improve our understanding and support the development of targeted interventions, addressing a crucial gap in existing research.

2 Methodology

2.1 Dataset

We use the 2018/2019 Canadian Student Tobacco, Alcohol, and Drugs Survey (CSTADS), which includes responses from over 60,000 students in grades 7–12 across Canada.

2.2 Exploratory Data Analysis

2.2.1 Research Question 1: How do demographic and social factors influence the probability of tobacco use among Canadian students in grades 7–12, based on the 2018/2019 CSTADS?. The analysis revealed significant demographic patterns in smoking behavior. Males were more likely to smoke, particularly when exposed to bullying or peer pressure. Household income distribution was right-skewed, with a larger proportion of students from lower-income areas and a few high-income outliers (above $120,000). Urban versus rural differences were statistically significant ($p \approx 0$), suggesting that location and local social norms may influence tobacco use.

Behavioral and social factors played a critical role. Students who experienced physical bullying were significantly more likely to smoke ($\chi^2 = 187.2893$, $p < 0.0001$). Smokers tended to perceive smoking as less risky than non-smokers, indicating a degree of desensitization. Most notably, social influence was strongly associated with smoking behavior—99.98% of smokers reported being socially influenced, compared to only 54.92% of non-smokers—highlighting its near-universal presence among youth smokers.

Substance use trends showed that the average age of smoking initiation was around 14, with socially influenced students tending to start even earlier. Smoking rates were highest in grades 7–8 and declined in later grades. Alcohol use followed a similar decline, while methamphetamine and cocaine usage remained relatively stable across grades.

2.2.2 Research Question 2: What is the association between social influence and the probability of tobacco, alcohol, and drug use among Canadian students, accounting for demographic confounders? Social influence had a strong association with smoking, particularly among males, who appeared more likely to use smoking as a coping mechanism for social stressors such as bullying. Interestingly, students in grade 7 showed relative resistance to peer pressure, but susceptibility increased in higher grades, suggesting that vulnerability to social influence grows with age.

For alcohol use, students who were socially influenced had a 50% likelihood of drinking, compared to only 10% among those without social influence ($p < 0.0001$). However, the presence of alcohol use among some uninfluenced students points to other underlying factors at play. In terms of illicit substances, bullied students were found to be 3.6 times more likely to use meth (11% vs. 3%), and cocaine use, while low overall, exhibited similar social-influence patterns.

Interaction effects further highlighted subgroup-specific dynamics. Males who experienced bullying had higher smoking rates than their female counterparts, underscoring gendered responses to stress. Additionally, early initiators (under age 14) were almost exclusively influenced by social factors, while older initiators demonstrated more diverse behavioral triggers.

2.3 Modeling

We employed five modeling techniques to evaluate the influence of demographic and social variables on youth substance use: Bayesian Logistic Regression (BLR) [6], Lasso-Penalized Logistic Regression [16], Multilevel Logistic Regression [7], Propensity Score Matching (PSM) [14], and Random Forest [3]. Each approach captures different aspects of the data: uncertainty quantification (Bayesian), feature selection (Lasso), hierarchical structure (Multilevel), causal inference (PSM), and variable importance (Random Forest).

2.3.1 Bayesian Logistic Regression (BLR). We modeled the binary outcome $Y_i \in \{0, 1\}$ (e.g., smoking) using a logistic function:

$$P(Y_i = 1 \mid \mathbf{X}_i, \boldsymbol{\beta}) = \frac{1}{1 + \exp(-\mathbf{X}_i^\top \boldsymbol{\beta})}$$

with priors $\beta_j \sim \mathcal{N}(0, \sigma^2)$. The posterior distribution,

$$p(\boldsymbol{\beta} \mid \mathcal{D}) \propto p(\mathcal{D} \mid \boldsymbol{\beta}) \cdot p(\boldsymbol{\beta}),$$

was estimated via the No-U-Turn Sampler (NUTS) within a Markov Chain Monte Carlo (MCMC) framework, allowing inference over 95% credible intervals.

2.3.2 Lasso-Penalized Logistic Regression. To identify the most predictive features, we applied Lasso-regularized logistic regression:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} Y_i \log \hat{p}_i + (1 - Y_i) \log(1 - \hat{p}_i) - \lambda \sum_{j=1}^{p} |\beta_j| \right]$$

with $\hat{p}_i = \frac{1}{1+\exp(-\mathbf{X}_i^\top \boldsymbol{\beta})}$ and λ tuned using 5-fold cross-validation. The Lasso penalty shrinks non-informative features to zero.

2.3.3 Multilevel Logistic Regression. To account for nested structures (e.g., students within schools), we used a random-intercept multilevel model:

$$\text{logit}(P(Y_{ij} = 1)) = \beta_0 + u_j + \sum_{k=1}^{p} \beta_k x_{ijk}$$

where $u_j \sim \mathcal{N}(0, \tau^2)$ captures group-level heterogeneity. This enhances generalization across schools or regions.

2.3.4 Propensity Score Matching (PSM). To estimate causal effects of peer influence, we implemented propensity score matching. For each student, we computed the probability of receiving treatment (i.e., being socially influenced), denoted:

$$e(\mathbf{X}_i) = P(T_i = 1 \mid \mathbf{X}_i)$$

where T_i is the treatment indicator. Treated and control units were then matched based on similar $e(\mathbf{X}_i)$ using nearest-neighbor matching without replacement. The average treatment effect on the treated (ATT) was computed by:

$$\text{ATT} = \frac{1}{n_T} \sum_{i:T_i=1} (Y_i - Y_i^{\text{matched}})$$

2.3.5 Random Forest. To explore nonlinear patterns and variable importance, we used a Random Forest classifier with B decision trees. Each tree was built on a bootstrapped sample of the data and a random subset of features. Prediction was done via majority voting:

$$\hat{Y}_i = \text{majority}\{T_b(\mathbf{X}_i)\}_{b=1}^{B}$$

Feature importance was ranked based on Gini impurity reduction aggregated across all trees.

2.3.6 Model Comparison and Evaluation. We compared models using the following metrics:

- **AUC (Area Under the ROC Curve)** – to assess classification performance.
- **WAIC and DIC** – for Bayesian model fit.
- **Sparsity and interpretability** – for Lasso.
- **ATT** – for estimating causal effect via PSM.
- **Feature importance and accuracy** – for Random Forest.

Overall, this modeling pipeline combines inference, prediction, and causal estimation to provide a comprehensive analysis of substance use predictors among Canadian youth.

3 Results

Our analysis identified key predictors of youth substance use across multiple modeling frameworks. Bayesian logistic regression showed that higher grade ($\beta = 0.759$), male sex ($\beta = 0.123$), and lower household income ($\beta = -0.179$) significantly increased the likelihood of smoking. The model exhibited a good fit with WAIC = −13901.97 and LOO = −13901.98 (all Pareto $k \leq 0.7$).

Lasso regression models demonstrated strong predictive performance, particularly for alcohol use (AUC = 0.7779), where mental health was associated with increased risk (OR = 1.27) and social influence appeared protective (OR = 0.17). For cocaine use (AUC = 0.7313), social influence increased the risk (OR = 1.10), while bullying had also slightly increased the risk (OR = 0.18). In the case of methamphetamine (AUC = 0.7142), grade level (OR = 1.40) and urban residence (OR = 1.21) emerged as significant predictors.

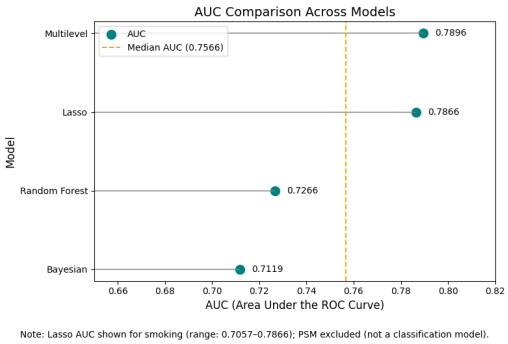

Figure 1. AUC Comparison Across Models. Multilevel logistic regression achieved the highest AUC (0.7896), followed closely by Lasso (0.7866), while Random Forest and Bayesian regression had AUCs of 0.7266 and 0.7119, respectively. The dashed line marks the median AUC (0.7566). PSM is excluded as it is not a classification model.

Multilevel logistic regression further emphasized the overwhelming role of social influence, with an odds ratio of 951.35 ($p < 0.001$), and also highlighted mental health (OR = 1.68) and bullying (OR = 1.14) as statistically significant. These findings were consistent with propensity score matching, which estimated that social influence increased smoking risk by 0.52 units.

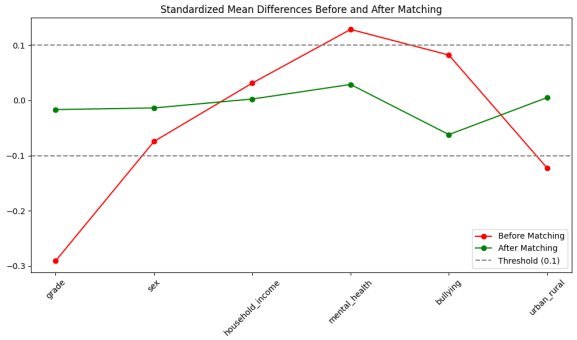

Figure 2. Standardized Mean Differences Before and After Matching. The plot shows covariate balance across six variables before (red) and after (green) propensity score matching. All post-matching values fall below the 0.1 threshold, indicating effective bias reduction and improved comparability between treatment and control groups.

Additionally, Random Forest modeling ranked grade (importance = 0.41), mental health (0.18), and social influence (0.16) as the top predictors, achieving an overall classification accuracy of 0.73.

Across all models, social influence consistently emerged as the most dominant predictor. This was also reflected in bivariate associations, where 99.98% of smokers reported

being socially influenced, compared to only 54.92% of non-smokers. A chi-square test confirmed that students who experienced bullying were significantly more likely to smoke ($\chi^2 = 187.29$, $p < 0.001$).

4 Discussion

4.1 Key Findings

Social influence emerged as one of the strongest predictors of tobacco and drug use. Gender differences were notable, with males showing slightly higher usage rates. Higher household income was negatively correlated with tobacco use but showed mixed results for alcohol consumption.

4.2 Implications

These findings emphasize the importance of peer-focused interventions in schools. Programs aimed at reshaping peer norms around substance use may prove effective. Socioeconomic disparities also highlight the need for tailored outreach in lower-income communities.

4.3 Limitations

Limitations include self-reported data, which may introduce bias, and the cross-sectional nature of the dataset, which limits causal inference.

References

[1] K. P. Arbour-Nicitopoulos and G. Faulkner. 2013. Physical Activity and Smoking Among Canadian Youth. *Journal of Adolescent Health* 52, 5 (2013), 614–620.

[2] Aharon Ben-Tal and Arkadi Nemirovski. 1998. Robust convex optimization. *Mathematics of operations research* 23, 4 (1998), 769–805.

[3] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[4] T. Elton-Marshall and S. T. Leatherdale. 2020. Cannabis and Tobacco Co-Use Among Youth. *Drug and Alcohol Dependence* 214 (2020), 108152.

[5] Tara Elton-Marshall, Scott T. Leatherdale, Robin Burkhalter, and Katelyn S. Brown. 2019. Tobacco, alcohol and marijuana use among Indigenous and non-Indigenous students in grades 9 to 12. *Health Promotion and Chronic Disease Prevention in Canada* 39, 3 (2019), 87–95.

[6] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian Data Analysis*. CRC press.

[7] Andrew Gelman and Jennifer Hill. 2006. Data analysis using regression and multilevel/hierarchical models. *Cambridge university press* (2006).

[8] A. C. Herciu and S. T. Leatherdale. 2014. Bullying Victimization and Smoking Among Youth. *Journal of School Health* 84, 8 (2014), 514–520.

[9] S. T. Leatherdale and R. Ahmed. 2012. The Substance Use Profile of Canadian Youth. *Addictive Behaviors* 37, 3 (2012), 225–230.

[10] R. Lowry and D. Corsi. 2020. Trends and Correlates of Cannabis Use in Canada. *CMAJ Open* 8, 4 (2020), E743–E750.

[11] A. Montreuil and J. L. MacDonald. 2017. Prevalence and Correlates of E-Cigarette Use Among Canadian Students. *CMAJ Open* 5, 2 (2017), E314–E321.

[12] K. A. Patte and S. T. Leatherdale. 2017. Mental Health and Substance Use Among Canadian Youth. *Health Promotion International* 32, 6 (2017), 1010–1020.

[13] W. Qian and R. Schwartz. 2019. Vaping and Tobacco Use Among Canadian Youth. *Canadian Journal of Public Health* 110, 4 (2019), 489–497.

[14] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.

[15] C. Sikorski, T. Lebel, and R. Afifi. 2019. Tobacco, Alcohol, and Marijuana Use Among Indigenous Youth. *Health Promotion and Chronic Disease Prevention in Canada* 39, 5 (2019), 150–157.

[16] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288. https://en.wikipedia.org/wiki/Lasso_(statistics)

[17] A. Zuckermann and S. T. Leatherdale. 2019. Prelegalisation Patterns and Trends of Cannabis Use Among Canadian Youth. *BMJ Open* 9, 10 (2019), e030451.