Predicting Factors of Teen Vape Use

Marisa Papagelis, Chandler Pettigrew, Sophie Rosas-Smith, Viki Zygouras

Wellesley College

Abstract:

This paper analyzes the factors contributing to teen vape use. Using data from the 2014-2018 New York Youth Tobacco Surveys, a classification tree was trained with eight factors to predict teen vape behavior and determine which contributed most to vaping.[1] Based on the analysis, the biggest contributing factors were having lived with a smoker, whether the student is in high school or middle school, and allowance money. Validation of the model yielded an accuracy of 74.13%.

[1] State of New York, *Youth Tobacco Survey: Beginning 2000*, (November 21, 2019), distributed by U.S. Department of Health & Human Services, https://healthdata.gov/dataset/youth-tobacco-survey-beginning-2000.

Background and Significance:

E-cigarettes, also known as “vapes” or “JUULs”, are defined as “electronic devices that heat a liquid and produce an aerosol, or mix of small particles in the air.”[2] Though e-cigarettes were introduced to the U.S. in the mid-2000’s, there has been a recent increase in teen e-cigarette use and related health concerns.[3] The Centers for Disease Control and Prevention confirmed “more than 450 possible cases and five confirmed deaths in 33 states and the U.S. Virgin Islands” resulting from e-cigarette use as of September 2019.[4]

Various studies have reported possible reasons for the rise in vaping among teenagers such as flavor, curiosity, and believing that e-cigarettes are less harmful than conventional cigarettes.[5] While there are studies that observe the health effects of vaping with a focus on traditional demographic data, there should also be a focus on investigating socioeconomic factors that have the potential to make certain teenagers more vulnerable to vape use. Identifying which factors contribute to teen vaping is important in understanding the rise of e-cigarettes and the challenges of improving public health. We hypothesize that teen vape use is influenced by the following factors: age, sex, race, whether a student attends high school or middle school, lives in New York City vs. New York State, lives with a smoker, and student income per week.

Methods:

A. Data collection: the New York Youth Tobacco Survey is a biennial survey that measures youth tobacco use, access, and perceptions. The number of participating schools in 2014, 2016, and 2018 were 72, 65, and 75 respectively. In 2018, schools were incentivized with $1000 grant to participate. In 2014, 2016, and 2018 the sample size of student respondents were 1,693, 1,452, and 1,637 respectively. Despite the survey having been conducted since 2000, the model uses data from 2014-2018 which coincides with the rise in teen vaping.
B. Variable Creation: Eight variables were hypothesized to contribute to teen vape use. These variables include: age, sex, race, attending high school vs. middle school, lives in NYC, lives with a smoker, student income per week, and year. The original survey data included over 100 response categories. One row corresponds to one student’s survey results.
C. Data cleaning: A new dataset was created and limited to the eight selected variables. Some questions in the survey were changed over time to aid student interpretation and response rate. Variables with consistent responses over the selected year range were carefully chosen for analysis. (i.e. ‘_race’ variable was empty but ‘_race2’ contained data within the year range). Due to the categorical nature of the variables, missing data was replaced with either an empty string or 999. Rows were not dropped for containing missing data. All string response values in the data set were converted to integers. The

---

[2] “Quick Facts on the Risks of E-cigarettes for Kids, Teens, and Young Adults,” cdc.gov, Centers for Disease Control and Prevention, November 26, 2019, https://www.cdc.gov/tobacco/basic_information/e-cigarettes.

[3] Ibid.

[4] J. Brainard. "CDC Warns Against Vaping," *Science* 365, no. 6458 (2019): 1062, 10.1126/science. doi: 365.6458.1062.

[5] U.S. Department of Health and Human Services. “E-Cigarette Use Among Youth and Young Adults: A Report of the Surgeon General,” 2016, https://www.cdc.gov/tobacco/data_statistics/sgr/e-cigarettes/pdfs/2016_sgr_entire_report_508.pdf.

year variable was constrained to 2014-2018 to produce a more relevant analysis of teen behavior.

D. Analytic Methods: To understand the driving factors of teen vape use, a binary classification tree was trained and tested with the dataset. The eight cleaned variables listed above were converted to factors and used to develop the results of the model. The cleaned data was split into randomly sampled training and testing groups with sizes of 3,000 and 1,780 responses respectively. According to the results of the tree, the most significant contributing factor to whether a student has vaped or not was if they live with a smoker with a p-value of less than 0.001 indicating strong statistical significance. Other important factors with p-values of less than 0.001 include the amount of weekly student income with amounts higher than $20 increasing predicted vape usage as much as 68%. High school students have an 8% higher prediction of vape use than middle school students. Two different implementations of the classification model was tested in R: ctree and rpart. The main difference is the criteria for splitting on a variable. Ultimately, the ctree algorithm yielded a 12% higher accuracy during validation than rpart and was selected for analysis.

Results:

The results were validated by measuring the accuracy of the predictions. Success rate, or accuracy, was calculated in two different ways. First, we checked equality in a 1:1 comparison of predicted value against actual value. We measured accuracy as the ratio of correctly predicted values to total predicted values. This resulted in an accuracy of 74.13%. An alternative validation method uses the residual values. The difference between the predicted values and the actual values were calculated. If the difference was equal to zero, the model correctly predicted the actual value. If the difference was not equal to zero, the value was incorrectly predicted. The total correctly predicted values were divided by total predicted values. The measured accuracy resulted in the same accuracy of 74.13%.

$$Accuracy = \frac{Number\ of\ correctly\ predicted\ values}{Total\ number\ of\ predicted\ values}$$

Conclusion:

The three biggest contributing factors to teen vape use are having lived with a smoker, whether the student is in high school or middle school, and student income per week. It is important to note that these results are not generalizable to the entire U.S. youth population. The survey was exclusively conducted in public schools in the state of New York which may not accurately represent students and teen tobacco use across the country. The fact that the survey was given to middle and high school students, who may not have answered the survey questions truthfully, is a potential source of sample bias. The survey, however, was multiple choice to prevent students from submitting a range of outrageous values. For potential future work, the model could be improved to include more variables and be trained on more years of youth tobacco data. Additionally, replicating our analysis with newer data from the latest survey could improve model accuracy.

References

Brainard, J. "CDC Warns Against Vaping." *Science* 365, no. 6458 (2019): 1062. doi: 10.1126/science.365.6458.1062.

“Quick Facts on the Risks of E-cigarettes for Kids, Teens, and Young Adults.” cdc.gov. Centers for Disease Control and Prevention, November 26, 2019. https://www.cdc.gov/tobacco/basic_information/e-cigarettes.

State of New York. *Youth Tobacco Survey: Beginning 2000*. November 21, 2019. Distributed by U.S. Department of Health & Human Services. https://healthdata.gov/dataset/youth-tobacco-survey-beginning-2000.

U.S. Department of Health and Human Services. “E-Cigarette Use Among Youth and Young Adults: A Report of the Surgeon General.” 2016. https://www.cdc.gov/tobacco/data_statistics/sgr/e-cigarettes/pdfs/2016_sgr_entire_report_508.pdf .