Emma Oriol
DATA 5332
4/11/2023

Factors Influencing Youth Substance Use and Overall Health

Abstract:

Youths navigate a complex world of family, social, and environmental influences that affect their health and decisions regarding substance use. The 2020 National Survey on Drug Use and Health (SAMHSA, 2021) collected data on 2,892 variables related to youth drug use and health, providing insight into the complex factors influencing youth. Models built using this data predict cigarette use (93% accuracy), self-reported health status (41% accuracy), and age of first alcohol use (on average within 4.3 years).

Introduction:

Youth substance use is a leading public health concern. According to the 2020 National Survey on Drug Use and Health, “8.2 percent of adolescents aged 12 to 17 drank alcohol in the past month, and 13.8 percent of adolescents aged 12 to 17 used illicit drugs in the past year” (Coady, 2022). The goal of this report is to better understand the complex content of youth drug and alcohol use and health by investigating the factors of highest influence in predicting youth cigarette use, overall health, and age of first alcohol use through predictive models. These models utilize the 2020 National Survey on Drug Use and Health (SAMHSA, 2021). The National Survey on Drug Use and Health contains data from 4,269 youth respondents in 2020 and measures 2,982 variables. These variables include measures of substance use, family life, youth involvement and academic achievement, and socioeconomic factors.

The models built use decision tree methods to create categorization and regression models. The first model predicts categorization of ‘history of cigarette use’ vs ‘no history of cigarette use’ using a decision tree model with pruning. The second model predicts classification of self-reported overall health status (“Excellent”, “Very good”, “Good”, “Poor”) using a random forest decision tree model. The third model uses a boosting decision tree method to produce a regression plot predicting age of first use of alcohol.

Theoretical Background:

The models in this report use a decision tree modeling method. Decision tree models use recursive binary splitting to repeatedly segment data. Each binary split considers only one variable, and the culmination of splits results in a terminal node that categorizes each data point. Decision tree models can be used to create both categorization and regression models and are notably useful for their high interpretability (James et al., 2022).

Several methods are available to improve decision tree models. First, pruning limits the number of variables considered at each split in order to reduce model complexity and improve interpretability. The optimal number of variables to be considered at each split can be determined by cross validation (James et al., 2022).

Second, a random forest model is an ensemble decision tree model that fits many decision tree models on random subsets of the data and averages the results. Each model uses a random subset of predictor variables in order to give more weight to a variety of variables. The number of predictors considered at each split is the square root of the total number of predictors. The default number of trees generated is 500, which is generally considered enough models without unnecessarily increasing computational costs. Random forest models are useful for reducing overfitting and decreasing model sensitivity to training data (James et al., 2022).

Lastly, boosting improves model accuracy and reduces test error by creating a series of decision trees that each learn from the previous decision tree's errors. Each tree is designed to fit to the data that was misclassified by the tree before it. This process is repeated until the error is small enough or the number of trees is met. The parameter lambda is the shrinkage/learning rate used to set the rate of change between trees. The interaction depth is used to set the number of splits in each tree, and is usually kept very low. Boosting can be prone to overfitting, so cross-validation is useful to determine the total number of trees (James et al., 2022).

Methodology:

A subset of 79 variables relating to substance use, family and home life, demographics, and education were chosen from the 2,892 variables in order to improve model simplicity and reduce computational costs. Every model was trained on a randomly selected subset of 80% of the data and tested on the remaining 20%. This split ratio was chosen to give most of the data to train a strong model but leave enough remaining data to test the models effectively.

The first model is a classification model that predicts smoking history (i.e. 'history of cigarette use' vs 'no history of cigarette use') using a pruned decision tree. Variables related to substance use, such as history of marijuana use or smokeless tobacco use, were excluded in order to investigate the factors that predict cigarette use outside of the use of other substances. The variable "smokehistory" was imputed using the "ircigfm" variable which measures number of cigarettes a month. Respondents that chose "never smoked" were categorized as "no", and all other respondents were categorized as "yes". A decision tree model was then built to predict "smokehistory" using the selected predicted variables. This model was then pruned down to consider only seven predictor variables at each split. Seven was chosen as the optimal number of variables through cross validation. The model was trained on the training data, and then the test classification accuracy rate was calculated using the test data.

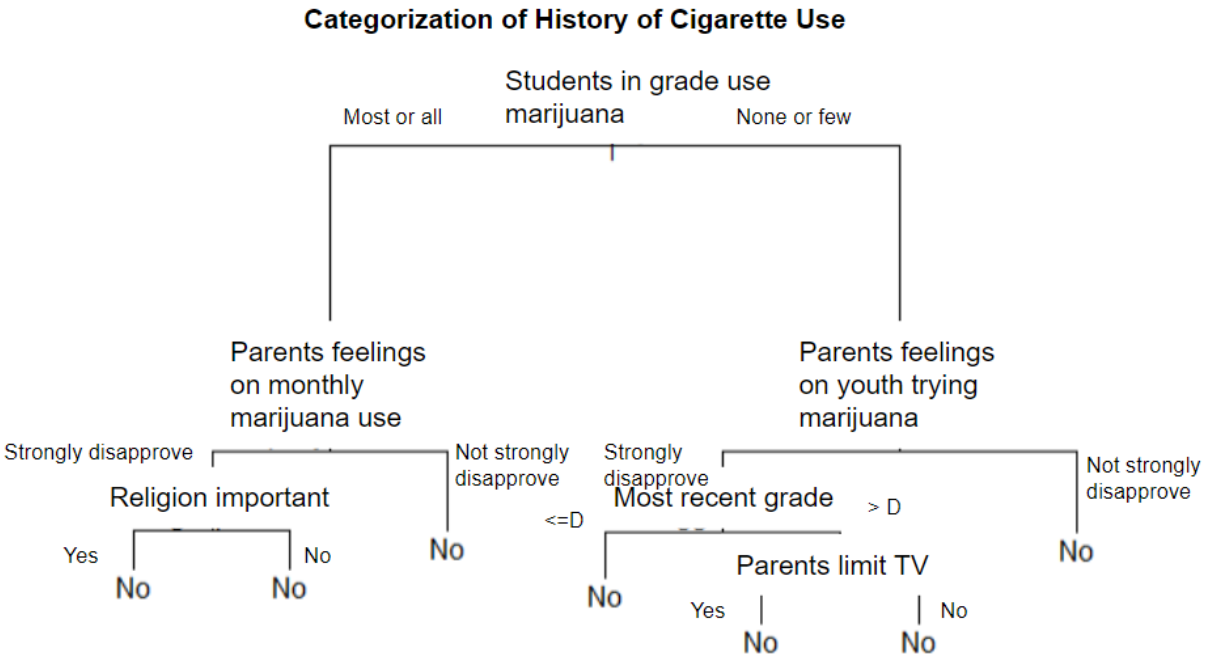
The second model is a multi classification model that predicts self-reported overall health status ("Excellent", "Very Good", "Good", or "Poor"). This is a random forest model that generated 500 decision tree models and averaged the results. The data set with 79 predictor variables was

used, but for each tree only 9 predictors were used (the square root of the total number of predictor variables rounded to the nearest whole number). The model was trained on the training data, and then the test classification accuracy rate was calculated using the test data.

The last model is a regression decision tree model that uses boosting. A total of 1,000 trees were generated, each with an interaction depth of four and a shrinkage factor of 0.01. 1,000 trees was chosen to be enough trees to get strong results without increasing computational costs further or overfitting. The interaction depth sets the size of each decision tree generated and the shrinkage factor sets the rate of learning/change between each generated tree. The values of four and 0.001 were chosen to allow the model to learn slowly and reduce overfitting. The model was trained on the training data, and then the mean squared error was calculated using the test data.

Computational Results:
The first model categorized the history of cigarette use on the test data with 93.2% accuracy (6.8% misclassification rate). However, the model only predicted “No” for all input data. The model selected the following variables as predictors; students in youth grade use marijuana, parents feel about youth use marijuana monthly, religious beliefs very important in life, youth think parents feel about youth try marijuana, grade average for last grading period completed, parents limit amount of TV in the past year.

Below is the decision tree.



The tree can be read by tracing down the nodes to the terminal node. For example, the first node of the tree splits by “students in youth grade use marijuana”. If the answer is “most or all”, the left branch is followed until the data is again split by “parents feel about youth use marijuana monthly”. If the answer is “strongly disapprove”, the left branch is again followed to “religious importance”. If the answer is “important”, then the final prediction categorization is “No history of cigarette use”. This means that a youth whose peers mostly use marijuana with parents who strongly disapprove of marijuana and religious importance in their life will be predicted to have never tried cigarettes.

The second model predicted self-reported health classification (“Excellent”, “Very Good”, “Good”, or “Poor”) on test data with an accuracy rate of 41.6% (58.5% misclassification rate).

The model selected the following variables, in order, as the most important for prediction; Parents tell youth they are proud of things they’ve done in the past year, How youth feels about peers trying marijuana, Parents tell youth they had done a good job in the past year, do youth talk with others about serious problems, youth participated in youth activities, How does youth think parents feel about youth use of marijuana monthly.

The last model predicts the first age of alcohol use with a test mean square error of 4.32, which means that on average the model predicts within 4.32 years of the actual age of first alcohol use. The top five variables with the highest relative influence are as follows; Students in youth grade drink alcoholic beverages, youth talked with parent about danger of tobacco/alcohol/drugs, how youth feels about peers trying marijuana, youth had serious fight at school/work, youth carried a handgun.

Discussion:

The first model categorizes the history of cigarette use on the test data with high accuracy (93.2%). However, the model only predicted “no history of cigarette use” for all input data, which means that the model is not very useful for predicting positive history of cigarette use. An improvement to the model could be to make it more sensitive to predicting positive history of cigarette use. The selected predictor variables indicate that factors like peer attitudes towards drugs, parents’ attitudes towards youth use of drugs, religious importance, and academic achievement are highly influential to youth’s decisions towards cigarette use.

The second model categorizes self-reported overall health status (“Excellent”, “Very good”, “Good”, or “Poor”) with relatively low accuracy (42%). This could be indicative of the complexity of factors that influence health which were not fully encapsulated by the subset of variables used in this model. Additionally, the high number of categories for health increases the difficulty and complexity of accurately predicting categorizations. The model could be improved and simplified by combining the “Excellent”, “Very good”, and “Good” categories in order to predict “Poor” or “Not poor” health.

Notably, predictor variables of high importance included several measures of parents expressing support and appreciation of youth, youth perception of how their peers and parents feel about marijuana use, youth participation in activities, and youths being able to talk about their problems. This could indicate that youths having a supportive home life is correlated with improved health.

Lastly, the third model predicts the age of first use of alcohol on average within 4.3 years of the true age. This mean squared error is high enough that the model predictions aren't particularly useful. However, it is still insightful to look at the predictor variables of highest importance to understand what factors are most correlated with age of alcohol use. Predictor variables include youth having a serious fight at school/work and carrying a handgun, which could be risk factors for alcohol use. Additionally, youths' perceptions of their peers and parents' views on marijuana and alcohol are useful in predicting age of alcohol use. This shows, again, that youth perceptions of what their parents and peers think about substance use is impactful to their decisions surrounding substance use.

The data type used to encode predictor variables can impact the type of machine learning model that is appropriate for making predictions. Binary encoding, where variables are represented as 1 or 0, is useful for indicating the presence or absence of a variable. Categorical or ordinal categorical encoding, such as with overall health ratings, is best suited for classification models, and numerical encoding is best for regression models. When constructing decision trees, models convert variables into a binary category to create a binary split in the data (i.e. age < 2 or age >=2). It's important to choose the appropriate data type for your variables based on the type of model you want to use for your predictions.

Conclusions:

In conclusion, several themes arose in evaluating model performance and variable selection. Parents and peers' attitudes towards drugs and alcohol were included as predictive variables with high importance in every model. This suggests that parent attitudes and peer attitudes towards substance use are highly influential in teen health and decisions surrounding substance use. Additional themes that showed up in several models were variables relating to peers having discussions about their problems and talking to their parents about substance. This suggests that the level of openness and trust between parents and their youths could influence youth health and substance use. More research is needed to understand the complex relationships between parents, peers, and youth and how it influences youth health and substance use. When presenting findings and pursuing further research, it's important to be cognisant and uncritical of the many parenting styles and cultural norms surrounding parenting. In some cultures, talking about substance use with youth may be considered inappropriate.

Bibliography:

Coady Jeffrey A. Coady, J. A. (2022, October 3). *Youth Substance Use Prevention Month*. SAMHSA. Retrieved April 10, 2023, from https://www.samhsa.gov/blog/youth-substance-use-prevention-month#:~:text=In%202021%2C%20more%20than%20100%2C000,died%20from%20a%20drug%20overdose.&text=According%20to%20SAMHSA's%202020%20National,drugs%20in%20the%20past%20year.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2022). *An introduction to statistical learning: With applications in R*. Springer.

Substance Abuse and Mental Health Services Administration. (2021). *Key substance use and mental health indicators in the United States: Results from the 2020 National Survey on Drug Use and Health* (HHS Publication No. PEP21-07-01-001, NSDUH Series H-56). Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration.