**Methods:** Data set name(s), dates of data collection, demographic description and size of the study population, inclusion/exclusion criteria, study design, exposures/interventions, covariates, outcomes, and statistical approach, including power calculations

**Study Population and Study Design**

The Youth Risk Behavior Surveillance System (YRBSS) includes representative samples of 9th through 12th-grade students from local school-based surveys as well as national, state, territory, tribe, and tribal government surveys. Every two years, usually during the spring semester, these surveys are carried out. The health and education departments of each state, territory, and locality undertake surveys to gather information typical of most public high school students in each jurisdiction. Based on the variables selected from the YRBSS, we were able to examine the pattern of variable importance from age, gender, school grade, race, unintentional injuries and violence, tobacco use, alcohol and other drug use, and risky sexual behavior with the outcome - STI testing other than HIV through a cross-sectional study. The study participants were collected from the Youth Risk Behavior Surveillance System (YRBSS) high school 2019 dataset. We included participants who were enrolled in high schools of selected states and local school districts aged from 12 years to 18 years or older without missing data. State, territorial, tribal government and local school district surveys with representative1 and no representative2 data are displayed. It is important to note that the YRBSS 2019 state survey does not conduct in Minnesota, Oregan, Washington, and Wyoming states. Participants were excluded from the study dataset if they had missing data from any of the survey questions.

The local high school district data were collected from 15 local high school districts provided by the YRBSS, including Broward County, FL, Chicago, IL, Eaton Consortium, MI, Fort Worth, TX, Genesee Consortium, MI, Hillsborough County, FL, Los Angeles, CA, Newark, NJ, Orange County, FL, Palm Beach County, FL, Pasco County, FL, Philadelphia, PA, Portland, OR, Shelby County, TN. **The local school district data included 2,723 participants.** The state data were collected from 12 states provided by the YRBSS state data after removing all the missing values, including Alabama, Arkansas, Illinois, Iowa, Kentucky, Michigan, Mississippi, Nebraska, Oklahoma, Pennsylvania, South Carolina, and West Virginia. **The state data included 3,797 participants.** After training the ML methods on the entire dataset, we applied the best-performed and second-best-performed ML methods to each local school district and each state to identify the pattern of variable importance. The YRBSS dataset provides de-identified data, which did not include any personal identifier and human subjects' direct contact. The Institutional Review Board (IRB) did not conduct the IRB review process. The consent process is waived.

**Measures**

All outcome and exposure variables were derived from the survey questions. The primary outcome is derived from the Youth Risk Behavior Survey question Q85:

During the past 12 months, have you been tested for a sexually transmitted disease (STD) other than HIV, such

as chlamydia or gonorrhea?

A. Yes

B. No

C. Not sure

The independent variables include demographic variables, unintentional injuries and violence factors, tobacco use, alcohol and other drug use, risky sexual behavior, and cognitive and health factors. The demographic factors include age, sex, grade, and race, which were collected from YRBS questions Q1, Q2, Q3, and Q5. BMI is calculated using Height and Weight in the following formula:

*BMI = kg/m2 = Weight (in kg)/[Height (in m)2]*

The unintentional injuries and violence factors include Birth Control Use (qnothhpl), physical fights (Q17), Dating-related forced sexual activities (Q21), Bullying on school property (Q23), and Suicidal thoughts (Q26). The tobacco use is collected from the question: Tried cigarette smoking (Q30). Alcohol and other drug use factors collected from survey questions: Alcohol consumption (Q41), Marijuana use (Q47), Prescription pain medicine misuse (Q49), Cocaine use (Q50), Heroin use (Q52), and Methamphetamine use (Q53). Risky sexual behavior factors collected from survey questions: Condom use (Q63) and Sexual contact with different genders (Q65). Cognitive and health factors collected from the survey questions: HIV testing (Q84), Asthma diagnosis (Q87), and School grades (Q89). The detailed survey questions are described in the supplemental material.

**Statistical Analysis**

**Data Cleaning and Feature Selection**

The state and local school district datasets were derived from the Combined YRBS High School Datasets by selecting related risk factors based on a clinical assessment of variables that could contribute to the outcome - STD testing. The independent variables mentioned previously as treated as categorical variables in all the models except the BMI. The BMI is treated as a continuous variable in the model. All machine learning methods implemented in this study aimed to identify the unbiased pattern between STI testing and risk factors. The variables related to drug use, including cocaine use (Q50), heroin use (Q52), and methamphetamine use (Q53), were transformed into binary variables due to their skewed to one of the categories. We used complete data for all the Machine Learning methods, and no further analysis was conducted for the missing data. To address the imbalanced data, we scaled and centered the data in the analysis process. Moreover, we assess the completeness of each category for all variables and remove categories with no values to ensure the scaling of data during fitting the ML methods.

**Model Selection, tuning, and testing**

We used multinomial logistic regression, random forest, support vector machines (SVM), Elastic Net regression, ridge regression, lasso regression, and classification tree methods in the data analysis. Our study aims to explain the variable importance and association between STI testing and demographic factors, sexual factors, unintentional injuries and violence factors, alcohol and drug use, and cognitive and health factors. As a result, we did not conduct the 70/30 data separation. However, we applied 10-fold cross-validation to minimize the overfitting and select the optimized hyperparameters for each ML method applied in the analysis. Besides the primary goal of analyzing the association, the secondary goal is to assess the performance of different ML methods on the same dataset. After the data cleaning process, the ML method was applied using the “caret” package in R version 4.2.3. And the “rpart. plot” library was used to plot the decision tree plot. The “confusionMatrix” function was used in this study to obtain the accuracy and its 95% CI for each ML method. Accuracy is the measure used to select the ML method with the best performance on the data. The comparison of accuracy across ML methods used in this study was made by the “resamples” function from the “caret” package. The “resamples” function provides the minimum, 1st quartile, median, mean, 3rd quartile, and maximum value of accuracy for each ML algorithm. The ML algorithms were trained on the whole dataset of State and local school districts. And then, we applied the best-performed ML algorithm and second best-performed to each State and each local school district dataset to explore the pattern of important variables associated with the outcome – STI testing (Q85) to ensure model accuracy and validation.

The default hyperparameter can be a good starting point for gaining a basic understanding of the ML method. However, the default parameters might not always be the best for the data, resulting in subpar performance and incorrect forecasts. Hyperparameter tuning is an important process to optimize the performance of the ML method. The type of hyperparameters is unique for different ML methods. The tuning is conducted by providing a range of values in the parameter setting “tuneGrid” of the “train” function, which can provide the best-tuned hyperparameter in the “besTune” column of the training results. Besides the accuracy and best hyperparameters, we also plot the variable importance table by “varImp” function. The more significant the variable, the more crucial it is to the model's effectiveness. **Figure. 1** show the data flow diagram. The dataset used in the ML methods training and R code can be found in the GitHub repository: <https://github.com/yh3430/CU_2023_thesis>