Data Science Final Project: Teens'n Drugs

1 Abstract

Background: Youth may experiment with or develop dependencies on illicit drugs, and on and tobacco and alcohol products; all are associated with negative health outcomes. Predicting youth who engage in substance use provides opportunity for prevention, harm reduction, or treatment interventions. **Objectives:** To develop predictive models for 1) illicit drug use and 2) alcohol and tobacco use for United States youth. **Methods:** Data from the Youth Risk Behavior Surveillance System Survey were split into training, validation, and testing sets. Three supervised learning techniques (Logistic Regression, Random Forest, Neural Network) were used to train classifiers on the training data for each objective, plus a dummy classifier. For each technique, two feature selection methods were compared: manual and data-driven, and two versions of training data used: regular and weighted whereby samples were duplicated according to survey weights. Performance was compared on the validation set using accuracy, sensitivity, specificity, and area under the curve. We selected one model for each objective to compute final performance using test data. **Results:** For both objectives logistic regression and random forest performed similarly. Performance was slightly better with data-driven feature selection; weighting the data had little impact. The neural network was inconsistent. We selected logistic regression for objective 1 and random forest for objective 2, achieving 0.897 (95% confidence interval: 0.886, 0.911) and 0.758 (0.742, 0.778) accuracy on testing data, respectively. Sensitivity was 0.213 (0.180, 0.231) for objective 1 and 0.530 (0.497, 0.566) for objective 2; Specificity was 0.988 (0.981, 0.987) for objective 1 and 0.893 (0.878, 0.895) for objective 2. **Conclusion:** We created classifiers that have decent accuracy and specificity but low sensitivity for predicting 1) illicit drug use and 2) alcohol and tobacco consumption. Feature selection and data weighting approaches had less impact on model performance than anticipated.

2 Introduction

Teenagers are susceptible to pressures that may lead to dangerous or harmful substance use. Substance use correlates with risky behaviour like vehicle accidents, higher suicide rates, homicide, high school dropout, and risky sexual behaviours [1]. Illicit drugs represent 'hard' drugs where any single episode is high risk, especially for youth who may not have knowledge of appropriate dosing and equipment hygiene. Alcohol and tobacco products have a lower single-time risk use but still may lead to addiction or other health consequences. Interventions to use or mitigate risk associated with future use may be different for the two categories. Therefore, the present study has two objectives: to develop classifiers that can be used to predict 1) high risk (illicit) drug use and 2) alcohol or tobacco consumption in youth living in the United States.

3 Methods

3.1 Data

The Center for Disease Control and Prevention has been conducting a bi-annual survey, the Youth Risk Behavior Surveillance System (YRBSS), since 1991 for youth in the United States [3]. The survey includes questions pertaining to violence and bullying; risk tendencies such as drug use and risky sexual behaviour; health behaviours such as diet and exercise; and demographics like age, and race. Most questions are categorical and ask about behaviour restricted to the last 30 days. The publicly available YRBSS data contain both the 'raw' single question answers as well as collapsed versions of some of the questions (e.g. the raw questions has 5 responses relating to frequency of a behaviour; the collapsed version is binary yes or no). We consider both of these options for our project, additionally exploring the utility of manually creating composites for questions that address a single construct.

The YRBSS is designed with a three-stage clustering approach to give a nationally representative sample of high school students (grades 9-12) in the 50 United States and the district of Columbia (not the territories). Participant responses are weighted according to non-response of schools and students within sampled schools, as well as to correct for oversampling of black and Hispanic students. Data are available for district, state, and national levels; the district and state level data are not subsets of the national level data–each has a unique sampling frame. The primary purpose of the data is to provide nationally representative descriptive statistics for youth the United States and no predictive work using the data has been found. We used national level data from 2015 and 2017.

3.2 Data Preprocessing

All data pre-processing and analyses were done in python 3.6 [8]. Missing data is due to non-response or failure of a YRBSS-conducted logical consistency test. First, questions asked in one year only (2015 or 2017) were removed. Next, we removed participants who had 50% or greater non-response across all remaining questions to eliminate surveys that were mostly incomplete ('non-responders'). While the time-frame provided for the questions should assist with recall bias, and the anonymized nature of the survey help with social desirability bias, the sensitive nature of many questions may have lead to question-specific non-responses. This type of missingness may be informative in and of itself, so we treated it as a category for categorical questions. We filled in missing continuous values with the mode (body mass index and height) and normalized from [0,1]. We removed participants with missing outcome values to keep the outcome binary and because predicting non-response is not meaningful. Questions pertaining to any drug uses were removed from the data set due to it being too closely coupled with the predictions.

3.3 Outcome Operationalization

We manually constructed outcome variables because the YRBSS is granular and has low frequency of endorsement for any given question (e.g. one question asks about cocaine use; another asks about methamphetamines. We also do not expect first-step interventions in high schools to be tailored to the subcomponents of either outcome.

Objective 1: Predicting High Risk (Illicit) Drug Use: The outcome is any illicit drug use, as defined by the YRBSS [3], within the past 30 days. A yes response to using any of cocaine (q49), inhalants (q50), heroin (q51), methamphetamines (q52), hallucinogens (q53), or ecstasy (q57) is coded as 1; otherwise, participants are assigned 0.

Objective 2: Predicting Alcohol or Tobacco Use: The second outcome is any alcohol and/or tobacco product use in the past 30 days. A yes response to using any of cigarettes (q32); electronic vapor product (q35); chewing tobacco, snuff, dip, snus, or dissolvable tobacco products (q37); cigars, cigarettes or little cigar (q38); or alcohol (q42) is coded as 1; otherwise, participants are assigned 0.

3.4 Feature Construction

Two approaches to selecting model features were used: 1) manual construction and selection, and 2) data-driven selection. A data-driven approach would be advantageous for identifying the most informative questions, but manually constructing composite predictors would improve interpretability of the model and ease of use for high school staff. Asking about a very specific behaviour, as individual YRBSS questions currently do, may be less likely to be answered accurately, or may miss a similar behaviour that a youth engages in that has the same predictive information as another question. Asking about broader categories of behaviour and experiences may overcome this.

Manual Feature Construction and Selection: We assessed raw YRBSS questions (not the pre-collapsed ones) for 1) similarity to other questions to identify a possible composite group and 2) relevance to the outcomes of interest. We selected features for any given composite by grouping questions that ask about experiences or behaviours that belong to the same broader, pertinent construct, such as bullying. The highest response across all questions was used as the final response value. For example, three questions (25, 26, 27) asked about frequency of depressive or suicidal thoughts in the past 30 days; each of these relate to a 'low mood' construct, which may be associated with drug use. The highest frequency across the three questions was recorded as the 'composite depressive or suicidal symptoms' feature. When questions within a composite had different response options, categories were selected for the composite feature such that multiple responses from any given question may belong to a single response option on the final feature. For example, questions relating to sexual assault were reduced to 1) experienced at least one kind of sexual assault within the past 30 days and 2) no sexual assault experienced within the last 30 days. This is consistent with literature showing even a single sexual assault experience can have negative impacts[4]. Other relevant questions with no suitable composite were kept in their original format. Table XX outlines all manually selected features. 1

Data-Driven Feature Selection: Our primary goal was to optimize predictive performance and not make causal inference towards the features in the model, so we also tried a purely data driven strategy for selecting features. We used a recursive feature elimination (RFE) algorithm with logistic regression as the estimator to select optimal features. The package used is in Table 6. RFE eliminates the weakest features from the model until the desired number of features is reached. Five fold cross-validation was done to reduce over-fitting. This was done repetitively for number of features in the model, until the optimal number of features was reached.

3.5 Classification Techniques

For each objective three classification techniques were compared: 1) logistic regression, 2) random forest, and 3) neural network. The first two handle categorical features well and provide interpretable output. The third was selected because while we suspect there are high dimensional relationships between variables, we were also skeptical of the utility of neural networks for our objectives and wanted to test it. Code packages for each model can be found in Table 6.

Logistic Regression: Logistic regression models the log-odds of a binary outcome as a linear combination of features. We explored different penalties and regularization strengths. The final model implemented ridge regression (l2 penalty) using sklearn's Limited-memory Broyden Fletcher Goldfarb Shanno (lbfgs) solver with c = 10.

Random Forest: Random forest is a non-parametric tree-based classification method that is essentially designed to handle categorical features. Multiple decision trees are created at training time and the final classification is based on aggregating results from the individual trees. This method mitigates single decision tree method concerns of overfitting and high variance. We explored combinations of number of trees, of features available for each split, and of maximum depth. For the first objective we decided to use 400 estimators with a maximum of 5 features shown at each split and no restriction on maximum depth. For the second objective we also used 400 estimators but with a maximum of 10 features shown at each split and a maximum depth of 10.

Neural Networks: Neural networks use the multiplication of weights and bias to build a network of nodes called neurons. We built a network with two hidden layers with 128 neurons, and 256 neurons for the first and second layer respectively that use relu activation functions. The data was fit using binary cross entropy as the loss function, 50 epochs, a batch size of 500 and a learning rate of 1e-3. Due to uneven data set classifications, the weights in the loss function were adjusted so a positive prediction was weighted 10 times more than a negative prediction. This value was chosen as it roughly estimates the uneven class size.

Dummy Classifier: The low prevalence of positive outcomes could result in deceivingly high accuracy. To help evaluate performance we created a dummy classifier that predicts the mean of the outcome; this resolved to predicting everyone as a negative case (0; no substance consumption).

3.6 Data splitting

Data was split into a training set for model fitting, a validation set to compare performance, and a testing set to estimate performance for the final selected model, as outlined below. Each observation in the data is associated with a sampling frame weight representing the probability of being sampled. These weights are designed to allow calculation of representative sample statistics for the population of interest; using weighted survey data for predictive modelling is not a well studied area. We used two approaches for the training data: 1) ignore weights (non-weighted) and 2) duplicate observations according to the weight multiplied by 100 (weighted).

Test Set: We used 4,444 participants for the test set based on a bounded loss approach with d = 0.015 and using Popoviciu's inequality to give an upper bound on the standard deviation of 0.5.

Validation Set: We wanted the validation set to be at least as large as the test set. For several candidate validation set sizes, we used the cumulative distribution function to calculate the probability of selecting the correct model. We assumed an approximately normal distribution with 0.5 standard deviation based on 0.5 as an estimation of the standard deviation of the loss, a 0.01 mean based on a 1% difference between two models. Further assumptions were common variance and independence. Although all assumptions hold so probabilities in Table 2 should be interpreted cautiously, this provides some guidance as to how confident in model selection we can be. Candidate sizes ranged from the test set size (4444) to over double (10000). We selected a validation set of 8500 samples, which gives a 0.90 probability of selecting the correct model, and is congruent with common 'rule of thumb' validation size recommendations.

3.6.1 Evaluation Metrics

For each trained model (and the dummy classifier), including variations of feature selection method and weighting of data, we calculated the accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) using predictions on the validation data. A final model for each objective was selected based on these metrics and performance was reassessed in the test set; confidence intervals and a confusion matrix is additionally reported. Especially given our classes imbalance, sensitivity and specificity interpret the model in a way accuracy cannot. Favouring sensitivity for a final model would capture more youth who truly used substances at the cost of more false positives and expending extra resources. Favouring specificity is expected to more correctly exclude youth who have not used substances at the cost of missing youth who could benefit from additional resources or intervention.

Bootstrapped 95% confidence intervals were constructed using 10000 samples with replacement of the test data. On each sample we obtained predicted probabilities to calculate the desired metrics, and then took the 0.025 and 0.975 quantile of each set (10000) metrics.

4 Results

4.1 Sample size

The 2015 and 2017 YRBSS had a combined 30,389 participants. After removing 'non-responders' and those without outcome values there were 28,702 participants for objective 1 and 28,695 for objective 2. This left 15758 (objective 1) and 15751 (objective 2) participants for training, 8500 for validation, and 4444 for testing. Overall sample demographics are in Table 3.

4.2 Impact of Feature Selection Method on Validation Performance

Classifiers using the same feature selection technique behaved similarly. Manual and RFE feature selection methods can be compared using results from the validation set provided in tables 4 and 5, and AUC plots in figures 1 and 2. In general, the RFE method performed slightly better than the manual feature selection method, more-so for objective 2 than objective 1. The neural network was an exception and did not perform better than the dummy classifier for either outcome with the manually selected features.

4.3 Impact of Weighted Data on validation Performance

Weighting the training data had little impact. This can be seen in figures 1 and 2 where the ROC curve overlaps for the classifiers of objective 1 and 2, for both manual and RFE feature selection techniques. Only the neural network showed meaningful difference on any other performance metric. For efficiency, non-weighted data is favourable.

4.4 Model Selection

We narrowed candidate models to those trained with the RFE feature selection method and non-weighted data. Even though only the neural network achieved okay levels of sensitivity, it was excluded due to lower performance on the remaining metrics (sometimes worse than the dummy classifier) and due to its questionable suitability for our objectives. There was no discernible difference between logistic regression and random forest based on visual inspection of the AUC plots for either objective. For objective 1 both had comparable accuracy (0.903 vs 0.904) and specificity (0.986 vs 0.995); we selected logistic regression based on improved sensitivity (0.204 vs 0.137). For objective 2 we selected the random forest because all performance metrics were better.

4.5 Final Model Results on Test Data

Objective 1: Our logistic regression achieved 0.897 accuracy, 0.213 sensitivity, 0.988 specificity, and 0.811 AUC (Table 7). The confusion matrix is in Table 8, AUC plot in figure 3, and regression coefficients in table 11.

Objective 2: Our random forest achieved 0.758 accuracy, 0.530 sensitivity, 0.893 specifity, and 0.815 AUC (Table 7). The confusion matrix is in Table 10, AUC plot in Figure 3, and feature importance plot to demonstrates which features contribute most to classifications in figure 4.

5 Discussion

We developed predictive models to classify 1) illicit drug use and 2) alcohol and tobacco consumption using YRBSS data. The models are expected to be applicable to United States youth.

In general, all candidate classification techniques performed similarly on validation data, except the neural network which was worse for both objectives. This is not unsurprising and neural networks are more commonly used for image data. Nonetheless, the neural network was the only model that showed a meaningful difference between RFE and manual feature selection.

The RFE feature selection performed better than manual feature selection and had more features: 83 for objective 1 and 24 for objective 2 compared to the 24 manually selected for both objectives. Poor performance of the neural network on manual features may be because it had more than 300 nodes, leading to overfitting. Another consideration is the RFE feature selection used logistic regression as the estimator, so results are biased to a logistic regression classifier. Nonetheless, we saw superior performance of the random forest for objective two. Using random forests as the estimator in RFE may have further increased it's performance for both objectives.

While model selection based on performance may prefer the RFE method, model selection based on feasibility or implementation concerns may prefer the manual feature selection method as the performance was not too far behind and the resulting classifiers require less input data, which may lead to more accurate results if implemented in high schools. Youth may only answer so many questions (response fatigue could lead to more missing inputs in a larger model) and be more willing to answer questions about broad domains of experience (i.e. manually constructed composite outcomes) rather than specific survey questions once no longer in an anonymized survey design.

Using survey weights to manipulate the training data had little impact on point estimates of performance in validation data. Further work could assess any impact of weighting, including different approaches, on certainty of estimates.

Finally, across all candidate and final classifiers, sensitivity was the lowest performance metric, indicating the potential for produce more false negatives than false positives. This is partially accounted for by the class imbalance, further demonstrated by the accuracy of the dummy classifier. Given this knowledge and resource restrictions on high schools, until more sensitive classifiers are developed it may be better to use group-level interventions that targets all students in a school (e.g. harm reduction programs, social norm altering programs), with options for self-referral to individual supports if needed [5, 6]. Group level interventions can be funded and run by outside organizations and move between schools.

5.1 Limitations

Limitations of our work include cross sectional data from 2015 and 2017. Period or cohort effects could have impacted results. Our classifiers predicted current or recent substance consumption with the assumption this would be associated with continued or future consumption patterns that may benefit from intervention. Other limitations are using weighted survey data collected for descriptive purposes, no confidence intervals or uncertainty estimates to aid model selection in the validation data, and no external validation.

References

[1] Shahid Ali, Charles P. Mouton, Shagufta Jabeen, Ejike Kingsley Ofoemezie, Rhan K. Bailey, Madiha Shahid, and Qiang Zeng. Early detection of illicit drug use in teenagers, 2011.

[2] François Chollet et al. Keras. https://keras.io, 2015.

[3] Centers for Disease Control and Prevention. Youth risk behavior survey data. www.cdc.gov/yrbs, 2015-2017.

[4] An Tong Gong, Sunjeev K. Kamboj, and Helen Valerie Curran. Post-traumatic Stress Disorder in Victims of Sexual Assault With Pre-assault Substance Consumption: A Systematic Review. *Frontiers in Psychiatry*, 10, 2019.

[5] G.Alan Marlatt and Katie Witkiewitz. Harm reduction approaches to alcohol use: Health promotion, prevention, and treatment. *Addictive Behaviors*, 27(6):867 – 886, 2002. Integrating Substance Abuse Treatment and Prevention in the Community.

[6] Simone A. Onrust, Roy Otten, Jeroen Lammers, and Filip Smit. School-based programmes to reduce and prevent substance use in different age groups: What works for whom? systematic review and meta-regression analysis. *Clinical Psychology Review*, 44:45 – 59, 2016.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[8] Python Core Team . *Python: A dynamic, open source programming language*. Python Software Foundation, 2015.

6 Appendix

6.1 Feature Selection

Feature Name	Rationale	Questions Used	Final Characteristics
Age	Exposure increases with age.	Demographic age	Raw question format
Sex	Associated with substance use patterns.	Demographic sex	Raw question format
Body Mass Index	Substance use impacts appetite and weight.	Demographic bmi	Raw question format (continuous)
Race	Associated with substance use patterns.	Demographic race7	Raw question format
Sexual Minority Status	Associated with substance use patterns.	67	Raw question format

Safety	Several questions addressed safety practices including those related to risky behaviour	8,9,10,11	Highest frequency response taken across all questions: never, rarely/non-habitual levels, frequent, always/a lot
Perceived Threat	Several questions pertain to how safe a youth feels, which may relate to their home or school environment. Although not asked directly, this may relate to availability or normative behaviour around substances.	15, 16, 17	Composite using highest frequency of response across questions: no threat, lowest level threat, higher threat, very frequent threat perceived
Sexual Assault	Multiple questions asked about sexual assault within the past 30 days. While each question asked about something with varying levels of intrusiveness or frequency, one event for any question can be considered a traumatic experience and could lead to substance use as a coping mechanism.	19,21,22	Binary composite to indicate at least once or never across all subquestions
Bullying	Two questions asked about bullying, which may relate to peer pressure or coping mechanisms associated with substance use	23, 24	Binary composite indicating whether or not the youth experiences bullying
Depressive and Suicidal Symptoms	Low mood is associated with substance use; while direction of causality is non-fixed, knowing low mood can be an indicator of use	25, 26, 27	Composite using highest frequency of response across questions: None, mild or infrequent, severe or frequent symptoms
Physical Activity	Physical activity is a protector and replacement activity for substance use.	79, 82	Composite using highest frequency of response across questions: ranging from 0 to most/all days of the past week
Weapon carrying	Associated with perceptions of danger and risk-taking tendencies	q12	Raw question format
Sexual contact	Associated with youth substance use.	q66	Raw question format
Number of sexual partners	Associated with risk behaviour and youth substance use	q62	Raw question format

Perceived Weight	Substances are associated with reduced appetite and weight; youth may turn to for appetite suppression.	q68	Raw question format
Intention to Lose Weight	Youth may turn to tobacco as appetite suppressant	q69	Raw question format
Fruit Consumption	Indicator of overall healthy behaviours and sociodemographic status	qnfr2	Raw question format
Vegetable Consumption	Indicator of overall healthy behaviours and sociodemographic status	qnveg2	Raw question format
TV Watching Amount	Indicator of overall activity level and may be an activity engaged in while using substances	q80	Raw question format
Video Game Amount	Indicator of overall activity level and may be an activity engaged in while using substances. Not grouped with TV due to less similarity and less of a sensitive question/less masking/privacy expected to be needed for accurate responses	q81	Raw question format
Dentist Check-Ups	General health and socioeconomic status indicator, which are associated with substance consumption	q86	Raw question format
Asthma Diagnosis	Associated with smoke inhalation	q87	Raw question format
Sleep Amount	Substance use often occurs later in the day/night and can impact sleep	q88	Raw question format
School Grades	Associated with focus and attendance, which are inversely associated with substance use	q89	Raw question format

Table 1: Manual Feature Selection Details

Note: No-response categories were added for all categorical variables, as described in our missingness strategy in the Methods section. Raw question format refers to the values in the original YRBSS data, which are categorical unless otherwise stated, and available through their website [3].

6.2 Validation Set Size Calculations

n	Probability of Selecting the Best Model
4444	0.827
5000	0.841
8000	0.897
8500	0.904
9000	0.910
10000	0.921

Table 2: Validation set size calculations

6.3 Demographic Characteristics

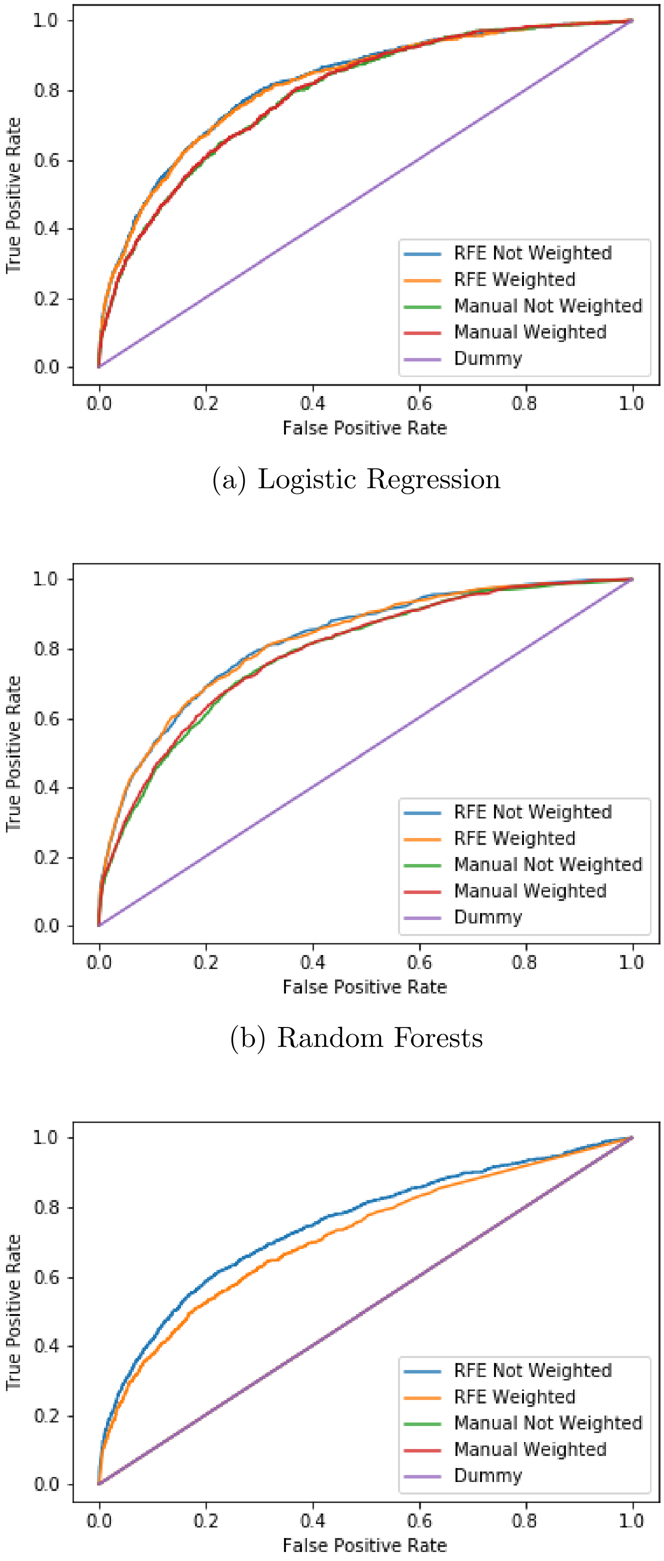
Demographic Information	Subcategory	Summary
Age	12 years or younger	42 (0.4%)
	13 years old	28 (0.09%)
	14 years old	3311 (11.5%)
	15 years old	6960)(24.2%)
	16 years old	7359 (25.6%)
	17 years old	7105 (24.7%)
	18 years or older	3777 (13.2%)
	Missing	127 (0.4%
Grade	9th	7367 (25.7%)
	10th	7230 (25.2%)
	11th	7190 (25.0%)
	12th	6693 (23.3%)
	Missing	229 (0.8%)
Sex	Female	14645 (51.0%)
	Male	13872 (48.3%)
	Missing	192 (0.7%)
Race	American Indian/Alaska Native	278 (1.0%)
	Asian	1232 (4.3%)
	Black or African American	3973 (13.8%)
	Hispanic/Latino	8343 (29.1%)
	Native Hawaiian/Other Pacific Islander	192 (0.7%)
	White	12670 (44.1%)
	Multiple Races (Non - Hispanic)	1471 (5.1%)
	Missing	550 (1.9%)

Table 3: Summary of Sample Demographics: 2015 and 2016 YRBSS after removing 'nonresponders'

6.4 Validation Data Performance Metrics

Model Type	Weighted	Feature Selection	Accuracy	Sensitivity	Specificity	AUC
Dummy Classifier	NA	NA	0.894	0	1	0.5
Logistic Regression	No	Data-Driven	0.903	0.204	0.986	0.819
		Manual	0.897	0.133	0.987	0.790
	Yes	Data-Driven	0.903	0.198	0.987	0.816
		Manual	0.896	0.127	0.988	0.789
Random Forest	No	Data-Driven	0.904	0.137	0.995	0.825
		Manual	0.901	0.119	0.994	0.792
	Yes	Data-Driven	0.903	0.133	0.994	0.824
		Manual	0.900	0.123	0.992	0.788
Neural Network	No	Data-Driven	0.862	0.326	0.926	0.715
		Manual	0.117	1	0	0.5
	Yes	Data-Driven	0.802	0.501	0.858	0.748
		Manual	0.117	1	0	0.5

Table 4: Validation Performance For Objective 1 (High-Risk/Illicit drug use)



(a) Logistic Regression

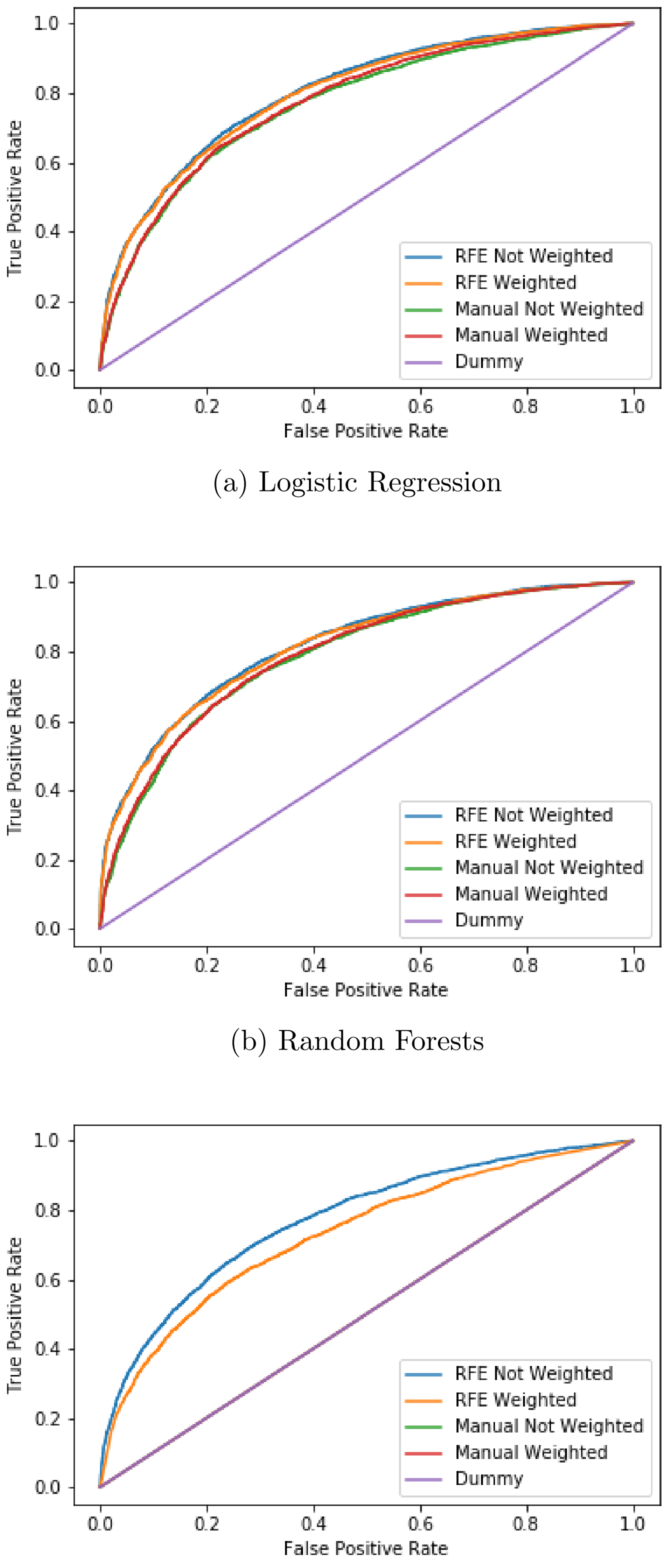
(b) Random Forests

(c) Neural Network

Figure 1: AUC plots of the objective 1 (High-Risk/Illicit drug use)

Model Type	Weighted	Feature Selection	Accuracy	Sensitivity	Specificity	AUC
Dummy Classifier	NA	NA	0.625	0.000	1.0	0.5
Logistic Regression	No	Data-Driven	0.743	0.520	0.877	0.804
		Manual	0.728	0.493	0.869	0.777
	Yes	Data-Driven	0.746	0.523	0.879	0.799
		Manual	0.725	0.481	0.871	0.770
Random Forest	No	Data-Driven	0.758	0.533	0.892	0.818
		Manual	0.739	0.547	0.853	0.792
	Yes	Data-Driven	0.754	0.534	0.887	0.813
		Manual	0.738	0.537	0.858	0.788
Neural Network	No	Data-Driven	0.693	0.601	0.728	0.732
		Manual	0.371	0	0	0.5
	Yes	Data-Driven	0.639	0.839	0.520	0.773
		Manual	0.371	1	0	0.5

Table 5: Validation Performance for Objective 2 (Alcohol and Tobacco use)



(a) Logistic Regression

(b) Random Forests

(c) Neural Network

Figure 2: AUC plots of the objective 2 (Alcohol and Tobacco use)

6.5 Testing Data Performance Metrics

Model Type	Package Used
Feature Selection	sklearn.feature_selection.RFECV [7]
Logistic Regression	sklearn.linear_model.LogisticRegression [7]
Random Forest	sklearn.ensemble.RandomForestClassifier [7]
Neural Network	Keras [2]
Dummy Classifier	sklearn.dummy.DummyClassifier [7]

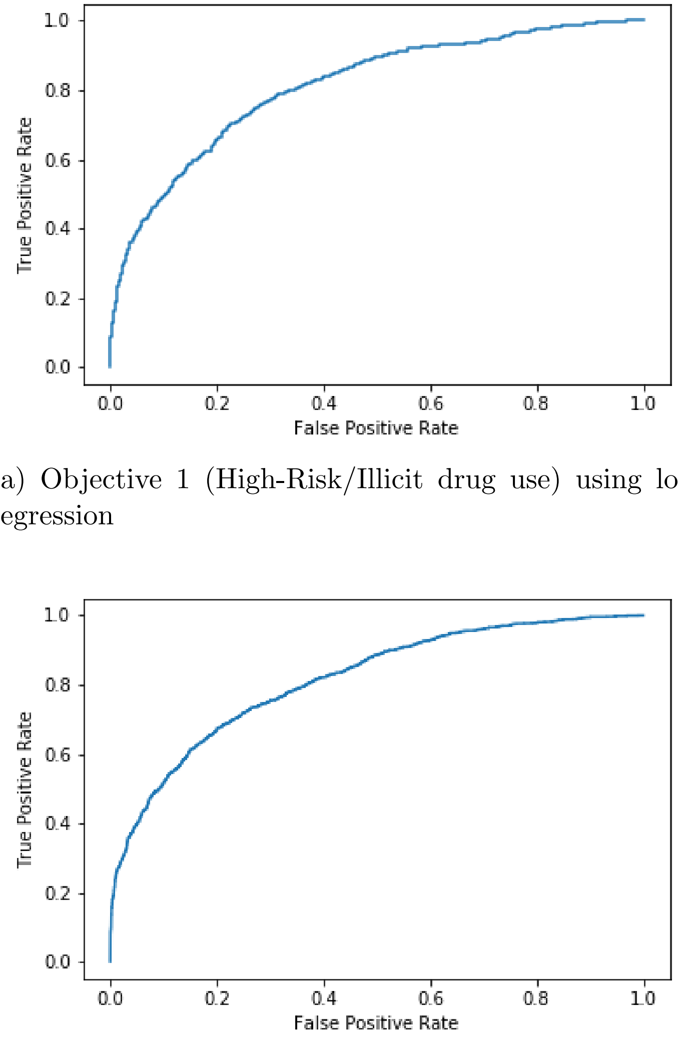
Table 6: Key Packages Used

Model Type	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	AUC (95% CI)
Objective 1	0.897 (0.886, 0.911)	0.213 (0.180, 0.231)	0.988 (0.981, 0.987)	0.811 (0.584, 0.635)
Objective 2	0.758 (0.742, 0.778)	0.530 (0.497, 0.566)	0.893 (0.878, 0.895)	0.815 (0.795, 0.814)

Table 7: Testing Performance

	Predicted Positive	Predicted Negative
Observed Positive	110	409
Observed Negative	49	3876

Table 8: Testing Performance Objective 1 (High-Risk/Illicit drug use) Confusion Matrix



(a) Objective 1 (High-Risk/Illicit drug use) using logistic regression

(b) Objective 2 (Alcohol/ Tobacco use) using random forests

Figure 3: AUC of the final models on the testing data

Question	**Coefficient**
age	-0.12996133
sex	0.14007833
grade	0.11656388
race4	0.08151893
race7	0.0656219
bmi	-0.93292594
q67	-0.01044678
q66	0.16914702
sexid	-0.01044678
sexid2	0.3309728
sexpart	-0.16576325
sexpart2	0.26947381
q8	-0.10515266

q9	0.18047593
q10	0.13520829
q12	0.14787795
q15	0.09327482
q16	0.11246571
q17	0.13689193
q18	0.06236822
q19	-0.03475479
q21	-0.04417338
q22	0.0627582
q23	-0.07786692
q24	-0.00945394
q25	-0.29589541
q26	-0.20502565
q27	0.01084526
q28	0.18809358
q58	-0.56045897
q59	-0.56977652
q60	-0.02605732
q61	0.20298169
q62	-0.00485803
q63	-0.28480446
q64	0.15468964
q68	-0.04774481
q69	-0.06963265
q72	0.05324012
q74	0.03449686
q76	0.0505714
q78	-0.05735116
q79	-0.01547327
q81	0.02161934
q85	-0.07199273
q88	0.00584107
q89	0.01228097
qwater	0.031487
qfoodallergy	-0.02878704
qindoortanning	0.14119372
qsunburn	0.05462804
qconcentrating	-0.24023756
qspeakenglish	0.24237949

Table 9: Coefficients and intercept from logistic regression model for objective 1 High-Risk/Illicit drug use) on testing data

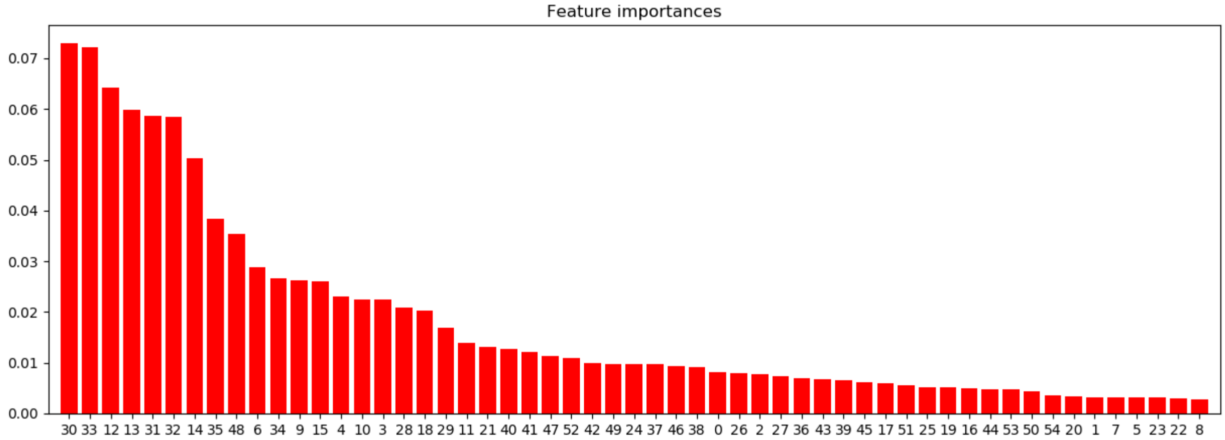


Figure 4: Objective 2 (alcohol and tobacco use) Random Forest: Feature Importance Plot

	Predicted Positive	Predicted Negative
Observed Positive	875	775
Observed Negative	299	2495

Table 10: Testing Performance Objective 2 (alcohol and tabacco use): Confusion Matrix

Feature	**Ranking**
33	0.080777
30	0.068744
32	0.066109
12	0.064934
13	0.060475
31	0.055210
14	0.047576
48	0.033059
34	0.032605
35	0.031119
15	0.026803
6	0.025586
10	0.023756
3	0.022602
4	0.022191
18	0.021214
29	0.021123
9	0.021014
28	0.020933
11	0.013276
40	0.012408

21	0.012345
41	0.011715
52	0.011271
47	0.011097
42	0.010029
37	0.009638
46	0.009319
49	0.009270
24	0.009147
38	0.009131
26	0.008381
0	0.008160
2	0.007572
27	0.007417
36	0.007024
43	0.006919
39	0.006637
17	0.006258
45	0.006147
51	0.005891
19	0.005407
25	0.005037
16	0.004910
53	0.004790
44	0.004772
50	0.004581
20	0.003799
54	0.003521
5	0.003259
1	0.003148
23	0.003093
7	0.003037
22	0.002948
8	0.002816

Table 11: Feature rankings from random forest classifier on objective 2 (alcohol and tobacco use) on testing data. Corresponds to Figure 4.