2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

A Psychosocial Approach to Predicting Substance Use Disorder (SUD) Among Adolescents

[1]Adway S. Wadekar
Saint John's High School
Shrewsbury, Massachusetts, USA
adway@adway.io

***Abstract*—Substance Use Disorder (SUD) affects about 5% of adolescents and can lead to many personal and societal problems. Risk factors such as peer pressure, permissive parenting, and impulsiveness make SUD more probable, whereas protective factors like community engagement alleviate this risk. No factor, however, is a sole determinant of SUD. The objective of this research is to build an ensemble learning framework to comprehensively predict adolescents at risk for SUD, considering the interplay between different factors. A data-driven model of 34 factors reflecting multiple dimensions of an adolescent's sphere of life, many of which comprise an adolescent's social network, is built from over 100 questions in the National Survey on Drug Use and Health. These factors are split into two groups; Proximal includes characteristics that are individual-centric, whereas Distal includes environmental influences. A labeled data set is curated by pooling the observations from the 2016 and 2017 editions of the survey. Two ensemble classifiers are trained based on the labeled data set, while applying the SMOTE algorithm to consider class imbalance. Both classifiers can distinguish between adolescents with and without SUDs exceptionally accurately, with Area Under the ROC curve over 0.90, outperforming multivariate logistic regression, a commonly used model in public health studies. Obesity combined with being approached with drugs poses the highest risk from over 1000 interactions. It is possible that the legalization of marijuana may exacerbate this problem. Based on these findings, we may infer that SUD among adolescents may not be exclusively attributed to natural tendencies or environmental influences but arises from their confluence.**

***Index Terms*—Substance Use Disorder, Adolescence, Machine Learning, Sociodemographic Factors, Psychosocial Factors**

I. Introduction

Substance Use Disorder (SUD) among adolescents can lead to a variety of issues including lower school performance, poor peer relationships, and motor-vehicle accidents, although it affects only approximately 5% of adolescents in the U.S. Predicting adolescents at risk for SUD can guide the design of intervention strategies and may be vital in alleviating many such public health concerns. Prior research suggests that several risk and protective factors including peer pressure, seeking peer approval towards the use of substances, perceived risk of harm, involvement in youth activities, participating in drug education programs, lack of impulse control and risk taking behaviors influence the occurrence of SUD [1].

Many studies estimate the likelihood of adolescents developing dependence on or abuse to one or more substances based on only certain portions or cross sections of their multi-dimensional lives by considering a few factors at a time. For example, one study explores the link between parenting style and drug abuse [1], another one links obesity to illegal substance use and misuse [2], [3], and a third one shows the connection of community attributes including disorganization and poverty to substance abuse [4]. To the best of our knowledge, however, no research has cohesively considered the entire sphere of an adolescent's life in a single, integrated model. Unless the factors are considered together, it is impossible to understand how they influence each other in exacerbating the risk of SUD. With such integrated consideration, we can move beyond explaining the contribution of risk and protective factors to the development of SUD to predicting its likelihood.

This research develops a psychosocial approach to predict adolescents at risk for SUD by considering a combination of their intrinsic physical and psychological traits, properties of their school and family environments, and attributes of their neighborhoods and communities. Central to the approach is a data-driven model, comprising of 34 factors that represent multiple dimensions of an adolescent's life, built using the responses from the National Survey on Drug Use and Health (NSDUH) [5], [6]. Many of these factors are a product of the interactions adolescents have with other people – their social network. A labeled data set is curated by pooling the responses from the 2016 and 2017 editions of the NSDUH, where the outcome is abuse of or dependence on many illicit substances and alcohol. Two ensemble classifiers (gradient boosting and random forest) are trained and tested using the labeled data set while accounting for class imbalance using the SMOTE algorithm. The performance of the classifiers is evaluated using sensitivity, specificity, and AUC; and is compared to the multivariate logistic regression model. Finally, to gain insights into what drives adolescents towards SUD, importance scores of features and their interactions are studied.

The rest of the paper is organized as follows. Section II presents the psychosocial approach. Section III summarizes and discusses the results. Section IV concludes the paper and offers directions for future research.

II. Psychosocial Approach

In this section, we present the steps in our pyschosocial approach to SUD prediction.

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

A. Building a Data Model

In this approach, whether or not an adolescent may develop SUD is viewed as an outcome that is made probable by certain risk factors, and this risk may be alleviated by certain protective factors. A data-driven model of the risk and protective factors encompassing adolescents' sphere of life was built using responses to over 100 questions from the National Survey on Drug Use and Health (NSDUH). The NSDUH is an annual survey of the civilian, non-institutionalized population of the United States aged 12 years or older. It is a large, nationally representative sample of the United States population.

"Simple" factors in the approach have a one-to-one mapping to a single question in the survey. "Compound" factors are composed from a many-to-one mapping of multiple related questions to a single factor. For a compound factor, an overall score is computed by adding the coded responses to those questions that compose the factor. For example, the compound factor "religious beliefs" is composed from the coded responses to the following questions:

1) 1. Number of religious services attended in the past 12 months. (25 or more = 1, less than 25 = 2);
2) Religious beliefs are very important. (Agree/Strongly Agree = 1, Disagree/Strongly Disagree = 2);
3) Religious beliefs influence my decisions. (Agree/Strongly Agree = 1, Disagree/Strongly Disagree = 2);
4) Important that friends share my beliefs. (Agree/Strongly Agree = 1, Disagree/Strongly Disagree = 2)

The sum of the coded responses to these four questions range from 4 to 8. They indicate a spectrum of religiosity ranging from strong to mild.

The literature does not suggest a consensus organization of these risk and protective factors into categories. One study classifies them into three groups, familial, social, and individual, whereas another classifies them into individual, family, peer, and school. Absent any standard classification scheme, we split these 34 factors into two groups: *Proximal (P)* and *Distal (D)*. The proximal group includes immutable factors such as race and gender. It also consists of those factors that are mostly owned by an adolescent, and are largely under their influence such as obesity, their attitude towards the use of substances by self and peers, religious beliefs, health, and personality traits. By contrast, the distal group consists of factors that are predominantly determined by the environment in which an adolescent is nurtured. These factors will undoubtedly impact the adolescent, and include family characteristics (economic status, family size and composition), the use of substances by peers, opportunities to participate in drug education and drug/violence prevention, self-esteem/confidence building programs either through the school or community, attitudes of parents about experimenting with different drugs occasionally or regularly, the degree to which parents are involved both by offering encouragement and support as well as setting limits and boundaries in a gentle yet firm manner, the extent to which adolescents participate in activities organized by the school, institutions of faith and community, characteristics of the community and neighborhood including state laws surrounding the legalization of marijuana, and whether they live in a large metro, small metro, or a non-metro environment.

It is noted that some factors do not exclusively belong to either proximal or distal groups, rather, they may belong to both. For example, religiosity may be a product of both inherent beliefs of an adolescent and the values and practices of their family. However, each factor is exclusively classified into proximal or distal groups based on whether it predominantly defines an adolescent's innate attribute or characterizes the environment in which the adolescent is being raised. The mapping of risk and protective factors to NSDUH questions is summarized in Tables I and II respectively. Overall, the model includes 12 simple and 22 compound factors. The proximal group consists of 4 simple and 11 compound factors, whereas the distal group consists of 8 simple and 11 compound factors.

A broad, two-dimensional view of SUD is taken. It considers a range of illicit substances, namely, marijuana, hallucinogens, inhalants, methamphetamine, cocaine, heroin, prescription pain relievers, prescription sedatives, prescription stimulants, prescription tranquilizers, and alcohol. Substance abuse is characterized by a pattern of substance use leading to neglect of roles or commitments, physical hazards, legal issues, or interpersonal problems. Substance dependence is a more extreme diagnosis wherein use of the substance is reoccurring to the point of reducing important social and occupational activities, causing tolerance of the substance and/or withdrawal symptoms and causing the user to devote a great deal of time to obtain and use the substance. Either abuse or dependence of at least one of these substances in the past year is labeled as SUD. This is consistent with the fourth edition of the Diagnostic and Statistical Manual of Mental Disorders [7], which has shown good accuracy in psychiatric research. Our analysis of SUD is limited to illicit drugs as the NSDUH does. This approach likely prioritizes substances that are more prevalently used, especially marijuana and alcohol [8].

B. Curating a Labeled Data Set

A labeled data set based on the data-driven model was curated, where each observation representing an adolescent consisted of an outcome to be predicted, that is, SUD vs. No-SUD. It also includes the 34 protective and risk factors defined in Section II-A that are germane to prediction. Initially, only the 2016 edition of the NSDUH survey was considered, which consists of about 56,000 responses. Of these, about 13,000 were from adolescents (age range 12-17). Many of the NSDUH variables used in the model are in their raw form, and hence, for many observations, these variables have missing data. Eliminating the observations with missing values decreased the size of the sample to 9,128 observations. Therefore, data from both the 2016 and 2017 editions of the NSDUH survey were pooled to increase the size of the sample. The NSDUH survey design is stable and consistent [8], [9],

TABLE I
PROXIMAL FACTORS & MAPPING TO NSDUH VARIABLES

ID.	Feature	Mapping/Description	# Var.
1.	Risk of Harm – Daily Use	Perceives people risk harming by binge drinking and smoking.	2
2.	Risk of Harm – Weekly	Whether an adolescent perceives that people risk harming themselves by binge drinking, or using cocaine, marijuana, LSD, Heroin.	2
3.	Risk of Harm – Monthly	Whether an adolescent perceives that people risk harming themselves by using cocaine, marijuana.	2
4.	Risk of Harm – Lifetime	Whether an adolescent perceives that people risk harming themselves by trying LSD and Heroin.	2
5.	Mental Illness, Depression	Feeling sad, empty, depressed, discouraged, or hopeless, lost interest and became bored with most things usually enjoyable.	9
6.	Special Drugs	At least once use of cough/cold meds, GSB.	2
7.	Health	Self-assessed overall health	1
8.	Obesity	Based on BMI (Obese, Overweight, Normal)	1
9.	Race	Seven levels (Non Hispanic White, Non Hispanic Black/African American, Non Hispanic Native American/Alaska Native, Non Hispanic Native Hawaiian/Pacific Islander, Non Hispanic Asian, Non Hispanic more than one race, Hispanic).	1
10.	Gender	Male/female	1
11.	Gang Affiliation, Violence	Whether an adolescent engaged in a serious fight at school/work, or fought with a group versus other group, carried a handgun, sold illegal drugs, stole/tried to steal item greater than $50, Attacked with intent to seriously harm anyone.	6
12.	Peer Approval	What an adolescent thinks close friends feel about them smoking 1 pack of cigarettes per day, trying marijuana, use marijuana monthly, drinking 1-2 alcoholic beverages/day.	4
13.	Religious Beliefs	Number of times attended religious services, whether religious beliefs are important, whether religious beliefs influence life decisions, whether it is important for friends to share the religious beliefs.	4
14.	Risk taking behavior	Whether the adolescent gets a real kick out of doing dangerous things, whether the adolescent likes to test limits by doing risky things, whether an adolescent wears a seatbelt when ride in front passenger seat.	3
15.	Youth Approval	What youth feels about peers trying marijuana, using it monthly, smoking more than 1 pack of cigarettes per day, or drinking 1-2 alcoholic drinks/day.	4

allowing for data from multiple years to be combined. The pooled data consisted of a total of 18,624 observations.

C. Selecting Machine Learning Models

Machine learning models offer the potential to account for complex interactions among factors that may influence an outcome, a recent review recognizes the promise of machine learning in substance abuse research [10].

Many machine learning models were considered. Among these, Support Vector Machines are not suitable, because most factors are categorical. Naive Bayes classifiers are useful only when each factor contributes independently to the prediction, but those in our model are likely to have associations. Artificial Neural Networks (ANNs) are suited for unstructured data such as text and images. For structured lightweight survey data ANNs can be too robust and therefore unnecessary [11].

Considering the characteristics of the problem (categorical factors with associations), decision trees seemed to be the natural choice. Decision trees are common because they are simple and provide a clear, visual guide to the decision-making process. Decision trees, however, may overfit, that is, they may try to explain the training data well instead of finding patterns that generalize [12]. Overfitting may lead to low accuracy when predicting outcomes using test data. Moreover, they are very susceptible to noise and small changes in the training data may cause large changes in the predictions.

These problems can be mitigated by building an "ensemble" of trees and then aggregating the predictions from this collection of trees into a final, singular prediction. Two ensembles that utilize decision trees to aggregate, namely, the random forest and gradient boosting methods, are common. These methods have two key differences. How and when the trees are built is the first difference. Random forest builds each tree in parallel, while gradient boosting builds each tree sequentially correcting upon the misclassifications from the prior trees. Another difference lies in when the predictions from the trees are aggregated. Random forest combines predictions at the end by a majority rule approach, whereas gradient boosting combines predictions along the way. Random forests and gradient boosting each excel in different areas. Random forests perform well for multi-class object detection and bioinformatics, which tends to have a lot of statistical noise. Gradient Boosting performs well when the data is unbalanced such as in real-time risk assessment. The problem of adolescent SUD is a bioinformatics application, but with class imbalance. As a result, it is difficult to estimate a priori, whether random forests will outperform gradient boosting or the other way round.

The multivariate logistic regression model that is commonly used in medical studies, was also included as a baseline for comparison [13]. The main limitation of the logistic regression model is that it can only indicate the significance of the risk

TABLE II
DISTAL FACTORS & MAPPING TO NSDUH VARIABLES

ID	Feature	Mapping/Description	# Var
1.	School, academics	Felt about going to school, whether the teacher let them know that they were doing a good job, grade average.	3
2.	Peer Drug Use	Associated with peers in grade who use cigarettes, marijuana, alcohol, or got drunk at least once/week	4
3.	Easy availability	Easy to obtain cocaine, crack, heroin, LSD, marijuana?	4
4.	Approach	Approached by someone selling drugs	1
5.	Parenting style involvement	Whether the parents checked and helped with homework, made them do chores, limit amount of TV and time out on school night, told them that they are proud, and talked about dangers of drugs, alcohol, and tobacco. Number of times argued and fought with parent.	8
6.	Parental attitudes towards substances	What the adolescent thinks parents feel about smoking one pack of cigarettes/day, trying marijuana, using marijuana monthly, drinking 1-2 alcoholic beverages/day.	4
7.	Mother	Presence of mother in the household	1
8.	Father	Presence of father in the household.	1
9.	Poverty	Poverty level of the family (poverty, income two or more times that of federal poverty threshold).	1
10.	County	Living in large metro, small metro, or a non-metro area.	1
11.	Insurance	Covered by private insurance, Medicare, Medicaid, Champus, VA or military, or any other health insurance.	5
12.	Family size	Number of people in household (1 through 6, or more).	1
13.	Govt Programs	Family participates in any government programs such as the Supplemental Social Security Income, Food Stamps, Cash/Non Cash Assistance.	4
14.	Youth Activities	Number of youth activities participated in school, community, church/faith-based, or other activities.	4
15.	Drug Education	Participated in any drug education in school either through a special class, or through films/lectures/discussions in or out of class such as a special assembly.	3
16.	Drug Prevention Message	Has heard any drug prevention message outside of school – through posters, pamphlets, billboards or Radio and TV.	1
17.	Drug/Self-help Programs	Participated in any groups that promote problem solving, communication skills, and self-esteem, or programs that prevent violence and drug use, or help substance abuse, counsel against pregnancy and STDs	5
18.	Support System	Who the adolescent talks to about serious problems parent, guardian, boyfriend/girlfriend, other adults, or no one.	5
19.	State Law about Marijuana	Whether the state in which the adolescent resides has legally approved marijuana.	1

and protective factors. The p-value obtained from this model has a binary interpretation, and it cannot be used to determine how significant the contribution of a factor is to the prediction. Therefore, logistic regression is not useful in determining the relative importance of various factors. Ensemble models, especially random forests, can rank factors and interactions according to their importance to prediction.

D. Training and Testing Models

In machine learning, the standard process is to split a data set with known outcomes into training and test sets, which respectively include anywhere from 60% to 80% and 40% to 20% observations [14]. The training set is used to train the classifier, and the test set is used to test its performance. Class imbalance prevents the use of this standard process in predicting SUD in adolescents. Class imbalance occurs because only 5% of the adolescents have SUD, and this class constitutes less than 10% of the total number of observations. If not handled, class imbalance will make the classification problem useless because a model could predict that all of the adolescents are unlikely to develop SUD, and it would be "correct" 95% of the time.

Various techniques can address class imbalance with the ultimate goal of producing a balanced training data set. Down-sampling randomly removes instances in the SUD class, whereas in up-sampling, instances are randomly replicated in the No-SUD class. We use the Synthetic Minority Over-sampling Technique (SMOTE), which combines up-sampling and down-sampling [15]. Furthermore, instances from the No-SUD class are down-sampled leading to a balanced data set. This prevents a massive loss of valuable data yet does not rely on a simple resampling. We implemented the SMOTE algorithm in the following steps:

- The data set was split into two partitions using stratified sampling, such that partitions #1 and #2 include 80% and 20% of the observations respectively. Stratified sampling preserves the original ratio of SUD to No-SUD classes

in both partitions.

- Partition #1 was SMOTE-ed, SUD observations are up-sampled, and No-SUD observations are down-sampled. The balanced training data set contains 1096 observations each from the SUD and No-SUD classes.
- Partition #2 is used to test and evaluate classifier performance. The test data set is left imbalanced to simulate a real-life condition in which the prevalence of SUD is only 5%.

E. Defining Performance Metrics

A model predicts whether an adolescent has SUD or not based on the risk and protective factors. This predicted outcome is compared against the true or the actual outcome by classifying into the following four groups, organized in the form of a confusion matrix.

- True Positive (TP): The model predicts SUD, when it is present.
- False Positive (FP): The model predicts SUD, when it is absent.
- True Negative (TN): The model predicts no SUD, when it is absent.
- False Negative (FN): The model predicts no SUD, when it is present.

Based on these four groups, we define sensitivity and specificity to evaluate the performance of each classifier. Sensitivity is the ability of a model to correctly identify those with SUD (true positive rate), while specificity is its ability to correctly identify those without SUD (true negative rate). If a highly sensitive model predicts an adolescent as having SUD, then we can be fairly certain that the adolescent does, in fact, have SUD. If a highly specific model says that an adolescent does not have SUD, then we can be fairly certain that the adolescent doesn't have SUD. Expressions for sensitivity and specificity are as follows:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

Generally, there is a tradeoff between sensitivity and specificity. A model with high sensitivity usually has low specificity, that is, it may find those with SUD, but it may also falsely identify those without SUD as having one. The ROC (Receiver Operating Characteristics) curve was used to assess this tradeoff [16]. ROC measures how well the model can distinguish between adolescents with or without SUD. AUC approaching 1.0 is desirable because it means that a model shows perfect discrimination. For psychological applications, such as in this research, the performance is considered fair for AUC between 0.70-0.79, good in the range 0.80-0.89, and exceptional if greater than or equal to 0.90 [17].

The split-downsample-train-test process was repeated 50 times, and the ensemble and logistic regression models were trained and tested on each combination of training and test sets. All analysis was completed using the packages in R [18].

TABLE III
PREDICTION PERFORMANCE

Model	Sensitivity	Specificity	AUC
DT	0.79 ($\pm$ 0.013)	0.79 ($\pm$ 0.008)	0.84 ($\pm$ 0.006)
RF	0.81 ($\pm$ 0.010)	0.83 ($\pm$ 0.003)	0.90 ($\pm$ 0.006)
GB	0.83 ($\pm$ 0.003)	0.82 ($\pm$ 0.001)	0.91 ($\pm$ 0.002)
LR	0.82 ($\pm$ 0.003)	0.82 ($\pm$ 0.011)	0.88 ($\pm$ 0.003)

F. Assessing Importance Measures

The relative importance of the factors was computed by exploiting the Gini impurity segregation mechanism of the random forest classifier [19]. Relative importance is determined by the mean decrease in Gini impurity, which is a measure that tree-based classifiers use to segregate the data into groups, across all the trees in a forest. The larger the Gini impurity, the higher is the segregation power of the factor. Importance of interactions is determined by how many times two given factors appear adjacent to each other in a forest. The importance of factors and their interactions was averaged across the 50 trials of the random forest classifier. Both the importance scores are normalized with respect to the topmost score in their respective categories being considered as the maximum score.

III. RESULTS AND DISCUSSION

This section summarizes and discusses the prediction performance and importance measures.

A. Prediction Performance

The average values and confidence intervals ($\alpha = 0.05$) of sensitivity, specificity, and AUC were computed and reported in Table III. Both ensemble classifiers outperform the decision tree model as expected. They can distinguish between adolescents who may or may not have SUD exceptionally accurately, with an average AUC of over 0.90. The prediction stability is also excellent as reflected in the narrow confidence intervals. The ensemble models produce mixed results, with no model decisively outperforming the other. Gradient boosting has better sensitivity and AUC, while random forest has better specificity. Gradient boosting may enjoy better sensitivity because it builds successive trees to improve the prediction of those observations that were missed in the prior trees.

Finally, both ensemble models have higher values of sensitivity, specificity, and AUC over the baseline logistic regression model and the decision tree weak learner. We note that our approach outperforms the prevalent approaches that predict misuse and lifetime use of pharmacological substances. Specifically, Vàzquez et. al. use a sample of Mexican fifth and sixth graders to predict their use of alcohol, marijuana, and tobacco, and inhalants [20]. Across the four substances considered in the study, their average AUC is 0.823. Moreover, it has been shown that adolescents more often use substances beginning in high school, not in middle school [9], [21]. While

TABLE IV
NORMALIZED RELATIVE IMPORTANCE SCORES OF THE FACTORS

Factor	Score	P/D
Seeking approval of peers for the use of substances	1.00	P
Easy availability of substances	0.98	D
Adolescent approving peer use of substances	0.93	P
Approached by someone selling substances	0.87	D
Peers using substances	0.85	P
Adolescent engaging in violent behavior	0.75	P
Risk taking personality of the adolescent	0.69	P
Obesity	0.57	P
Permissive attitudes of the parents	0.55	D
Parenting style and involvement	0.47	P
Perception of risk – weekly use of substances	0.35	P
Race	0.32	P
School environment of the adolescent	0.31	D
Depression and mental illness	0.30	P
Religious beliefs	0.29	P
Size of the family	0.28	P
Overall health of the adolescent	0.27	P
Participation in youth programs, Perception of risk – daily use of substances	0.24	D, P
Use of special drugs	0.23	P
Perception of risk – lifetime use of substances, Metro/Non-Metro residence	0.22	P, D, D
Perception of risk – monthly use of substances	0.19	P
Poverty level of the family	0.18	D
State law whether marijuana is legal, Presence of mother in the household	0.11	D
Participation in drug education programs, Participation of the family in government programs, Seeing drug prevention message outside of school.	0.10	D
Gender	0.09	P
Availability of a support system, Presence of father in the household	0.08	D, D
Presence of father in the household	0.08	D
Participation in youth activities	0.07	D
Family covered by government or private medical insurance	0.03	D

stopping substance use at the source is important, using a prediction model based on fifth and sixth graders may miss those adolescents that initiate substance use in high school. Han et. al. predict opioid misuse among US adolescents with an AUC of 0.811 [22].

This is particularly significant because misuse and lifetime use are comparatively more probable than SUD (20% vs. 5%). Thus, these outcomes should be easier to predict, whereas identifying adolescents with SUD is akin to finding a "needle in a haystack". We believe that our approach shows better performance because the underlying data-driven model is more comprehensive, and we specifically handle class imbalance using the SMOTE procedures.

B. Importance of Factors

Because the classifiers, especially the random forest classifiers, perform extremely well and consistently on an average, it is appropriate to draw inferences about the importance of various factors in predicting SUD among adolescents.

Table IV shows the normalized relative importance scores of factors (listed in a descending order). From the table it can be seen that the top ten factors are split evenly into proximal (P) and distal (D) groups. The table also shows that their relative importance is very close, and hence, no one factor is likely to be dispositive in predicting at risk adolescents.

There is a considerable debate about the relative importance of different risk and protective factors on substance abuse among adolescents, and their respective capabilities to predict at risk adolescents. Cleveland et al. [4] find that the risk factors were stronger predictors of substance use outcomes compared to the mitigation offered by protective factors. Researchers have also found that peer and individual factors had the strongest association with past year marijuana, cigarette, and

TABLE V
NORMALIZED RELATIVE IMPORTANCE SCORES OF THE FACTORS

Factor A	Factor B	Rel. Score
Approached by some selling substances	Obesity	1.00
Easy availability of substances	Obesity	0.95
Seeking approval for use of substances	Obesity	0.94
Approached by someone selling substances	Risk-taking attitude	0.94
Approached by someone selling substances	Parenting style and involvement	0.94
Engaging in violent behavior	Obesity	0.93
Approval for the use of substances by peers	Obesity	0.92
Approached by someone selling substances	Family size	0.90
Easy availability of substances	Risk-taking attitude	0.90
Approached by someone selling substances	Religious beliefs	0.89
Peer approval of the use of substances	Parenting style	0.89
Easy availability of substances	Parenting style	0.89
Peer approval of the use of substances	Risk-taking attitudes	0.88
Approached by someone selling substances	Risk of harm by weekly use	0.88
Permissive parental attitudes	Obesity	0.87
Approached by someone selling substances	Peer use of drugs	0.87
Engaging in violent behavior	Parenting style	0.87
Approached by someone selling substances	Race	0.87
Engaging in violent behavior	Risk-taking attitude	0.87
Approval of use of substances by peers	Risk-taking attitude	0.87

alcohol use [23]. However, other studies have found that family or school factors were among the strongest predictors of adolescent outcomes [24], [25]. This study finds similar results in that risk factors are stronger predictors than mitigating factors. It also affirms that peer and individual factors are more important compared to the characteristics of family, parenting style and school environment. Finally, it is interesting that various drug and violence prevention programs, and youth activities do not rank very high in dominance. This suggests that the threat posed by the risk factors is not easily mitigated or offset by preventive factors such as drug education in schools and participation in various drug prevention programs.

Obesity and risk taking are the two proximal factors that rank highly among the top ten. Obesity may be a manifestation of chronic loss of control and compulsivity. It is possible that it is triggered by easy availability of food and substances. This confirms prior studies, which suggest that food and drugs compete for the same reward pathways in the brain [3].

C. Importance of Factor Interactions

Because no one factor is entirely dominant, as suggested by the closeness of importance of the top ten factors, it is prudent to study the importance of the interactions among these factors. Each factor can interact with every other factor including itself leading to over 1000 interactions.

The normalized relative importance of the top 25 factor interactions is listed in Table V. The most common interactions are being approached with drugs/obesity, easy availability of drugs/obesity, adolescent approval of substance use/obesity, and being approached with drugs/risk taking behavior. Overall, obesity and risk-taking attitudes collide with many characteristics of the adolescents' environment pertaining to neighborhoods, parents and peers. This research thus highlights that neither proximal nor distal factors are superior predictors, but rather SUD in adolescents has a complex etiology and may occur when adolescents with risk-taking and impulsive personalities are placed in environments where substances are easily available.

The relative importance of the interactions between risk and protective factors can inform how intervention strategies must be designed and deployed to derive maximum benefit. For example, heavy premium should be placed on limiting access to and the availability of substances, which will be challenging given the changing landscape on the legalization of marijuana [26]. This is especially important because prior research has shown that use of drugs in adolescence is a strong determinant of drug abuse later in adult life [27]–[29]. Additionally, drug prevention and education programs should not only raise awareness of the ill-effects and dangers of substances, but they should also counsel adolescents on controlling their impulsivity and risk-taking tendencies.

IV. CONCLUSIONS AND FUTURE RESEARCH

This paper presents an approach that explores the relationships between many factors leading to a holistic view of an

adolescent's sphere of life and SUD to predict at-risk adolescents with exceptional accuracy. It finds that when factors reflecting the environment in which an adolescent is nurtured collide with inherent risk-taking and impulsive personality traits of an adolescent, it incurs a significant threat to the development of SUD. It is thus inferred that limiting easy access to substances in adolescents' communities should be of prime importance to alleviate drug abuse. On a broader note, this research, among others, demonstrates that vast amounts of information could be mined from public domain data sets to improve our understanding of SUD in various cohorts.

Certain key factors, however, which may influence substance use in adolescents are missing from the NSDUH survey. One such variable is the use of substances by parents and guardians. Integrating alternative sources of data to remedy this shortcoming is a topic of the future.

REFERENCES

[1] H. B. Shakya, N. A. Christakis, and J. H. Fowler, "Parental Influence on Substance Use in Adolescent Social Networks," *Archives of Pediatrics & Adolescent Medicine*, vol. 166, no. 12, pp. 1132–1139, Dec. 2012. [Online]. Available: https://jamanetwork.com/journals/jamapediatrics/fullarticle/1377497

[2] F. Denoth, V. Siciliano, P. Iozzo, L. Fortunato, and S. Molinaro, "The Association between Overweight and Illegal Drug Consumption in Adolescents: Is There an Underlying Influence of the Sociocultural Environment?" *PLOS ONE*, vol. 6, no. 11, p. e27358, Nov. 2011. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0027358

[3] R. A. Sansone and L. A. Sansone, "Obesity and Substance Misuse: Is There a Relationship?" *Innovations in Clinical Neuroscience*, vol. 10, no. 9-10, pp. 30–35, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3849872/

[4] M. J. Cleveland, M. E. Feinberg, D. E. Bontempo, and M. T. Greenberg, "The role of risk and protective factors in substance use across adolescence," *The Journal of Adolescent Health: Official Publication of the Society for Adolescent Medicine*, vol. 43, no. 2, pp. 157–164, Aug. 2008.

[5] Substance Abuse and Mental Health Services Administration, "National Survey on Drug Use and Health 2016 Edition," 2016. [Online]. Available: https://www.datafiles.samhsa.gov/

[6] ——, "National Survey on Drug Use and Health 2017 Edition," 2017. [Online]. Available: https://www.datafiles.samhsa.gov/

[7] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition ed. American Psychiatric Association, 2000. [Online]. Available: https://dsm.psychiatryonline.org/dsmPreviousEditions

[8] R. N. Lipari, "Key Substance Use and Mental Health Indicators in the United States: Results from the 2018 National Survey on Drug Use and Health," p. 82, 2018.

[9] R. N. Lipari, S. L. Van Horn, A. Hughes, and M. Williams, "State and Substate Estimates of Nonmedical Use of Prescription Pain Relievers," in *The CBHSQ Report*. Rockville (MD): Substance Abuse and Mental Health Services Administration (US), 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK448248/

[10] E. Barenholtz, N. D. Fitzgerald, and W. E. Hahn, "Machine-learning approaches to substance-abuse research: emerging trends and their implications," *Current Opinion in Psychiatry*, vol. 33, no. 4, pp. 334–342, Jul. 2020.

[11] J. M. Benitez, J. L. Castro, and I. Requena, "Are artificial neural networks black boxes?" *IEEE transactions on neural networks*, vol. 8, no. 5, pp. 1156–1164, 1997.

[13] S. Sperandei, "Understanding logistic regression analysis," *Biochemia Medica*, vol. 24, no. 1, pp. 12–18, Feb. 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3936971/

[12] M. Bramer, "Avoiding Overfitting of Decision Trees," in *Principles of Data Mining*, ser. Undergraduate Topics in Computer Science, M. Bramer, Ed. London: Springer, 2013, pp. 121–136. [Online]. Available: https://doi.org/10.1007/978-1-4471-4884-5-9

[14] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction," *BMC Medical Research Methodology*, vol. 19, no. 1, p. 64, Mar. 2019. [Online]. Available: https://doi.org/10.1186/s12874-019-0681-4

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, arXiv: 1106.1813. [Online]. Available: http://arxiv.org/abs/1106.1813

[16] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145 – 1159, 1997. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320396001422

[17] M. E. Rice and G. T. Harris, "Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r," *Law and Human Behavior*, vol. 29, no. 5, pp. 615–620, Oct. 2005.

[18] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://www.R-project.org/

[19] S. Nembrini, I. R. König, and M. N. Wright, "The revival of the Gini importance?" *Bioinformatics*, vol. 34, no. 21, pp. 3711–3718, Nov. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6198850/

[20] A. L. Vázquez, M. M. Domenech Rodríguez, T. S. Barrett, S. Schwartz, N. G. Amador Buenabad, M. N. Bustos Gamiño, M. d. L. Gutiérrez López, and J. A. Villatoro Velázquez, "Innovative Identification of Substance Use Predictors: Machine Learning in a National Sample of Mexican Children," *Prevention Science: The Official Journal of the Society for Prevention Research*, vol. 21, no. 2, pp. 171–181, 2020.

[21] National Institute on Drug Abuse, *Monitoring the Future Survey*, Jan. 2019.

[22] D.-H. Han, S. Lee, and D.-C. Seo, "Using machine learning to predict opioid misuse among U.S. adolescents," *Preventive Medicine*, vol. 130, p. 105886, 2020.

[23] D. A. Wright and M. Pemberton, *Risk and protective factors for adolescent drug use: Findings from the 1999 National Household Survey on Drug Abuse*. Department of Health and Human Services, Substance Abuse and Mental Health ..., 2004.

[24] S. Case, "Indicators of adolescent alcohol use: a composite risk factor approach," *Substance Use & Misuse*, vol. 42, no. 1, pp. 89–111, 2007.

[25] W. Kliewer and L. Murrelle, "Risk and protective factors for adolescent substance use: findings from a study in selected Central American countries," *The Journal of Adolescent Health: Official Publication of the Society for Adolescent Medicine*, vol. 40, no. 5, pp. 448–455, May 2007.

[26] A. Joffe and W. S. Yancy, "Legalization of Marijuana: Potential Impact on Youth," *Pediatrics*, vol. 113, no. 6, pp. e632–e638, Jun. 2004, publisher: American Academy of Pediatrics Section: ELECTRONIC ARTICLES. [Online]. Available: https://pediatrics.aappublications.org/content/113/6/e632

[27] E. Silins, L. J. Horwood, G. C. Patton, D. M. Fergusson, C. A. Olsson, D. M. Hutchinson, E. Spry, J. W. Toumbourou, L. Degenhardt, W. Swift, C. Coffey, R. J. Tait, P. Letcher, J. Copeland, R. P. Mattick, and Cannabis Cohorts Research Consortium, "Young adult sequelae of adolescent cannabis use: an integrative analysis," *The Lancet. Psychiatry*, vol. 1, no. 4, pp. 286–293, Sep. 2014.

[28] A. S. Wadekar, "Understanding Opioid Use Disorder (OUD) using tree-based classifiers," *Drug and Alcohol Dependence*, vol. 208, p. 107839, Mar. 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0376871620300041

[29] R. L. DuPont, B. Han, C. L. Shea, and B. K. Madras, "Drug use among youth: National survey data support a common liability of all drug use," *Preventive Medicine*, vol. 113, pp. 68–73, 2018.