Identifying Homeless Youth At-Risk of Substance Use Disorder: Data-Driven Insights for Policymakers

Maryam Tabar
The Pennsylvania State University
mfg5544@psu.edu

Heesoo Park
Sungkyunkwan University
righ120@skku.edu

Stephanie Winkler
The Pennsylvania State University
sxw96@psu.edu

Dongwon Lee
The Pennsylvania State University
dongwon@psu.edu

Anamika Barman-Adhikari
University of Denver
Anamika.BarmanAdhikari@du.edu

Amulya Yadav
The Pennsylvania State University
amulya@psu.edu

ABSTRACT

Substance Use Disorder (SUD) is a devastating disease that leads to significant mental and behavioral impairments. Its negative effects damage the homeless youth population more severely (as compared to stably housed counterparts) because of their high-risk behaviors. To assist policymakers in devising effective and accurate long-term strategies to mitigate SUD, it is necessary to critically analyze environmental, psychological, and other factors associated with SUD among homeless youth. Unfortunately, there is no definitive data-driven study on analyzing factors associated with SUD among homeless youth. While there have been a few prior studies in the past, they (i) do not analyze variation in the associated factors for SUD with geographical heterogeneity in their studies; and (ii) only consider a few contributing factors to SUD in relatively small samples. This work aims to fill this gap by making the following three contributions: (i) we use a real-world dataset collected from ~1,400 homeless youth (across six American states) to build accurate Machine Learning (ML) models for predicting the susceptibility of homeless youth to SUD; (ii) we find a representative set of factors associated with SUD among this population by analyzing feature importance values associated with our ML models; and (iii) we investigate the effect of geographical heterogeneity on the factors associated with SUD. Our results show that our system using adaptively boosted decision trees achieves the best predictive accuracy out of several algorithms on the SUD prediction task, achieving an Area Under the ROC Curve of 0.85. Further, among other things, we also find that both Post-Traumatic Stress Disorder (PTSD) and depression are very strongly associated with SUD among homeless youth because of their propensity to self-medicate to alleviate stress. This work is done in collaboration with social work scientists, who are currently evaluating the results for potential future deployment.

CCS CONCEPTS

• **Applied computing** → **Health care information systems**; • **Computing methodologies** → *Classification and regression trees.*

KDD '20, August 23–27, 2020, Virtual Event, CA, USA

ACM ISBN 978-1-4503-7998-4/20/08…$15.00
https://doi.org/10.1145/3394486.3403360

KEYWORDS

AI for Social Good, Data Science for Social Good, Substance Use Disorder Prediction, Homeless Youth

ACM Reference Format:
Maryam Tabar, Heesoo Park, Stephanie Winkler, Dongwon Lee, Anamika Barman-Adhikari, and Amulya Yadav. 2020. Identifying Homeless Youth At-Risk of Substance Use Disorder: Data-Driven Insights for Policymakers. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20), August 23–27, 2020, Virtual Event, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394486.3403360

1 INTRODUCTION

Substance use disorder (SUD) refers to a pattern of harmful substance use (e.g., alcohol, marijuana, street and prescription opioids, stimulants, etc.) resulting in significant impairments [1]. Despite their negative side effects, sufferers continue to use these substances. SUD is a widespread and costly issue in the United States with abuse of tobacco, alcohol, and illicit drugs imposing over $740 billion each year [26]. In fact, about 19.7 million adults were reportedly suffering from SUD in 2017 [39]. More importantly, SUD-related mortality rate has been increasing every year - it rose from 16 cases per 100,000 people (in 2002) to 27.5 cases per 100,000 (in 2015) [35].

In particular, SUD is more prevalent among the homeless youth population compared to the general public. For example, Busen and Engebretson [5] found that about 46% of their surveyed homeless youth suffered from SUD. Thus, any attempt at tackling SUD at a national level crucially depends on our success at minimizing the rates of SUD among homeless youth.

To tackle this problem, policymakers often design and implement state-level programs and initiatives to mitigate the prevalence of SUD among homeless youth. Yet, their policies are often not backed by data-driven insights about how different environmental, psychological and other possible factors play a role in the prevalence of SUD in their communities. Thus, it is often hard to ascertain the accuracy and effectiveness of their planned initiatives. To overcome this research gap, it is necessary to critically analyze different factors associated with SUD among homeless youth, which can be used to provide principled insights to policymakers in their efforts to developing long-term policies to tackle SUD. Further, the factors associated with SUD may vary geographically (from state to state), hence it becomes necessary to analyze this variation in a principled manner so that policymakers in different states can be provided different insights on associated factors.

There has been some prior work at the intersection of artificial intelligence (AI) and social sciences with the goal of mitigating problems faced by homeless population, e.g., Yadav et al. [48] and Rahmattalabi et al. [29] focused on enhancing HIV and substance abuse prevention interventions among the homeless youth population, respectively. Similarly, Tyler et al. [43] focused on the factors associated with illicit drug use among homeless young adults. Unfortunately, there is no definitive data-driven study to predict susceptibility to SUD among homeless youth, and to analyze factors associated with SUD among this population. Existing studies either (i) do not analyze variation in the associated factors for SUD with geographic heterogeneity in their studies; and (ii) only consider a few contributing factors to SUD in relatively small samples. Due to these deficiencies, results obtained from existing studies may not be generalizable.

This paper tackles the aforementioned challenges by addressing three research questions:

RQ1: Can we build accurate ML models that can predict homeless youth's susceptibility to SUD?

RQ2: What set of environmental, psychological, and behavioral factors is highly associated with SUD among homeless youth?

RQ3: Do factors associated with SUD among homeless youth vary with geographical differences?

In particular, this paper makes three novel contributions. First, we use a real-world dataset collected from ∼1,400 homeless youth across six states in USA and build accurate ML models that can predict each homeless youth's susceptibility to SUD. Our best performing model achieves an Area Under the ROC Curve (AUC) of 0.85, which illustrates the high accuracy of our ML model. Second, we find a representative set of environmental, psychological, and behavioral factors associated with SUD among this population by analyzing feature importance values associated with our ML models. Finally, we investigate the effect of geographical heterogeneity on the factors associated with SUD. Based on our study results, we find that PTSD and depression are very strongly associated with SUD among homeless youth because of their propensity to self-medicate to alleviate stress. In addition, we find that states' gun control policies and incarceration rates can potentially influence the level of association between SUD and certain relevant factors.

2 RELATED WORK

In this section, we survey recent studies on alleviating the problems faced by the homeless population. These studies fall into two broad scientific areas: AI and social science.

Artificial Intelligence Research. To the best of our knowledge, there has been no prior work on building and understanding models for predicting SUD among homeless youth. There has been a lot of interest in predicting substance use from social media data. Ding et al. [12] took advantage of several ML and text mining techniques to predict SUD. Hassanpour et al. [15] utilized a deep learning approach to predict the risk of substance use from Instagram profile data. However, the focus of these studies was mainly on the general population, and thus, their results might not apply readily to homeless youth. Also, there is a growing body of work in AI research on tackling problems faced by homeless youth. Yadav et al. [48, 49] and Rahmattalabi et al. [29, 30] focused on preventing HIV, substance abuse, and suicidal tendencies among the homeless youth population. However, most prior work in this space is concerned with finding prescriptive solutions, e.g., Yadav et al. [48] prescribes the selection of key influential homeless youth to spread awareness about HIV. On the other hand, our work is predictive, i.e., we aim to identify those at high risk of SUD and uncover factors associated with this disorder among homeless youth.

Social Science Research. Research with homeless populations is conducted in multiple social science disciplines with much of the work coming from sociology and psychology. While some of this work examines the effectiveness of interventions to address problems associated with homelessness, prior work also examines the experience of being homeless and how this relates to other aspects of an individual's life and well-being. Specifically, research investigates factors associated with an individual developing SUD. These factors can help identify at risk individuals, which is important for outreach centers as they intervene in homeless populations. In particular, any form of child maltreatment (especially physical or sexual abuse) is shown to be a factor strongly associated with SUD [11, 43, 45]. While on the streets, trauma remains an associated factor for SUD irrespective of whether the individual witnessed a friend or loved one being victimized (indirect victimization), or if they had experienced the trauma themselves (direct victimization) [42]. Mental health disorders are also factors associated with SUD [11]. Other factors linked to SUD include demographic characteristics such as gender and age with young homeless men considered as one of the highest risk groups [43, 45]. Typically, prior studies in this space only choose two or three groups of factors to investigate. In contrast, our work is the first attempt at examining multiple factor groups at the same time (e.g., adverse childhood experiences, victimization, stress, mental health, demographics etc.), which can reveal their association to SUD. More importantly, examining multiple factors at once enables us to compare the relative importance of different factors.

3 PREPARATION

Real-world Dataset. The dataset was collected from 1,426 homeless youth across six states in USA, namely California (CA), Arizona (AZ), Colorado (CO), Missouri (MO), Texas (TX), and New York (NY), from June 2016 until July 2017. Each homeless youth was given a questionnaire to fill up, which consisted of questions about various topics. Table 1 represents a couple of those topics, along with the features corresponding to a couple of sample questions under these topics. This survey was approved by institutional review boards in all six states. For more information regarding the data collection procedures, please refer to Barman-Adhikari et al. [2].

Data Pre-processing. We pre-process the original dataset in two steps. First, as there are a lot of missing entries (∼18.5%) in our dataset (as homeless youth could choose not to answer a question that made them feel uncomfortable), we used the MissForest algorithm [37], an off-the-shelf data imputation method to impute missing feature values in our dataset. Second, we apply feature standardization (i.e., Z-score normalization) to all features in our dataset. Finally, we create the training and test sets. To address RQ1 and RQ2, we randomly select 80% of samples as the training set and

Table 1: Summary of questionnaire topics with a couple of sample questions.

Topic/Feature Block	Feature	Explanation
Socio-demographic (SD)	gender	Male, Female, Transgender, Gender queer, and other
Criminal History (CH)	jail_homeless	Any jail or prison experiences since becoming unstably housed or homeless
	gunaccess	Having access to a gun or knowing how to access a gun easily
	avoid_police	Purposely avoiding situations that may expose you to interaction with police
Sexual-risk behaviors (SR)	life_sexpartners	The number of sex partners in life
	last_sui_di	Drinking alcohol or using drugs before having sexual intercourse
	online_sexpart	Having sex with someone you met online
Victimization Experiences (VE)	ace	Experience of trauma and stress in childhood
	anyst_phy_vict	Any physical street victimization (e.g., assaulted with a weapon)
	witness_gun_di	Witnessing someone get attacked by a gun
Gang Involvement (GI)	Juggalo_di	Ever been a Juggalo or a Juggalette?
Mental Health Characteristics (MH)	depression	The 9-item questionnaire (PHQ-9) is used to assess the level of depression
	ptsd	A 4-item questionnaire is used to measure PTSD
	perc_stress	Perceived stress during the past month
	unmet_ever	History of unmet mental health needs
	hospit_ever	History of staying in a hospital to treat mental health conditions
	medication_ever	Using medication to treat mental health conditions
	cope_8	How often do you use anger to get out of painful situations
	cope_9	How often do you use drugs or alcohol to deal with problems
Technology Access (TA)	soc_media_prof	Having a profile on a social media site

Table 2: Summary/Examples of dataset features.

Feature Type	Number	Examples
Numeric	19	life_sexpartners
Ordinal	32	cope_9
Nominal (Binary)	161	arrest_unstable
Nominal (Non-binary)	20	gender, ethnicity

consider the remaining 20% as the test set. To address RQ3, first, the dataset is split into six smaller datasets, each of which includes the samples of a particular state. Then, for each of these six datasets, 80% of samples are selected as the training set and the remaining 20% make the test set. For each RQ, the class distribution in the training and test sets is set to be the same as in the full dataset. At the end of this process, our dataset had 1,367 data points (one for each homeless youth), each of which had 231 features and a binary label for predicting SUD. Table 2 represents a summary of the features in our final dataset.

4 RQ1: PREDICTION MODEL FOR SUD

Predicting SUD. We formulate the problem of predicting the susceptibility of homeless youth to SUD as a binary classification problem. To find the best performing model, we compared the predictive performance of the following classification models:

- Logistic Regression (Logit)
- Classification And Regression Tree (CART) [4]
- Conditional Inference Forest (CForest) [18], which is an ensemble method using conditional inference trees as base learners.
- Adaptive Boosting (AdaBoost) [13]
- eXtreme Gradient Boosting (XGBoost) [9]
- Support-Vector Machine (SVM) with Radial kernel [7].
- Multi-Layer Perceptron (MLP) [16] with the ReLU activation function and two hidden layers; the number of neurons in each hidden layer is half of that in the previous layer.

Further pre-processing steps need to be taken before fitting Logit, SVM, and MLP. As SVM internally uses the Euclidean distance metric, it cannot be applied to categorical variables (with more than two levels) in theory. Also, the inputs of Logit and MLP need to be numeric or binary categorical. Therefore, to fit these three models, we convert the categorical features in our dataset into numeric ones, by using one-hot encoding, i.e., we represent a categorical variable with K (> 2) levels using K different binary variables, only one of which is allowed to have a value of 1 at a given time.

The hyper-parameters for all our models are tuned using K-fold cross-validation ($K = 10$). In addition, for training the MLP model, we use Adam [21] as the optimizer, the batch size was set to 32, learning rate was 0.001, and we trained for 50 epochs. Table 3 compares the predictive performance of all our ML models across several widely used evaluation metrics. The rows in this table represent different classification algorithms and the columns represent different evaluation metrics (Accuracy, Precision, Recall, F1, and AUC). According to the results in Table 3, AdaBoost is the best performing model in terms of all evaluation metrics. In particular, it achieves an AUC of 0.8546 which indicates its excellent class separation capability. Surprisingly, Table 3 shows that AdaBoost is much more accurate than our MLP model, although given our small-sized dataset, it is difficult to train a more accurate MLP model.

In summary, the results from Table 3 show that it is indeed possible to train highly accurate ML models to predict the susceptibility

Table 3: Performance of different ML models on predicting the susceptibility of homeless youth to SUD.

Model	Accuracy	Precision	Recall	F1	AUC
Logit	0.7032	0.5729	0.5789	0.5759	0.7776
CART	0.7289	0.6779	0.4210	0.5194	0.6850
CForest	0.7728	0.7619	0.5052	0.6075	0.8507
AdaBoost	**0.7985**	**0.7702**	**0.6000**	**0.6745**	**0.8546**
XGBoost	0.7545	0.7000	0.5157	0.5939	0.8304
SVM	0.7692	0.7285	0.5368	0.6181	0.8360
MLP	0.7362	0.6575	0.5052	0.5714	0.7010

of homeless youth to SUD, which helps answer RQ1 in the affirmative. Given AdaBoost's superiority over all other models, finally, we use AdaBoost as our model of choice in the rest of the paper.

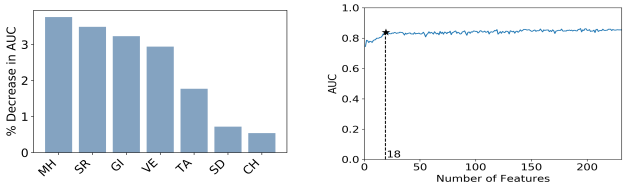
Predicting Alternative Substance Use. SUD is associated with the recurrent use of several different substances (drugs), e.g., heroin, ecstasy, etc. Efficiently tackling SUD requires a determination and consideration of which (and how many) alternative substances/drugs are used by homeless youth [25]. Therefore, having an accurate ML model capable of predicting the use of different kinds of alternative substances along with an accurate SUD predictor could be helpful in delivering more appropriate intervention programs to the homeless youth population. Next, we show that it is possible to train highly effective ML models to predict alternative substance use for six different substances: alcohol, marijuana, heroin, crack, methamphetamine (meth), and ecstasy.

Similar to SUD prediction, we model the substance use prediction problem as a binary classification problem in which the label indicates whether the homeless youth has used that specific drug during the past 30 days or not. Then, the same data preparation procedure is applied to prepare a separate training/test dataset for each of the six substances. For heroin, crack, meth, and ecstasy, the resulting dataset was highly imbalanced - only ~13% of homeless youth in our dataset had used each of these specific drugs, on average. We account for this data imbalance by applying the SMOTE oversampling technique [8] as a pre-processing step, before training our AdaBoost model. Table 4 represents the performance of AdaBoost in predicting alternative substance use. The rows show the different alternative substances and the columns represent evaluation metrics. This table shows that on average, the AdaBoost model has a relatively high AUC (> 0.7), which shows that in addition to accurate prediction of SUD, it is also possible to develop highly accurate models for predicting alternative substance use.

Ablation Studies for SUD Prediction Next, we conduct a preliminary investigation into the relative importance of different sets of features in the predictive accuracy of our AdaBoost model. Specifically, we conduct an ablation study as follows: (i) we divide the features in our dataset into seven separate feature blocks (as shown in Table 1); each feature block consists of features related to a specific topic, e.g., one feature block ascertains involvement with gangs (GI), another block ascertains criminal history (CH), etc.; (ii) we remove one feature block from the feature space (at a time), and then re-train an AdaBoost model on the remaining set of features;

Table 4: Performance of AdaBoost on predicting alternative substance use among homeless youth.

Substance	Accuracy	Precision	Recall	F1	AUC
Alcohol	0.6605	0.6540	0.7375	0.6933	0.7098
Marijuana	0.6948	**0.7277**	**0.7939**	**0.7594**	0.7196
Heroin	0.8917	0.3000	0.1200	0.1714	0.7764
Crack	**0.8955**	0.3500	0.3181	0.3333	**0.8128**
Meth	0.8171	0.4583	0.2340	0.3098	0.7498
Ecstasy	0.8764	1.000	0.1951	0.3265	0.7952



(a) % Decrease in AUC After Ablation. **(b) AUC with different number of features.**

Figure 1: Ablation Results & Finding Important Features.

(iii) finally, we report the percentage decrease in AUC values for our model.

Figure 1a shows the result of ablating different feature blocks. The X-axis shows the ablated feature block and the Y-axis shows the percentage decrease in AUC. According to the results, among all feature blocks, removing *mental health characteristics* (MH) leads to the greatest decrease in the model's predictive accuracy. At the same time, *sexual risk behavior* (SR), *gang involvement* (GI), and *victimization experiences* (VE) also lead to large decreases in the model's AUC. These findings are consistent with a large body of literature that has established strong connections between mental health [17], sexual risk behavior [46], and victimization experiences [10] and SUD [19, 33]. Inspired by these findings, we now delve deeper to build a comprehensive understanding of the different kinds of factors associated with SUD.

5 RQ2: UNCOVERING FACTORS ASSOCIATED WITH SUD

Our high-level goal in this paper is to find a representative set of environmental, psychological and behavioral factors associated with SUD among the homeless youth population, which can be used to provide principled insights to policymakers and practitioners in their efforts to developing long-term policies to tackle SUD. In this section, we attempt to achieve this goal by analyzing feature importance values associated with our SUD prediction model. We use the *Mean Decrease in Impurity* (MDI) [24], a well-known metric for tree-based ML models, to find feature importance values. However, due to the lack of space, we restrict our attention to analyzing importance of only those features that play a key role in identifying homeless youth susceptible to SUD.

To discover this subset of "important" features, we do the following: (i) we rank all features in our dataset based on their MDI values; (ii) starting from the most important feature, we add features one-by-one in the decreasing order of importance to the dataset and re-train a separate AdaBoost model (with only the restricted set of features). Figure 1b shows the AUC of the AdaBoost models trained with the increasing number of features. The X-axis shows the (increasing) number of features used to train the model and the Y-axis shows the AUC of the resulting AdaBoost model. This figure exhibits diminishing returns (in terms of the increases in AUC) beyond the addition of the 18 most important features in our dataset. Thus, we restrict our attention in RQ2 to these 18 features.

Figure 2 shows these 18 features ranked according to their normalized MDI (NMDI) values. The definition of these features can be seen in Table 1. Overall, we categorize these 18 features into three broad categories: environmental factors, psychological factors, and sexual-risk behaviors, and further analyze these three categories in detail.

Environmental Factors. According to our study, environmental factors play a key role in SUD among homeless youth population. In particular, our study indicates that some specific types of direct victimization, e.g., experience of physical street victimization, (`anyst_phys_vict`, NMDI=0.508) and indirect gun victimization, i.e., witnessing another person being victimized by gun, (`witness_gun_di`, NMDI=0.500) are highly associated with SUD among homeless youth. On the other hand, we observe that the alternate types of victimization (e.g., sexual assault) are less strongly associated with SUD (Average NMDI=0.369). In addition, adverse childhood experiences (`ace`, NMDI=0.547) and perceived stress (`perc_stress`, NMDI=0.662) are also strongly associated with SUD. This finding is consistent with existing literature as follows: (i) there is a lot of prior work which hypothesizes that homeless people's lifestyle (e.g., sleeping outside) increases the likelihood of experiencing victimization [22]. For example, Stewart et al. [38] shows that ~85% of the homeless population have experienced trauma and victimization. (ii) These victimization experiences are shown to be significantly related to psychological distress and painful situations among youth [14]. (iii) The importance of these factors along with factors related to coping strategies (`cope_8` and `cope_9`, average NMDI=0.749) are consistent with prior work on SUD in homeless populations which shows that these youth self-medicate substances to alleviate the effect of painful situations and to cope with stressful situations [44].

In addition, factors associated with law enforcement such as avoiding police officers or the places where police officers might be found (`avoid_police`, NMDI=0.468), and being arrested since being homeless (`jail_homeless`, NMDI=0.498) are also strongly associated with SUD among homeless youth. Intuitively, this is possible since SUD involves the use of illicit substances (e.g., non prescription use of opioids, crack, cocaine), and an encounter with law enforcement could result in the individual being arrested and sent to jail. Even someone using a legal substance (e.g., alcohol) could be arrested for being intoxicated in public. Given the high punitive cost of engagement with law enforcement agencies, therefore, it is reasonable to expect that youth suffering from SUD would prefer to avoid encounters with law enforcement, or else they might end up in jail at some point during their time on the streets.

Psychological Factors. According to our study, psychological factors play a key role in SUD among homeless youth. In particular, certain mental health disorders (e.g., `ptsd`, `depression`) and mental health needs (e.g., `unmet_ever`, `hospit_ever`, `medication_ever`) are highly associated with SUD among this population. Our study indicates that PTSD (`ptsd`, NMDI=0.589) and depression (`depression`, NMDI=0.687) are more important than other mental health disorders (Average NMDI=0.213). This is consistent with prior literature [23], which suggests that people struggling with PTSD self-medicate and use substances to cope with PTSD symptoms. Furthermore, the simultaneous presence of PTSD and depression among the highly associated factors along with the victimization experiences feature block is consistent with prior work [20], which shows that the comorbidity of depression and PTSD are highly likely among adolescents with victimization experiences.

Our dataset consists of information about the following eight disorders: PTSD, depression, attention deficit hyperactivity disorder (ADHD), oppositional defiant disorder (ODD), conduct disorder (CD), bipolar disorder (BD), schizophrenia, and anxiety disorder. Previous research has shown that in the general population, the risk of developing SUD in individuals with BD and ODD is higher than that in individuals with PTSD [40]. Also, in comparison with those having other mental disorders, individuals struggling with depression have a lower risk of developing SUD [40]. However, we observe that even though ADHD, BD, depression, and PTSD have been shown to be prevalent among homeless youth [3, 5], PTSD and depression are more associated with SUD than the other mental disorders.

Sexual-risk behaviors. According to our study, factors pertaining to sexual risk behaviors also play a key role in SUD among homeless youth. In particular, we observe that the number of sex partners (`life_sexpartners`, NMDI=0.729), using substances before sex (`last_sui_di`, NMDI=0.724), having sex with someone you met online (`online_sexpart`, NMDI=0.461) are highly associated with SUD among homeless youth. This is consistent with existing literature [32], which studied the relationship between drug use and sexual risk behaviors. In particular, they explained that sex partners of drug users are highly likely to use drugs, and in this case, factors pertaining to sexual risk behaviors can be related to SUD. In summary, our feature importance analysis shows that there is a strong association between certain sexual-risk behaviors and SUD among homeless youth and it should be considered during intervention programs to deliver effective services.

In conclusion, we find many intertwined factors playing a key role in SUD among homeless youth. These factors can be categorized into environmental, psychological factors, and sexual risk behaviors. Our study indicates that adverse childhood experiences (NMDI=0.547), physical street victimization (NMDI=0.508), indirect gun victimization (NMDI=0.500), and perceived level of stress (NMDI=0.662) are more strongly associated with SUD as compared to other types of victimization (NMDI=0.369). Finally, PTSD (NMDI=0.589) and depression (NMDI=0.687) are found to be more strongly associated with SUD than the other mental health disorders among homeless youth (Average NMDI=0.213).

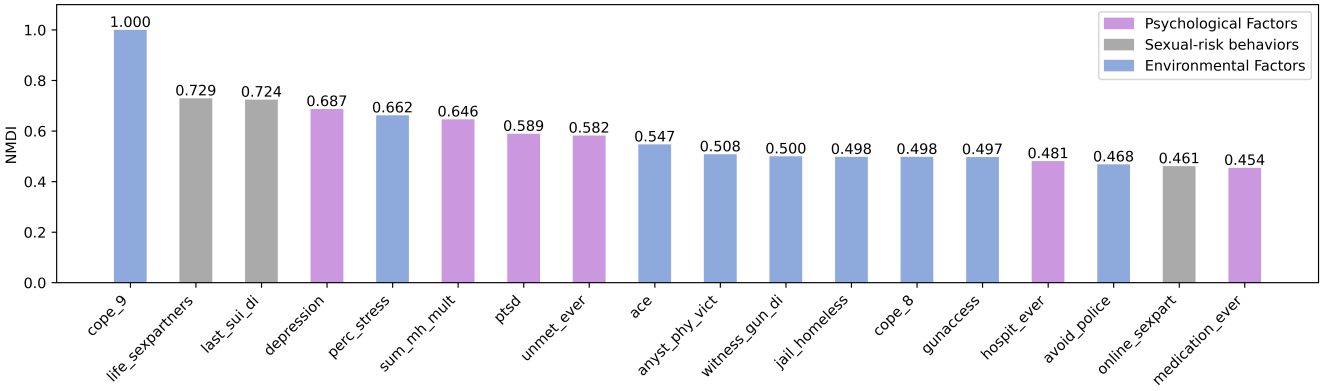


Figure 2: Factors associated with SUD among homeless youth.

Table 5: Performance of AdaBoost across various states.

State	Accuracy	Precision	Recall	F1	AUC
CA	0.6883	0.5909	0.4642	0.5199	0.6727
AZ	0.7000	0.6153	0.5333	0.5714	0.7093
CO	0.6578	0.6470	**0.6111**	**0.6285**	0.7972
MO	0.7027	0.5000	0.4545	0.4761	0.7167
TX	0.7435	**0.6666**	0.3333	0.4444	**0.8487**
NY	**0.7631**	0.5000	0.1111	0.1818	0.6513

6 RQ3: VARIATION IN ASSOCIATED FACTORS WITH GEOGRAPHICAL DIFFERENCES

There is a large amount of diversity among different states in USA in terms of (i) legalization of different drugs; (ii) strictness of gun laws; (iii) rates of homelessness and conditions of homeless youth; (iv) other cultural and environmental conditions, etc. Given these differences, it is not unreasonable to expect that policies and solutions for mitigating SUD among homeless youth that work well in one city/state may not necessarily generalize to other states. Thus, the factors associated with SUD may vary geographically (from state to state), hence it becomes necessary to analyze this variation in a principled manner, so that policymakers and practitioners in different states can be provided different insights on associated factors. This enables policymakers to come up with state-specific policies for mitigating SUD.

In this section, therefore, we analyze these geographical differences by: (i) dividing our dataset into six smaller datasets, each of which contains data from homeless youth belonging to a particular state. (ii) Next, we trained a separate AdaBoost model for SUD prediction in each state. Table 5 represents the predictive performance of AdaBoost for different states. This table shows that across all states, the average AUC is ~0.7, which indicates a high class separation capacity. This shows that we are able to successfully train accurate ML models to predict the susceptibility of homeless youth to SUD across different states. Next, we analyze feature importance values to understand differences between states.

Importance of Criminal Justice History. One major issue among the homeless youth population is the high rate of arrest and incarceration. In fact, past literature has shown that being homeless increases the likelihood of criminal offenses and consequently, multiplies the risk of arrest[1] [36]. This pattern has motivated us to investigate the association of criminal justice history and SUD across different states. To measure criminal justice history (after turning 18), the participants were asked to answer the following questions, each of which corresponds to a predictor variable (feature) of the prediction model.

- ever_arrest: Have you ever been arrested since turning 18?
- arrest_unstable: Since becoming unstably housed or homeless, have you been arrested?
- ever_jail: Have you ever been in jail or prison since turning 18?
- med_jail: Did you ever receive medication for your behavior or mood while you were in jail or prison?
- jail_homeless: Since becoming unstably housed or homeless, have you been in jail or prison?

The total importance of criminal justice history in predicting SUD has been defined as the average of the normalized importance of these features. Figure 3 compares the level of association between criminal justice history and SUD. In this figure, the background color of each state shows the incarceration rate with dark red representing the highest rate and yellow representing the lowest rate. This rate is originally obtained using the Bureau of Justice statistics data for 2017 [41]. Each circle shows the degree of association between criminal justice history and SUD for that specific state; the bigger the radius of the circle, the stronger is the association. According to the results, in general, states with higher incarceration rates have a stronger association between SUD and criminal justice history. In general, this is consistent with a couple of previous studies. Past literature has shown that homeless people are more likely to get incarcerated[1] and further, about 65% of inmates struggle with SUD which often co-occurs with mental illnesses[2][31]. Unfortunately, there is also a lack of sufficient treatment for SUD

[1]https://nlchp.org//wp-content/uploads/2018/10/Housing-Not-Handcuffs.pdf
[2]https://www.centeronaddiction.org/newsroom/press-releases/2010-behind-bars-II

inside prisons and only a few inmates receive treatment[2]. This can increase the risk of relapsing to drug use after release and consequently, returning to criminal activities[3]. Note that there is one exception to this pattern. The association between SUD and criminal justice history in Missouri is weaker than that in Colorado and California, while the incarceration rate in Missouri is higher than Colorado and California. We have not found a comprehensive explanation for this anomaly, though we cannot discount demographic differences in the individual samples for each state.

Given the concerning rate of incarceration and SUD among the homeless population, immediate action needs to be taken to alleviate this problem to break the cycle. One possible way to tackle this issue would be providing more effective SUD treatment for incarcerated individuals, especially in those states with a higher level of association.

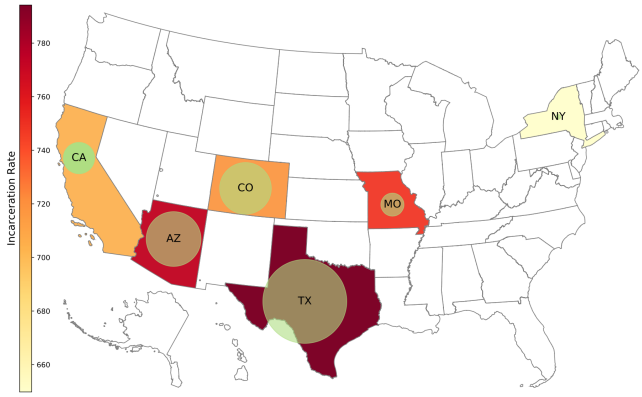


Figure 3: Comparison of the importance of criminal justice history across states (AZ=0.513, CA=0.293, CO=0.484, MO=0.216, TX=0.786, NY=0.045).

Importance of Juggalo Experiences. Gang involvement, while a major concern for the United States as a whole, is of particular importance to the homeless population. Specifically, gangs support the homeless youth population by providing basic needs, such as food, medicines, etc. As a result, gang-involvement tends to be prevalent among homeless youth. In addition, gang membership is typically accompanied by violence and substance use [34, 50]. While there are many different gangs, Juggalos are important to discuss when examining the homeless youth population. Juggalos are defined as fans of groups associated with the Psychopathic Records label and are popular among homeless youth because of their tendency to embrace poverty and a lifestyle outside of mainstream life [28]. Classified as a gang by the FBI in 2011, Juggalos are stereotyped as being violent, young criminals[4] [27]. Past literature has shown that Juggalos are at higher risk of misuse of certain substances (such as marijuana and meth) and victimization [27]. Therefore, this has motivated us to compare the association between SUD and Juggalo-specific experiences. For this purpose, we use the NMDI value of the `Juggalo_di` feature, which indicates whether a homeless youth considers themselves a Juggalo or not. Figure 4 compares the level of association between SUD and Juggalo experiences; the bigger the radius of the circle, the stronger is the association between SUD and `Juggalo_di` in that state. In this figure, stars show the venues of *Gathering Of The Juggalos* (GOTJ), which is the Juggalos' main annual festival [47]. According to the result, this association is far stronger in Missouri as compared to the other states. This is consistent with the geographical location of GOTJ venues, which are clustered around Midwestern states (where Missouri is located). In summary, previous studies have shown that gang-involved individuals are at a higher risk of substance use [34, 50]. As a result, taking a step towards reducing gang prevalence would be helpful for both alleviating SUD among homeless youth, and also, increasing safety in the society.

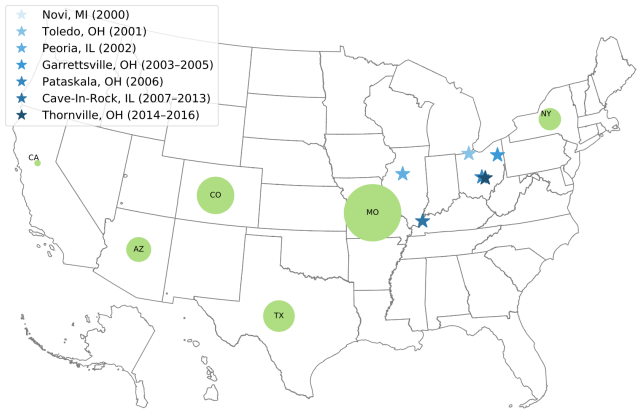


Figure 4: Comparison of the importance of Juggalo experiences across states (AZ=0.159, CA=0.037, CO=0.242, MO=0.373, TX=0.203, NY=0.142).

Importance of Gun-Related Victimization. Because of their specific lifestyle, homeless people are highly vulnerable to victimization [22]. Gun-related victimization refers to being assaulted by gun or witnessing another person assaulted by gun. We calculate the total importance of gun-related victimization in predicting SUD as the average of the normalized importance of features pertaining to gun-related victimization in our dataset. Specifically, we have the following features in our dataset:

- `vict_ass_gun`: In your lifetime, has anyone shot at you with a gun on purpose?
- `witness_gun_di`: In your lifetime, have you ever seen someone being injured or killed by a gun?
- `vict_ass_gun_inj`: In case someone has shot at you with a gun on purpose, have you ever been injured by that?

Each state of the United States has its own gun control legislation. These state laws are mostly intended to limit access to certain guns for certain individuals. Figure 5 shows the level of association between SUD and gun-related victimization as well as the weakness of gun control legislation in each state [6]. In this figure, the background color of each state shows the weakness of gun control law with dark red representing the weakest laws and yellow representing the strongest laws. Each circle shows the degree of association between gun related victimization and SUD for that specific state;

[3]https://isr.unm.edu/reports/2011/jail-based-substance-abuse-treatment-literature-review..pdf
[4]www.fbi.gov/stats-services/publications/2011-national-gang-threat-assessment

the bigger the radius of the circle, the stronger is the association. According to the results, the strength of the association is almost consistent with that of gun control laws in each state, except in Missouri. Thus, states with stronger gun control legislation see weaker association of gun related victimization with SUD, and vice versa (with the exception of Missouri). Intuitively, this is possible because if gun control legislation in a state are stringent, we would expect less gun-related violence, and consequently, less traumatic experiences (irrespective of whether someone is homeless or not). Therefore, gun-related victimization would have a lesser impact on people's lifestyle, and thus, we see a smaller association between SUD and gun-related victimization in those states with stronger gun control laws.

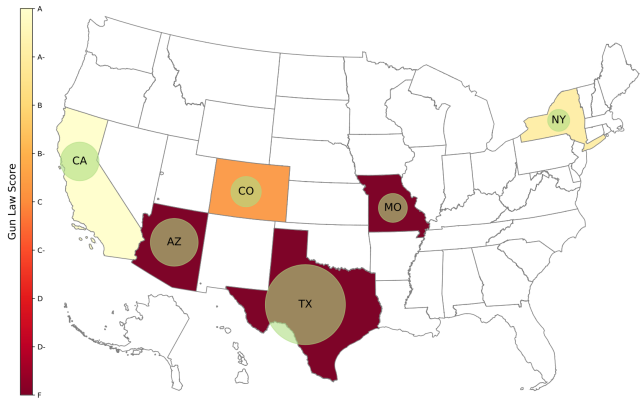


Figure 5: Comparison of the importance of gun-related victimization across states (AZ=0.448, CA=0.363, CO=0.285, MO=0.271, TX=0.751, NY=0.206).

7 LIMITATIONS

There are a few limitations of our study, many of which stem from the dataset that we use. The nature of the homeless population necessitates some decisions that limit the claims we can make with this research. The youth population surveyed for this study was not randomly selected which makes it more difficult to generalize our results to the entire population of homeless youth. Our data also relies on self-report measures, which have their own set of limitations. With self-report data, participants may not be completely honest when responding to the survey. Circumstances in this study make this more likely because the questions in the survey related to different conditions that have a stigma, making it possible that the individual would give a more socially acceptable answer instead of truth. As such, it is possible that conditions like SUD are under-reported in this dataset.

Our dataset is also cross-sectional, which has been collected from homeless youth at one specific time. As a result, we are not able to infer causal relationships among factors. Therefore, a future pathway would be collecting longitudinal data to follow the participants' conditions over time and inject them into a model, though this scheme may not be feasible with a transient population such as homeless youth.

8 CHALLENGES IN IMPLEMENTATION

In regard to future work, it is important to consider how the findings of this study can be applied to tackling substance use disorder in real world settings. In theory, our data-driven insights in both RQ2 and RQ3 can be used as weak guidelines by policymakers, as they formulate new state-level policies to tackle SUD. Yet, fully trusting the uncovered insights may be undesirable (as they correspond to correlative associations), unless we can validate these insights by finding causal associations. Thus, the next step of our work is to collect longitudinal data from homeless youth to be able to follow the participants' conditions over time. Such data can be used to infer causal associations, which can be used to validate the insights that have been uncovered in our work.

However, a few implementation challenges need to be solved before our longitudinal data collection procedure can be conducted with homeless youth. First, many homeless youth are highly suspicious due to suffering neglect/abuse over a long period of time. There is also a secondary issue about the protection of privacy for the involved youth. Unfortunately, most homeless youth drop-in centers (i.e., non-governmental organizations working with homeless youth) collect information about their youth, most of which is not to be shared with third parties, including researchers, etc. We propose encapsulating our data collection, pre-processing, prediction and feature importance analysis modules into a Google Chrome extension that the drop-in centers could use without providing identifying information to our team. Finally, public awareness campaigns in the drop-in centers working with this program would help overcome fears and suspicions held by homeless youth to encourage their participation.

9 CONCLUSION

This study takes an advantage of a real-world dataset to predict substance use disorder (SUD) among homeless youth. In addition, we analyze our predictive models to derive insights into the factors highly associated with SUD. For instance, we find that PTSD and depression are highly associated with SUD among homeless youth. Finally, we analyze variation in the associated factors with varying geographic locations and find that there is a great deal of location-specific variation in the factors associated with SUD. In future, we plan to collect longitudinal data to infer causal associations with SUD, which can be used by policymakers and practitioners to derive improved policies for tackling SUD.

ACKNOWLEDGMENTS

We thank Hsun-Ta Hsu, Robin Petering, Diane Santa Maria, Sarah Narendorf, Jama Shelton, Kimberly Bender, and Kristin Ferguson for sharing the dataset with us.

This work was in part supported by NSF awards #1742702, #1820609, #1915801, #1934782, and the High-Potential Individuals Global Training Program (2019-0-01590) by IITP and MSIT, Korea.

REFERENCES

[1] American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5).* American Psychiatric Association.

[2] Anamika Barman-Adhikari, Hsun-Ta Hsu, Daphne Brydon, Robin Petering, Diana Santa Maria, Sarah Narendorf, Jama Shelton, Kimberly Bender, and Kristin Ferguson. 2019. Prevalence and correlates of nonmedical use of prescription

drugs (NMUPD) among Young adults experiencing homelessness in seven cities across the United States. *Drug and Alcohol Dependence* 200 (2019), 153–160.

[3] Kimberly Bender, Samantha M. Brown, Sanna J. Thompson, Kristin M. Ferguson, and Lisa Langenderfer. 2015. Multiple Victimizations Before and After Leaving Home Associated With PTSD, Depression, and Substance Use Disorder Among Homeless Youth. *Child Maltreatment* 20, 2 (2015), 115–124.

[4] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. 1984. *Classification and Regression Trees*. CRC press.

[5] Nancy H. Busen and Joan C. Engebretson. 2008. Facilitating risk reduction among homeless and street-involved youth. *Journal of the American Academy of Nurse Practitioners* 20, 11 (2008), 567–575.

[6] Giffords Law Center. 2020. Annual Gun Law Scoreboard. https://lawcenter.giffords.org/scorecard/. [Online; accessed 2-February-2020].

[7] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2, 3, Article 27 (2011).

[8] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of artificial intelligence research* 16, 1 (2002), 321–357.

[9] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. Association for Computing Machinery, 785–794.

[10] Jordan P. Davis, Emily R. Dworkin, Jesse Helton, John Prindle, Sadiq Patel, Tara M. Dumas, and Sarah Miller. 2019. Extending poly-victimization theory: Differential effects of adolescents' experiences of victimization on substance use disorder diagnosis upon treatment entry. *Child Abuse & Neglect* 89 (2019), 165–177.

[11] Tracy L. Dietz. 2007. Predictors of reported current and lifetime substance abuse problems among a national sample of U.S. homeless. *Substance Use & Misuse* 42 (2007), 1745–1766.

[12] Tao Ding, Warren K Bickel, and Shimei Pan. 2017. Multi-view unsupervised user feature embedding for social media-based substance use prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2275–2284.

[13] Yoav Freund and Robert E Schapire. 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. System Sci.* 55, 1 (1997), 119–139.

[14] Robin M Hartinger-Saunders, Barbara Rittner, William Wieczorek, Thomas Nochajski, Christine M Rine, and John Welte. 2011. Victimization, psychological distress and subsequent offending among youth. *Children and Youth Services Review* 33, 11 (2011), 2375–2385.

[15] Saeed Hassanpour, Naofumi Tomita, Timothy Delise, Benjamin Crosier, and Lisa A. Marsch. 2019. Identifying substance use risk based on deep neural networks and Instagram social media data. *Neuropsychopharmacology* 44, 3 (2019), 487–494.

[16] Simon Haykin. 1998. *Neural Networks: A Comprehensive Foundation* (2nd ed.). Prentice Hall PTR.

[17] Laura M. Heath, Lise Laporte, Joel Paris, Kevin Hamdullahpur, and Kathryn J. Gill. 2018. Substance misuse is associated with increased psychiatric severity among treatment-seeking individuals with borderline personality disorders. *Journal of Personality Disorders* 32, 5 (2018), 694–708.

[18] Torsten Hothorn, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J. Van Der Laan. 2006. Survival ensembles. *Biostatistics* 7, 3 (2006), 355–373.

[19] Thomas M. Kelly and Denis C. Daley. 2013. Integrated Treatment of Substance Use and Psychiatric Disorders. *Social work in public health* 28 (2013), 388–406.

[20] Dean G Kilpatrick, Kenneth J Ruggiero, Ron Acierno, Benjamin E Saunders, Heidi S Resnick, and Connie L Best. 2003. Violence and risk of PTSD, major depression, substance abuse/dependence, and comorbidity: results from the National Survey of Adolescents. *Journal of consulting and clinical psychology* 71, 4 (2003), 692–700.

[21] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[22] Barrett A. Lee and Christopher J. Schreck. 2005. Danger on the Streets: Marginality and Victimization Among Homeless People. *American Behavioral Scientist* 48, 8 (2005), 1055–1081.

[23] Murdoch Leeies, Jina Pagura, Jitender Sareen, and James M Bolton. 2010. The use of alcohol and drugs to self-medicate symptoms of posttraumatic stress disorder. *Depression and anxiety* 27, 8 (2010), 731–736.

[24] Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. 2013. Understanding variable importances in forests of randomized trees. In *Advances in neural information processing systems*. 431–439.

[25] National Institute on Drug Abuse. 2019. Treatment Approaches for Drug Addiction. www.drugabuse.gov/publications/drugfacts/treatment-approaches-drug-addiction. [Online; accessed 2-February-2020].

[26] National Institute on Drug Abuse. 2020. Costs of Substance Abuse. www.drugabuse.gov/drug-topics/trends-statistics/costs-substance-abuse. [Online; accessed 14-June-2020].

[27] Robin Petering, Harmony Rhoades, Hailey Winetrobe, David Dent, and Eric Rice. 2017. Violence, Trauma, Mental Health, and Substance Use Among Homeless Youth Juggalos. 48, 4 (2017), 642–650.

[28] Christopher J. Przemieniecki, Samantha Compitello, and Josiah D. Lindquist. 2020. Juggalos - Whoop! Whoop! A family or a gang? A participant observation study on an FBI defined 'hybrid' gang. *Deviant Behavior* 41, 8 (2020), 977–990.

[29] Aida Rahmattalabi, Anamika Barman-Adhikari, Phebe Vayanos, Milind Tambe, Eric Rice, and Robin Baker. 2019. Social Network Based Substance Abuse Prevention via Network Modification (A Preliminary Study). *arXiv preprint arXiv:1902.00171* (2019).

[30] Aida Rahmattalabi, Phebe Vayanos, Anthony Fulginiti, Eric Rice, Bryan Wilder, Amulya Yadav, and Milind Tambe. 2019. Exploring algorithmic fairness in robust graph covering problems. In *Advances in Neural Information Processing Systems*. 15776–15787.

[31] Darrel A. Regier, Mary E. Farmer, Donald S. Rae, Ben Z. Locke, Samuel J. Keith, Lewis L. Judd, and Frederick K. Goodwin. 1990. Comorbidity of Mental Disorders With Alcohol and Other Drug Abuse: Results From the Epidemiologic Catchment Area (ECA) Study. *JAMA* 264, 19 (1990), 2511–2518.

[32] Michael W Ross and Mark L Williams. 2001. Sexual behavior and illicit drug use. *Annual review of sex research* 12, 1 (2001), 290–310.

[33] Stephen Ross and Eric Peselow. 2012. Co-Occurring Psychotic and Addictive Disorders. *Clinical neuropharmacology* 35 (2012), 235–43.

[34] Bill Sanders. 2012. Gang youth, substance use patterns, and drug normalization. *Journal of Youth Studies* 15, 8 (2012), 978–994.

[35] Julia Thornton Snider, Margaret E. Duncan, Mugdha R. Gore, Seth Seabury, Alison R. Silverstein, Mahlet G. Tebeka, and Dana P. Goldman. 2019. Association Between State Medicaid Eligibility Thresholds and Deaths Due to Substance Use Disorders. *JAMA Network Open* 2, 4 (2019).

[36] Richard Speiglman and Rex S. Green. 1999. Homeless and Nonhomeless Arrestees: Distinctions in Prevalence and in Socio Demographic, Drug Use, and Arrest Characteristics across DUF Sites. (1999).

[37] Daniel J. Stekhoven and Peter Buehlmann. 2012. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28 (2012), 112–118.

[38] Angela J Stewart, Mandy Steiman, Ana Mari Cauce, Bryan N Cochran, Les B Whitbeck, and Dan R Hoyt. 2004. Victimization and Posttraumatic Stress Disorder Among Homeless Adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry* 43, 3 (2004), 325–331.

[39] Substance Abuse and Mental Health Services Administration. 2018. *Key Substance Use and Mental Health Indicators in the United States: Results from the 2017 National Survey on Drug Use and Health*. Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration.

[40] Joel Swendsen, Kevin P. Conway, Louisa Degenhardt, Meyer Glantz, Robert Jin, and Kathleen R. Merikangas. 2010. Mental disorders as risk factors for substance use, abuse and dependence: results from the 10-year follow-up of the National Comorbidity Survey. *Addiction* 105 (2010), 1117–1128.

[41] The Sentencing Project. 2019. State-by-State Data. https://www.sentencingproject.org/the-facts/map. [Online; accessed 2-February-2020].

[42] Sanna J. Thompson, Kimberly Bender, Kristin M. Ferguson, and Yeonwoo Kim. 2015. Factors associated with substance use disorders among traumatized homeless youth. *Journal of social Work Practice in the Addictions* 15 (2015), 66–89.

[43] Kimberly Tyler, Lisa Kort-Butler, and Alexis Swendener. 2014. The Effect of Victimization, Mental Health, and Protective Factors on Crime and Illicit Drug Use Among Homeless Young Adults. *Violence and victims* 29, 2 (2014), 348–362.

[44] Kimberly A. Tyler and Katherine Johnson. 2006. Pathways in and out of substance use among homeless-emerging adults. *Journal of Adolescent Research* 21 (2006), 133–157.

[45] Kimberly A. Tyler and Lisa A. Melander. 2015. Child Abuse, Street Victimization, and Substance Use Among Homeless Young Adults. *Youth & Society* 47, 4 (2015), 502–519.

[46] Luis Villalobos-Gallegos, Maria Elena Medina-Mora, Corina Benjit, Silvia Ruiz-Velasco, Carlos Margis-Rodriguez, and Rodrigo Marin-Navarrete. 2019. Multidimensional patterns of sexual risk behavior and psychiatric disorders in men with substance use disorders. *Archives of Sexual Behavior* 48, 2 (2019), 599–607.

[47] Wikipedia contributors. 2020. Gathering of the Juggalos — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Gathering_of_the_Juggalos. [Online; accessed 2-February-2020].

[48] Amulya Yadav, Hau Chan, Albert Xin Jiang, Haifeng Xu, Eric Rice, and Milind Tambe. 2016. Using Social Networks to Aid Homeless Shelters: Dynamic Influence Maximization under Uncertainty. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. 740–748.

[49] Amulya Yadav, Bryan Wilder, Eric Rice, Robin Petering, Jaih Craddock, Amanda Yoshioka-Maxwell, Mary Hemler, Laura Onasch-Vera, Milind Tambe, and Darlene Woo. 2017. Influence Maximization in the Field: The Arduous Journey from Emerging to Deployed Application. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. 150–158.

[50] Kevin A. Yoder, Les B. Whitbeck, and Dan R. Hoyt. 2003. Gang Involvement and Membership among Homeless and Runaway Youth. *Youth & Society* 34, 4 (2003), 441–467.