Andrew Chen
Meghana Kalle
Group 1

Proactive Risk Assessment of Substance Abuse using Multi-Label and Multi-Class Machine Learning on Demographic and Psychological Data

Introduction

Substance abuse is a problem that has plagued society since ancient times, and dangerous drugs are more accessible now more than ever. In 2022, over 100,000 lives were claimed by drug overdoses in the U.S. alone (National Institute on Drug Abuse).

According to the 2023 United States National Survey on Drug Use and Health, 48.5 million, over 1 in 6 Americans over the age of 12 reported battling with a substance abuse disorder within the last year (American Addiction Centers). This highlights that substance abuse is not a fringe issue, but a crisis that affects millions of families, workplaces, and communities. Beyond the personal impact, substance abuse contributes to increased healthcare costs, loss of productivity, and increased crime rates. The opioid epidemic alone cost the United States $1.5 trillion in 2020, according to the U.S. government's Joint Economic Committee (JEC). As dangerous drugs become more accessible, the urgency to identify at-risk individuals steadily increases.

Despite the scale of this issue, current methods for identifying individuals at risk of substance abuse rely largely on self-reported assessments or intervention after symptoms of substance abuse disorder are already present. These traditional methods are inefficient because they are reactive and fall victim to human bias and prone to underreporting, meaning that at-risk individuals may remain hidden until it is too late. By leveraging machine learning tools, risk assessment tools can provide proactive insights rather than reactive responses, allowing healthcare professionals to respond quickly.

By creating a drug use risk assessment tool utilising the Drug Consumption Dataset from UCI, we hope to better understand the relationships between people and drugs, reducing the damages it can cause. This insight can support stronger intervention efforts and inform public health strategies.

Problem Statement

Identifying people who are likely to use specific drugs is a complex and often inaccurate process. Screening is the main method of identification which is difficult to administer and relies heavily on self-reporting. As a result, it is difficult to effectively intervene. To address this, we apply machine learning techniques to more reliably predict potential users of drugs based on demographic and psychological data.

Data Description

The data details information of 1885 participants. For each respondent, there are 12 known attributes which include participants' age; gender; educational level; country of origin; ethnicity; NEO-FFI-R personality scores to measure openness to experience, conscientiousness, extraversion/introversion, agreeableness and neuroticism; BIS-11 score to measure impulsiveness; ImpSS score to measure impulsive sensation seeking.

Respondents were also questioned regarding their use of 19 drugs (both legal and illegal) that participants have specified if they ever used the drug and if they have, then the last time they used that drug. These responses will serve as the target variables for this task where the goal is to predict an individual's risk level for each drug. The drugs included are: Alcohol, Amphetamines, Nitrite, Benzodiazepine, Caffeine, Marijuana, Chocolate, Cocaine, Crack, Ecstasy, Heroin, Ketamine, Legal Highs, LSD, Methadone, Magic Mushrooms, Nicotine, Semer*, Volatile Substances)

*: Semer, a fictitious drug was included to identify 'over-claimers'

Gender, education level, country of origin, ethnicity, and drug usage are categorical features.
Nscore, Escore, Oscore, AScore, Cscore, Impulsive, SS are continuous features.
This is a classification task.

Ratings for Drug Use:

CL0: Never Used
CL1: Used over a Decade Ago
CL2: Used in Last Decade
CL3: Used in Last Year
CL4: Used in Last Month
CL5: Used in Last Week
CL6: Used in Last Day

Additional processing steps include separating data into different columns according to their features and replacing the drug usage scale with integers. To do this we can use grep to replace “CL0” and “CL1” with 0 because we will assume those who have never used or used over a decade ago are fairly safe from substance abuse; “CL2” and “CL3” with 1 as used in last decade/year is more recent than CL0 and CL1, but not necessarily as vulnerable as those who have used within the last month; and “CL4”, “CL5”, and “CL6” with 2 as these are extremely recent. This would represent “not at risk”, “potentially at risk”, and “highly at risk” categories for users. Doing this will make the results more interpretable for a risk assessment tool.

We will use one-hot encoding to convert categorical data (education, country of origin, and ethnicity) to a numerical scale so that it is interpretable by the machine learning model. To do this we will get a list of the unique values using uniq, and create columns using awk.

There are no duplicate data points or missing values. (Tested using grep ‘,,’ and uniq -d)

Additionally, entries related to Semer will be excluded to maintain the integrity of the analysis. To do this we can use awk to remove rows where the column that corresponds to the fictitious drug are equal to 1.

Methodology

This project addresses a multi-label, multi-class classification problem where each individual is assessed for risk levels across 18 drugs. For each drug, the model will predict not at risk (0), potentially at risk (1), or highly at risk (2). We will use logistic regression, random forest, and multi-layer perceptron to do this.

Although the specific architecture of the model is not set, we can use k-fold validation to test various combinations of widths and depths as well as different activation functions (ReLU, Leaky ReLU, tanh) within the hidden layers. We will likely use ADAM as the optimization

function and the model will have 18 x 3 output neurons with a softmax activation function to determine whether an individual is not at risk, potentially at risk, or at risk for each drug. Our loss function will be categorical cross-entropy since we are using one-hot encoding.

For the logistic regression and random forest models, we will use Scikit-learn's MultiOutputClassifier to handle the multi-label structure.

Libraries:

- TensorFlow
- Keras
- Pandas
- Numpy
- SciKit-Learn
- Seaborn

Objectives

We aim to analyze data to find patterns in demographics of drug users to evaluate the risk of drug consumption. In addition, the model can be used to improve identification of at-risk individuals, allowing for more effective intervention strategies by healthcare officials.

Evaluation Metrics

- Recall: especially for dangerous substances, it is very important to correctly identify users (positives) of the drug
- Precision: important to have high precision so that predicted users are actually users (more important for less dangerous drugs)
- F1-score: involves both precision and recall
- Accuracy: defines how accurate our model is

Timeline

Data preprocessing: April 25

Model training: April 28

Evaluation: May 2

Documentation: May 4

Conclusion

This project aims to leverage machine learning to improve the identification of individuals at risk of drug use based on both demographic and psychological data found in the UCI Drug Consumption dataset. By creating a multi-label classification model, we hope to uncover patterns that can support intervention and prevention strategies.

References

https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates

https://www.kaggle.com/datasets/obeykhadija/drug-consumptions-uci

https://mmuratarat.github.io/2020-01-25/multilabel_classification_metrics

https://www.ncbi.nlm.nih.gov/books/NBK53214/

https://www.jec.senate.gov/public/index.cfm/democrats/2022/9/the-economic-toll-of-the-opioid-crisis-reached-nearly-1-5-trillion-in-2020

https://americanaddictioncenters.org/rehab-guide/addiction-statistics-demographics