SAFE DOSE

Team - Data Divers

Sanjana Kothari, Neetha Sherra, Devi Priya, Ananya Joshi, Ki Sung Park

1. Abstract

The degree of drug abuse has increased significantly over the past several years owing to the circumstances such as addiction, leisure, medical, and lack of knowledge pertaining to the usage of the drug. According to NIDA (The National Institute on Drug Abuse), a national research leader and information provider on substance use and addiction in the United States, many teenagers have exploited drugs at least once and recent events show an increased drug usage across all socio-economic groups. Drug abuse is life-threatening and additional research is required to identify the underlying root cause and how this misuse can be curtailed.

2. Problem Statement

This project aims to utilize data from drug related events to explore drug consumption patterns that lead to incidents like suicide, overmedication, etc. and further determine whether it is a case of drug abuse or not.

3. Datasets

DAWN is a nationally representative public health surveillance system that continuously monitors drug-related visits to hospital emergency departments (ED). This project uses data provided by DAWN (Drug Abuse and Warning Network) which collects data samples based on Stratified Simple Random Sampling of eligible hospitals representative of the entire nation.

The DAWN Drug Reference Vocabulary relies on Multum Lexicon, a drug vocabulary and classification system developed and maintained by Multum Information Services, Inc., a private sector firm. The dataset consists of 285 features which describe ED visit information, Sample design variables, Patient characteristics and Drug information for 229211 data points. The dataset is a mixture of discrete and continuous numeric features. Mappings of categorical variables to nominal variables have been provided by DAWN.

Drug Abuse Warning Network 2011 (DAWN-2011-DS0001) | SAMHDA (samhsa.gov)

4. Methodology

4.1. Exploratory Data Analysis

We will be using multiple python libraries for Visualizations and statistical analysis to understand the data.

4.2. Feature Engineering

Dataset contains numerous columns. With respect to answering this classification problem, we plan to assess and resolve

- Class imbalance : Handle class imbalance wrt drug abuse/not and the different case types to ensure unbiased model performance.
- Handle missing data: Impute missing/unavailable data
- Encodings : Categorical columns will be encoded for modeling purposes.
- Feature Selection : Feature selection for dimensionality reduction.

4.3. Binary Classification

To determine whether it is a drug abuse/misuse case or not using binary classifiers like Support Vector Machine (SVM), Decision Tree, etc. The outcome will be 0/1 representing not a misuse/is a misuse episode.

4.4. Multiclass Classification

To predict the type of case (suicide attempt/adverse reaction/overmedication, etc.) using multiclass classification techiniques like Naive Bayes and KNeighbors. The resultant output will be a label between 1 and 8 that represents the possible case type which can be used for controlling/monitoring similar cases in the future.

5. Evaluation Metrics

5.1. Binary

Our goal is to minimize False Negatives (drug abuse case but not identified) as well as minimize False Positives (not a drug abuse case but flagged as positive). Metrics - confusion matrix, ROC-AUC curve, F1 score

5.2. Clustering

If the initial analysis shows a demarcation between the cases, Silhouette score will be used to denote confidence in the classification of cases.

6. Deliverables

6.1. Web-App

- Stack - Django/React and Python
- Data Processing pipeline - Apache Beam
- Visualization - d3.js/plotly

6.2. Project Report

6.3. Git Link

CMPE255-SafeDose

7. Applications

- To identify emerging new drugs of abuse
- Assess alcohol use by minors that leads to ED visit
- Monitor trends and changing patterns in major substances of abuse
- Assess the potential effects of different components in prescription/OTC drugs
- Assist ongoing research on the reasons for increasing drug consumption