SafeDose

Project Report
Presented to
CMPE 255
Fall 2022

By
Team Data Divers

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

December 15, 2022

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.

'SafeDose' is a tool driven by data science algorithms to assist in the above-identified problem. By taking a data-driven approach, we aim to identify the different kinds of substance abuse occurring in the country as well as the overcoming the challenges of lacking medical documentation that leads to failure in identifying the cause of the drug misuse episode.

The solution is developed in accordance with the CRISP-DM methodology and the final outcome is presented as a web application.

Acknowledgments

We would like to express our gratitude and appreciation to our respected Professor Vijay Eranti, Computer Engineering Department, San Jose State University for his constant support and guidance throughout the semester.

Table of Contents

Chapter 1 Introduction

1.1 Project goals and objectives

This project aims to utilize data from drug-related events to explore drug consumption patterns that lead to incidents of suicide, overmedication, etc. While we agree that drug abuse is one of the top problems of our day and age it is also one that is difficult to identify and treat. With this project, we would like to improve upon the process of identifying the kinds of substance abuse and also address the issue of human bias and subjectivity using data science.

1.2 Problem and motivation

The motivation for this project stems from two major problems which are discussed in detail below

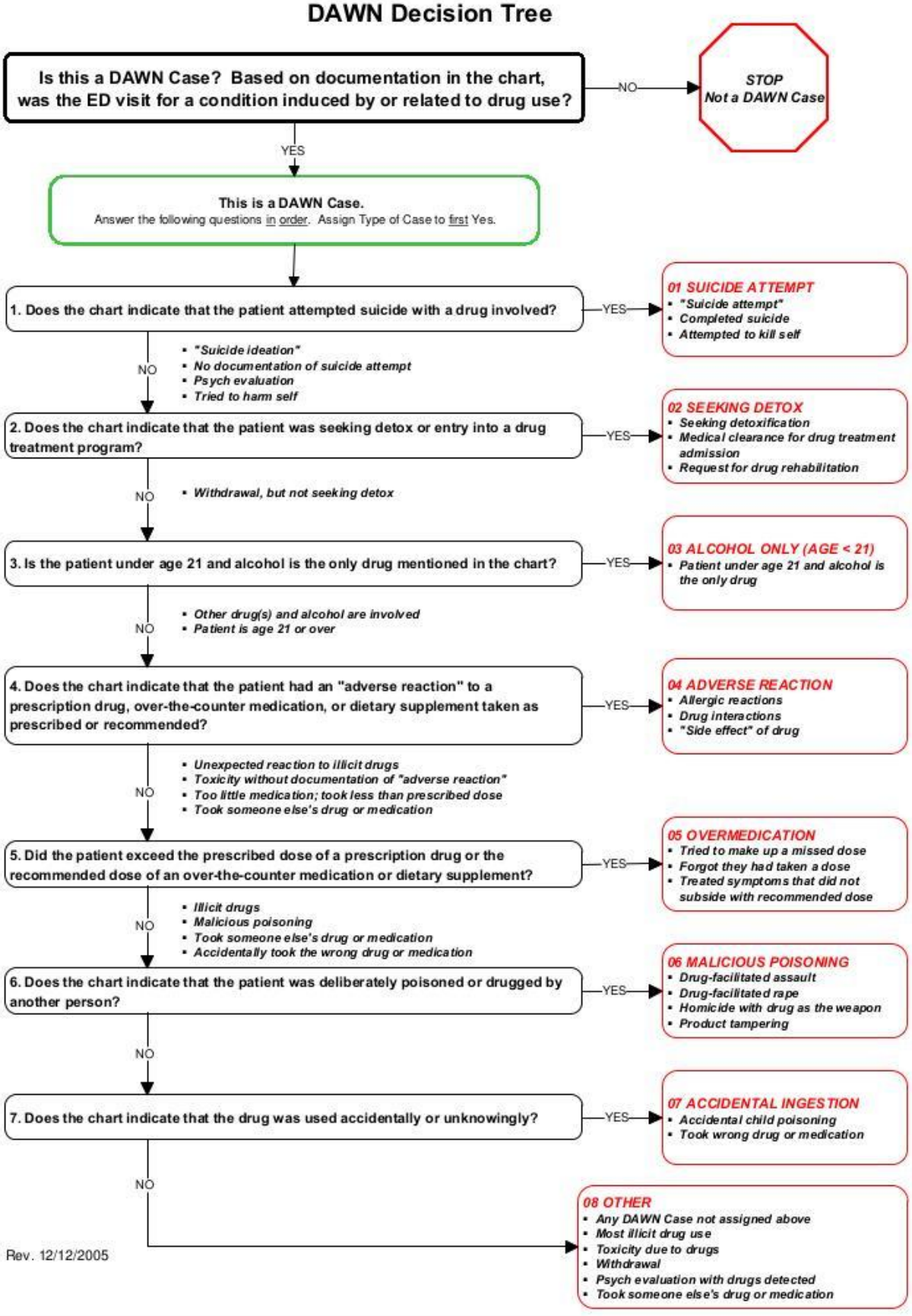
Problem A: Abuse indicator

When a hospital receives a potential drug abuse case, the case could have multiple abuse indicators such as Alcohol abuse, Illicit drugs (without alcohol) abuse, non-medical use of pharmaceuticals abuse, any pharmaceutical drug abuse, and all misuse and abuse. Given certain information on the case such as the demographic data of the patient involved, the drugs ingested, and the method of ingestion our model classifies a case into the type of abuse involved. Thus not only does our model output indicate widespread abuse of drugs but it also identifies the class of abuse.

Problem C: Case-type identification

During a visit to the Emergency Department, the patient fills out a case report. In the section of the report titled Type of Case the visit is required to be assigned to any one of the eight following case types Suicide attempt, Seeking detox, Alcohol only (Age < 21), Adverse reaction, Overmedication, Malicious poisoning, Accidental ingestion and Other. To avoid confusion, the ‘Dawn decision tree’ is used to assign an ED visit to the first applicable case type.

DAWN Decision Tree



Rev. 12/12/2005

Since ‘Other’ is reserved for cases that do not meet the requirements for the first seven types, most of the cases are assigned to the ‘Other’ type. Reasons for this include lack of documentation in medical records, insurance-related problems, and subjectivity or bias. Our classification model endeavors to eliminate subjectivity and other problems by classifying all the ‘Other’ cases into one of the more informative (seven) categories.

1.3 Project Application

- To identify new and emerging drugs of abuse. This could include broad medical categories of drugs, individual illicit drugs, alcohol, and pharmaceutical drugs.
- Assess incidents that involve the use of alcohol by minors and which result in Emergency Department (ED) visits.
- Address and minimize the problems caused by bias and human subjectivity thus eliminating the need for the ‘Dawn decision tree’ currently used to identify cases.
- Assist ongoing research on the reasons for increasing drug consumption.
- Assess the potential effects of different components in prescription/OTC drugs considering pharmaceutical drug abuse is also one of the identified types of abuse.

Chapter 2 CRISP-DM - BUSINESS UNDERSTANDING

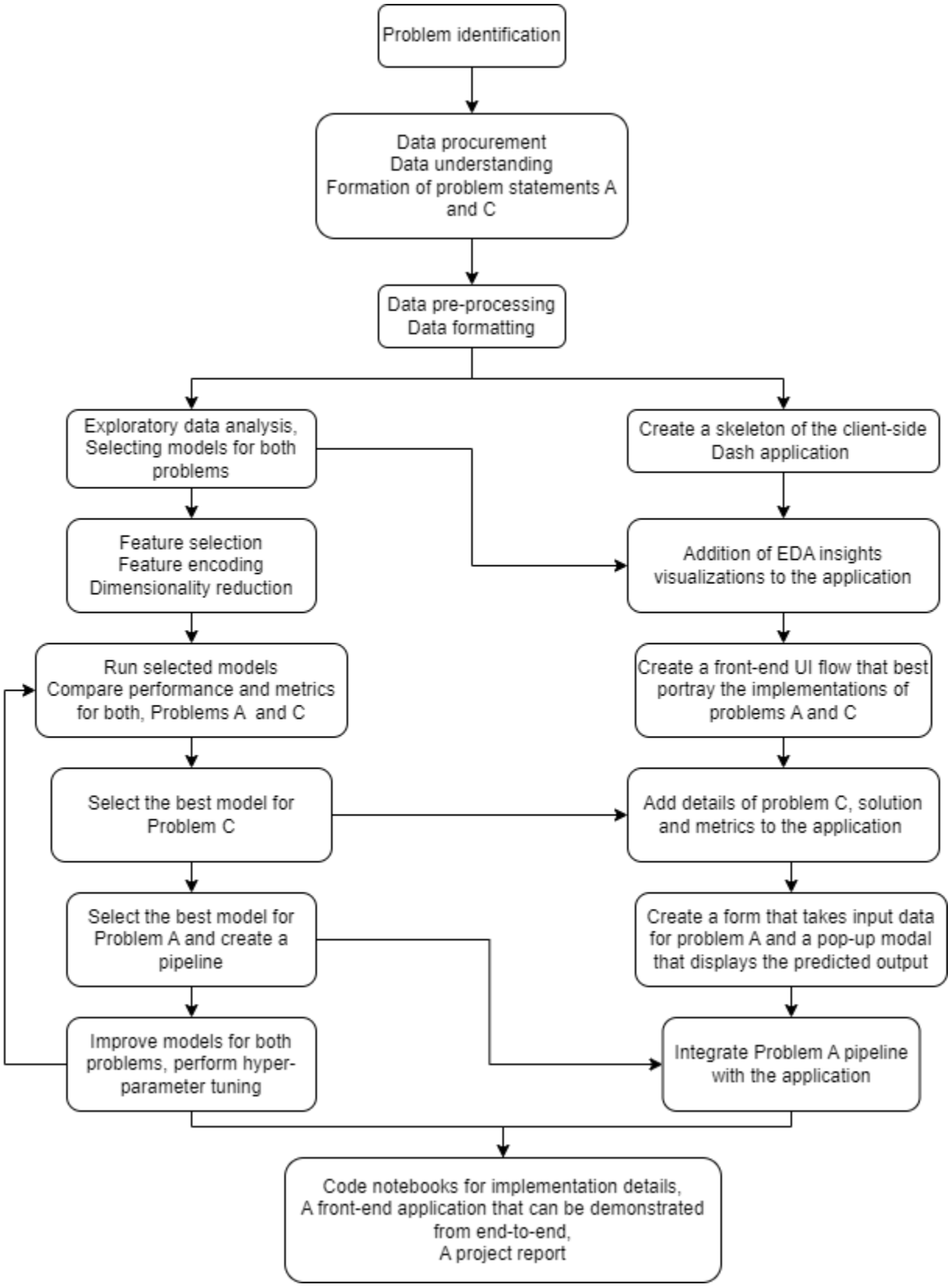
2.1 Business Objective

Providing statistics on drug usage by the public over a period of time and thereby utilizing data from drug-related events to explore drug consumption patterns that lead to a particular incident like suicide, overmedication, etc. This analysis helps to curb the misuse of drugs and other related events associated with them. In addition to that, it aids in creating and promoting wellness awareness among the public related to the drugs consumed for medical reasons. It also enhances further research activities pertaining to drug usage.

2.2 Data Mining Goals

Identifying and understanding the features related to the drug data and discovering patterns pertaining to drug abuse. The problems stated above are solved using classification techniques and the confidence in these classifications comes from performance metrics like precision and F1 score. Our goal is to achieve high performance in terms of the generalizing power of the model.

2.3 Project Plan



Chapter 3 CRISP-DM - DATA UNDERSTANDING

3.1 Data Collection

We have imported and extracted the data from the DAWN(Drug Abuse Warning and Network) website which is a nationwide public health surveillance system that captures data on Emergency Department visits related to substance abuse. The dataset represents medical health records of hospitals from different demographics in the United States of America pertaining to drug information and the case type related to it. We will use the DAWN datasets recorded in the year 2011 for our analysis and prediction of drug data.

3.2 Data Description

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. The dataset provides information for all types of drugs, including illegal drugs, prescription drugs, over-the-counter medications, dietary supplements, anesthetic gasses, substances that have psychoactive effects when inhaled, alcohol when used in combination with other drugs (all ages), and alcohol alone (only for patients aged 20 or younger). DAWN is the only data system providing estimates of the number of ED admissions associated with drug misuse and abuse and the particular drugs involved both for the United States as a whole and also for selected metropolitan areas.

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

1. Public-use dataset variables describe and categorize up to 16 drugs contributing to the ED visit, including toxicology confirmation and route of administration.
2. Administrative variables specify the type of case, case disposition, categorized episode time of day, and quarter of the year. Metropolitan areas are also included.
3. Created variables include the number of unique drugs reported and case-level indicators for alcohol, non-alcohol illicit substances, any pharmaceutical, non-medical use of pharmaceuticals, and all misuse and abuse of drugs.
4. Demographic items include age category, sex, and race/ethnicity.

3.3 Data Exploration

To understand the dataset better, we have done exploratory data analysis using pandas and plotly for visualizations.

We divided the drug dataset into two parts, demographic data that involved age categories, regions where the case was admitted, the race of the person involved in the case, gender, etc. Drug data involved all the drug ids and their subsequent catids and sdleds.

We mapped drug data on each of the demographic data to derive insights like

- How many cases of substance abuse occurred in each gender.
- Substance abuse cases collected in each metro region to understand which part of the US is likely to use our application depending on the number of cases it receives.
- If substance abuse cases have some correlation with different races
- There are 7 case types and we learned if a certain case type was majorly found in age categories.
- We found that the age below 21 which is regarded as substance abuse solely if there is alcohol involvement also had other drugs involved too, in the majority of the cases.
- We figured that there were 80000 plus cases that were not categorized in any of the case types and that became another problem statement that we solved by categorizing these cases into one of the other case types.
- We found the top 3 drugs that appear consistently in most of the substance abuse cases.
- We found that the drug mentioned was mostly always there in each case.
- We analyzed the number of substances involved for each drug through 1-22 mentions of it. However, they took the values in the range of (1,4). This led to a huge reduction in the number of rows where the number of substances was greater than 3.

All of the above data-related explorations helped us find the problems that we could solve, and figure out the indicators in the dataset that were useful to solve the problems we were looking at. It helped us approach the right methods in preparing our data for feature engineering.

Chapter 4 CRISP-DM - DATA PREPARATION

4.1 Data Selection

The exploratory data analysis revealed some interesting notions about the dataset.

1. The number of unique drugs reported in each case (NUMSUBS) takes values between 1 and 22. However, the range of values of this feature that lie between (Q1-1.5*IQR) and (Q3+1.5IQR) is only 1, 2, and 3. So, the dataset is reduced by removing rows that have NUMSUBS > 3. This reduces the dataset by 4.4% bringing it down from 229211 rows to 218950 rows.
2. The next step involves reducing the number of columns by dropping features that are not relevant to solving the problem such as features denoting the sampling process, the time-related features that have been randomized for public disclosure protection, etc.

4.2 Data Cleansing

The dataset contains negative values that represent the following:
-7: Not applicable
-8: Not documented
-9: Missing
All these are replaced by 0 as these values cannot be imputed or estimated by interpolation. Substituting them as 0 tells the classification model to treat them as one category of variables.

4.3 Data Transformation

To solve the 2 problems at hand, we create two datasets. **Dataset A** denotes the dataset used to determine the type of abuse that has occurred, and **Dataset C** is used for determining the type of case a record can represent from the 'Other' case types bucket.
A two-step process is followed to transform the data into the final datasets that are then input to the classification models.

1. Hashencoding - Hashencoding is a process of converting categorical features with very high cardinality into numerical features. Hash encoders hash every value in the feature column and the hash value determines the bucket that the value falls into. By taking 7 buckets, we encode all drug-related columns such that every feature is expanded into 7 columns with binary data.

2. For demographic and some drug-related columns, one-hot encoding is followed as the number of categories is small. One hot encoding ensures no information is lost, unlike hash encoding where information loss occurs due to hashing collisions. However, hash encoding offers a compressed encoding that is computationally efficient.
3. For dataset A, the encodings are then combined and provided to Principal Component Analysis. PCA finds the principal components that explain 80% variance in the data. The resultant dataset has reduced dimensions with minimal information loss which is then input to the model for multilabel classification of the type of abuse.
4. For dataset C, the encoded data is passed through MCA (Multiple Correspondence Analysis) which is similar to PCA but is specific to categorical data. This too reduces the dimensionality of the dataset.
5. For dataset C, the final step involves performing SMOTE (Synthetic Minority Oversampling Technique) due to the class imbalance of CASETYPE. Performing oversampling ensures that they are nearly an equal number of records for every CASETYPE which prevents bias towards the majority class and reduces the possibility of poor classification of the minority classes. Below is the distribution of classes in the original dataset.

1	SUICICDE ATTEMPT:(1)	9033	3.9 %
2	SEEKING DETOX:(2)	14841	6.5 %
3	ALCOHOL ONLY (AGE < 21):(3)	7421	3.2 %
4	ADVERSE REACTION:(4)	88096	38.4 %
5	OVERMEDICATION:(5)	18146	7.9 %
6	MALICIOUS POISONING:(6)	793	0.3 %
7	ACCIDENTAL INGESTION:(7)	3253	1.4 %
8	OTHER:(8)	87628	38.2 %

Chapter 5 CRISP-DM - MODELING

5.1 Modeling Techniques

Problem A, when predicting the types of abuse that a case record falls into, is a multilabel classification problem. A case can show more than one type of drug abuse from among the five types of abuse:

- ALLABSUE: all misuse and abuse indicator
- ALCOHOL: visit includes alcohol mention indicator
- PHARMA: pharmaceuticals indicator
- NONMEDPHARMA: non-medical use of pharmaceuticals indicator
- NONALCILL: non-alcohol illicit drugs indicator

Some algorithms for multilabel classification are:

- K Neighbors Classifier
- Random Forest Classifier

For Problem C that involves the predicting the type of visit, i.e. what incident of drug consumption led to this episode, is a multiclass classification problem of classifying the case types under the ‘other’ bucket into one of the possible categories using algorithms. The categories include:

- 'CASETYPE_1' : 'Suicide Attempt',
- 'CASETYPE_2' : 'Seeking Detox',
- 'CASETYPE_3' : 'Alcohol consumed below 21 years,
- 'CASETYPE_4' : 'Adverse Reaction',
- 'CASETYPE_5' : 'Overmedication',
- 'CASETYPE_6' : 'Malicious Poisoning',
- 'CASETYPE_7': 'Accidental Ingestion'

Some algorithms for multiclass classification are:

- Naive Bayes
- Logistic Regression
- Light Gradient Boosting Machine
- Random Forest Classifier
- K Neighbors Classifier

5.2 Generating Test Design

The dataset for Problem A is split into training and test set using train_test_split where test_size=0.20. The classifier is trained on the training dataset and evaluated on the test set.

The dataset for Problem C is split into three sets - training, testing, and validation. The training set is all the records where CASETYPE != 8 (Others). The training set is then split into train and validation sets where test_size is 0.2. The validation set is used to test the performance of the classifier on unseen data since we do not have the CASETYPE labels in the test set. The test set is the set of records where CASETYPE == 8 (Others).

5.3 Model Development

The multilabel classification problem of identifying the substances that have been abused is done using a RandomForestClassifier followed by MultiOutputClassifier.

```
forest=RandomForestClassifier(n_estimators=100,
min_samples_leaf=1, min_samples_split=2, random_state=11)
multi_target_forest=MultiOutputClassifier(forest, n_jobs=10)
```

For determining the type of case, several models have been implemented and the best performing model is chosen for final predictions on the test set.

- Naive Bayes

  ```
  nb=GaussianNB()
  nb.fit(X_train, y_train)
  ```

- Logistic Regression

  ```
  lr=LogisticRegression()
  lr.fit(X_train, y_train)
  ```

- K Neighbors Classifier

  ```
  knn=KNeighborsClassifier(n_neighbors=4)
  knn.fit(X_train, y_train)
  ```

- Light GBM

```
gbc=LGBMClassifier(n_estimators=100,learning_rate=0.01,
max_depth=5, random_state=11)
gbc.fit(X_train, y_train.values.ravel())
```

- Random Forest Classifier

```
forest=RandomForestClassifier(n_estimators=50,
min_samples_leaf=1, min_samples_split=2, random_state=1)
multi_target_forest=MultiOutputClassifier(forest,n_jobs=2)
multi_target_forest.fit(X_train, y_train)
```

5.4 Model Assessment

For problem A two metrics are evaluated:

- F1 score = (2*Precision*Recall)/(Precision+Recall)
  F1 score achieved with RandomForestClassifier = 99.9

  Precision = TP/(TP+FP)
  Out of all predicted true for each ABUSETYPE, how many are actually true for each ABUSETYPE? This is required as it helps to determine what substances are actually being abused and what is the pattern of consumption of such drugs so that measures can be taken to prevent their abuse/ misuse.

  Recall = TP/(TP+FN)
  Out of all the actual positives for an ABUSETYPE, how many are correctly identified as positive? Flagging incorrect cases as positive might result in incorrect decisions regarding treatment, policy implementation, etc.

  Since both precision and recall are crucial, the F1 score is considered as it takes the HM of Precision and Recall.

- Accuracy = (TP+TN)/(TP+FP+FN+TN)
  Accuracy obtained with RandomForestClassifer = 99.9

For problem C, several models are tried on the validation set and the best performing model is chosen. In predicting CASETYPES, precision is an important metric as the type of case would determine the next steps of treatment for the patient. The precision score tells us out of the predicted true/positive cases, how many were actually true/ positive for that particular case type. Recall should also be considered as it is important that out of the actual positive case of each case type, how many were correctly identified, which is given by recall. However, precision is more significant in our case, and hence precision is the primary metric and F1 score is the secondary metric.

	precision	recall	f1-score	support
Accidental Ingestion	0.98	0.98	0.98	17156
Adverse Reaction	0.95	0.90	0.92	17156
Alcohol consumed below 21 years	1.00	1.00	1.00	17155
Malicious Poisoning	0.99	1.00	1.00	17155
Overmedication	1.00	0.99	1.00	17156
Seeking Detox	0.96	0.96	0.96	17155
Suicide Attempt	0.91	0.95	0.93	17155
accuracy			0.97	120088
macro avg	0.97	0.97	0.97	120088
weighted avg	0.97	0.97	0.97	120088

Classification report on the validation set

Chapter 6 CRISP-DM - EVALUATION

To determine the types of substances that have been abused by an individual visiting the ED, the hospital needs to perform a toxicology test to evaluate possible accidental or intentional overdose or poisoning. It is also used to determine the presence of substances in the body for medical or legal purposes. Using data-driven algorithms, the 'SafeDose' platform can help determine the types of abuse without requiring a toxicology test. By simply entering demographic and medical information gathered from the patient, the abuse types can be determined. This can help speed up the process in the hospitals and also help to identify and handle unknown cases.
The results from the model predictions on the test set provide confidence in the ability of the algorithm to give out the correct ABUSETYPES.

As for the classification of 'Others' into one of the 7 CASETYPES, the lack of documentation of substance abuse in medical records makes it difficult to determine the cause of the event like adverse reactions, malicious poisoning, consumption of illicit drugs, etc. Thus, the element of ambiguity in such cases can be eliminated by having an algorithm identify the CASETYPE by learning from the historical patterns of each of the CASETYPES. This can also help in reducing human bias (for any reason) in claiming an episode to be a case of drug abuse.

Chapter 7 CRISP-DM - DEPLOYMENT

We are using a Dash application in order to demonstrate the working of our project. Dash is a framework that is built on top of plotly.js, react.js, and flask in Python.

Our Dash app *SafeDose* is rendered in the browser and consists of multiple web pages:

- Dashboards related to our Exploratory Data Analysis phase segregated into Demographic and Drug related findings.
- A form that is intended to mimic data collection during an ED visit. The data in this form serves as input to our 'Abuse' indicator classification model which then performs multi-label classification. The results of our model which indicate the type of abuse involved are displayed as a pop-up dialog box.
  Since the input from the form is real-time, it is passed through a data transformation pipeline where the encodings and dimensionality reduction objects (stored as pickles) trained on the training set are applied to it. The pickled model is then used to predict the type of substances abused.
- A page that walks the user through our motivations and solution to the problem of human subjectivity explained in section 1.2.