

RISE OF DRUG ABUSE AMONG YOUNG PEOPLE

1:PROBLEM DEFINITION

This case study examines the likelihood of a young person abusing drugs , the causes and consequences of the same

Objectives

- ✓ Identify and analyze the root causes: The core objective is to move beyond surface-level observations and systematically uncover the interconnected psychological, social, economic, and environmental factors driving drug abuse among youth. This involves distinguishing between correlation and causation.
- ✓ Understand the lived experience of young people: This helps to understand the "why" from their perspective, rather than just imposing an external viewpoint.
- ✓ Evaluate the effectiveness of current interventions: It seeks to assess existing strategies
- ✓ to determine what is working, what is not, and where the critical gaps in the system lie.

Stakeholders

- ✓ Young people: They are both the primary actors and the most affected group. Their well-being, health, and future potential are directly at stake.
- ✓ Parents, guardians , family and friends: They are the first line of defense and support. They are deeply affected emotionally and often financially by a child's substance abuse

Key performance metrics

Reduction in First-Time Drug Use Among 12-17 Year-Olds.

- ✓ **Description:** This metric measures the effectiveness of **prevention efforts**. It tracks the percentage of adolescents in a specific region who report using an illicit drug for the first time within the past year.
- ✓ **Why it's a good KPI:** Preventing initiation is the most powerful way to curb the pipeline of addiction. A decline in this number indicates that education, community support, and mental health initiatives are successfully protecting the most vulnerable age group before habits form.

2:DATA COLLECTION AND PRE PROCESSING

The following are links to the datasets for the study;

- ✓ <https://www.kaggle.com/datasets/sheemazain/students-drugs-addiction-dataset-2024>

- ✓ <https://www.kaggle.com/datasets/bgallamoza/national-survey-of-drug-use-and-health-20152019>

The potential bias in these datasets would be reporting bias as respondents used during data collection may lie to avoid being judged or face legal action.

Data preprocessing steps

- 1.handle missing values by removing rows or columns if missing data is minimal and random and removing duplicate entries
- 2.encoding categorical data by converting non-numerical data into numerical format that machine learning algorithms can understand
- 3.normalization, standardizes the range of numerical features so that no single feature dominates the model simply because of its scale.

3:MODEL DEVELOPMENT

Model choice: random forest

It provides feature importance scores, which are crucial for this problem. Knowing that "history of depression," "peer substance use," and "low parental supervision" are the top predictors is far more valuable for designing interventions than a black-box prediction. In addition to that, **random forest** excels with the type of data we have: structured, tabular data from surveys It often achieves the best performance on such datasets without extensive tuning.

Splitting of data;

Given the datasets I have provided above,

- I. Sort the data by year.
- II. Split the data into training, validation and test set
- III. Training set(e.g 2015-2017)
- IV. Validation set(2018-2019)
- V. Test set(2024)

This simulates a real-world deployment where the model is trained on past data to predict future cases. It tests the model's ability to handle temporal drift

Hyperparameters;

1.n_estimators-The number of decision trees in the forest

Too few trees: The model may be unstable and have high variance, leading to poor performance (underfitting).

Too many trees: Increases computational cost with diminishing returns and, in extreme cases, can lead to overfitting if the trees are too deep.

2.max_depth-The maximum allowed depth for each individual tree.

This is the primary lever to control overfitting vs. underfitting.

Shallow trees (low max_depth): Simple models that might not capture all the complex patterns in the data (high bias, underfitting).

Deep trees (high max_depth): Complex models that can learn the training data perfectly, including its noise. This causes the model to fail on new data

4:EVALUATION AND DEPLOYMENT

Metric 1: F1-Score

The harmonic mean of Precision and Recall. It provides a single score that balances two competing concerns.

Relevance:

Precision: Of all the kids we *flag* as at-risk, what percentage actually are? **High precision means we minimize false alarms.** This is crucial for building trust with counselors and ensuring they don't waste limited resources on false leads.

Recall: Of all the kids who *are actually* at-risk, what percentage did we successfully flag? **High recall means we miss very few at-risk youth.** This is our

primary ethical imperative—failing to identify a child in need (a False Negative) is the worst outcome

Metric 2: Area Under the Precision-Recall Curve (AUPRC)

A plot of precision (y-axis) against recall (x-axis) for different probability thresholds, and the area under this curve.

Relevance: This is **the most important metric for imbalanced datasets**. In our case, the number of non-users (negative class) will vastly outweigh the number of users (positive class). The AUPRC focuses almost entirely on the model's performance in correctly identifying the rare, positive cases.

Concept drift occurs when the statistical properties of the **target variable** (what we are trying to predict) change over time in the real world, making the model's predictions less accurate

Since we often cannot get immediate, real-world labels we rely on **drift detection in the input data** by continuously monitoring the statistical distribution of the *input features* to the live model and compare them to the training data distribution. Also, Monitor the distribution of the model's *output scores* (the risk probabilities).

Technical challenge;

Data pipeline and integration. The model cannot exist in a vacuum. To be useful, it must be integrated into the existing IT systems of a school or health department

This involves:

Building a robust, automated data pipeline to securely pull student data from various siloed sources (student information systems, survey platforms, counseling notes) and transform it into the features the model expects.

Ensuring strict data security and privacy (FERPA, HIPAA compliance) as highly sensitive student data moves through this pipeline.

Handling missing data and schema changes in real-time. For example, if a new version of a school survey removes a question that was a key feature for the model, the pipeline must not break, and the model's performance must be re-evaluated.