Project Overview

Problem Statement: Predict heroin consumption based on drug usage patterns.

Proposed Solution: Apply machine learning models to analyze substance use and demographic data.

Potential Impact: Inform targeted interventions and public health strategies..

Dataset Overview

The dataset includes demographic information and usage patterns for various substances.

Preprocessing Procedures:

- **Data cleaning:** Handling missing values and outliers.
- **Feature engineering:** Encoding categorical variables and normalizing numerical features.

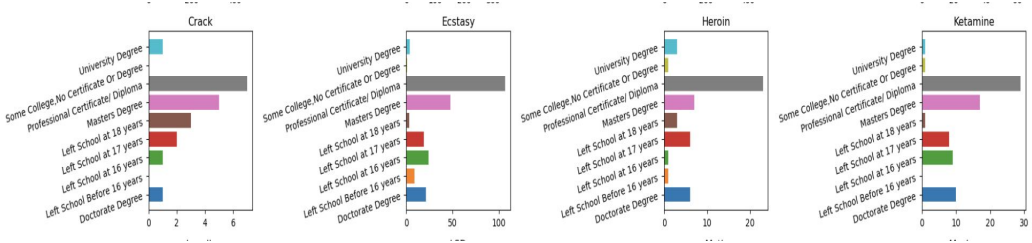
Features:

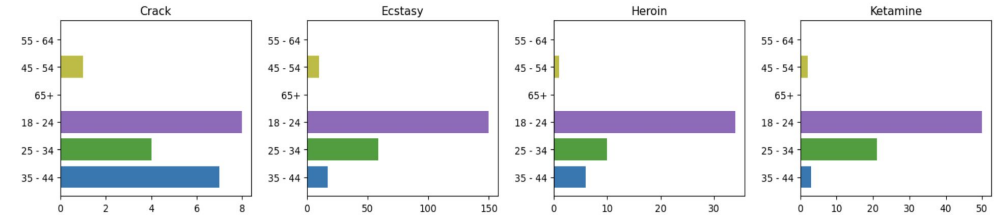
- Age Group
- Country
- Personality types
- Different classes of Drugs

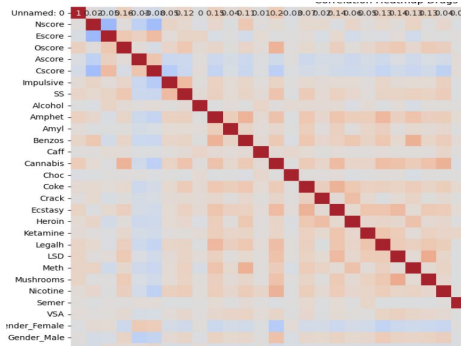
Target Variable:

- Heroin Usage

Important EDA findings



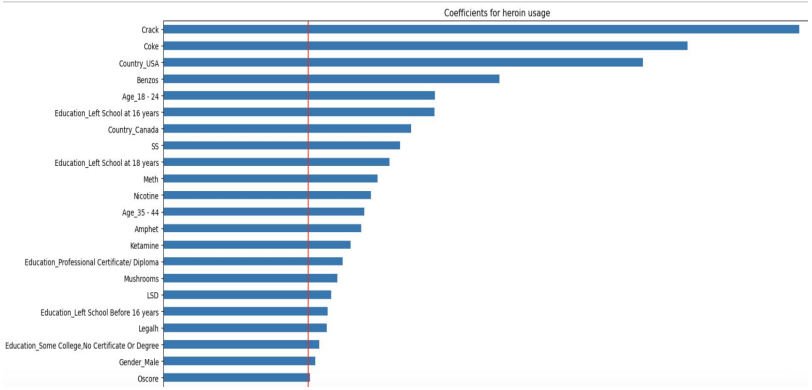




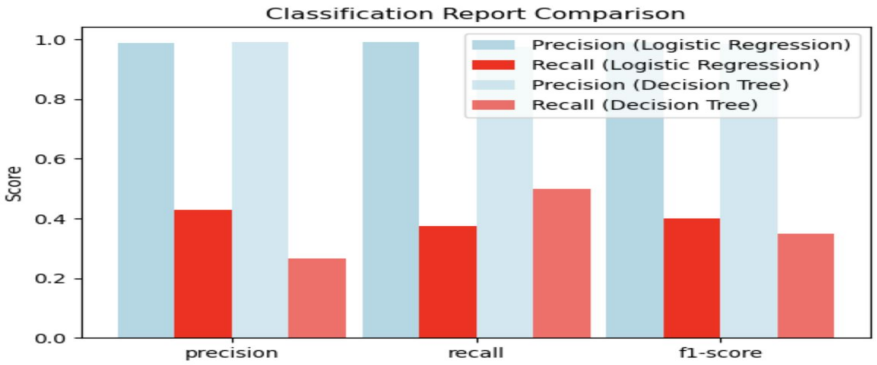
Correlation between Drug Use: Strong correlations between heroin use and the substances like crack, cocaine, and benzodiazepines.

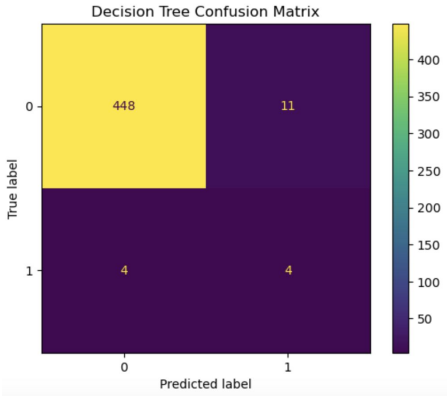
Demographic Influence: younger age groups (18-24) and certain education levels (e.g., professional certificates).

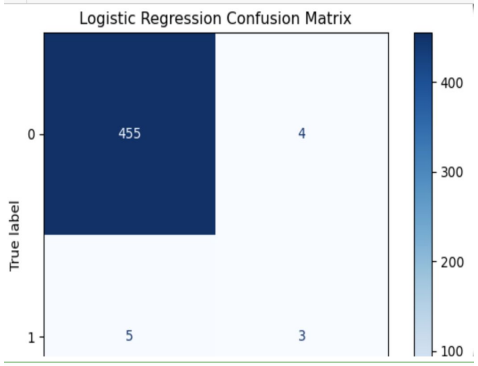
Geographical Trends: With higher prevalence in countries like the USA and Canada compared to others.



Baseline models and evaluation metrics







Logistic Regression: Achieves higher precision (99%) for non-heroin users (class 0) compared to the Decision Tree.

Decision Tree: Shows better recall (42%) for heroin users (class 1) than the Logistic Regression model.

Overall: Logistic Regression outperforms the Decision Tree with higher accuracy (98% vs. 96%) in predicting heroin usage.

Accuracy of the models are good but Future work could involve further tuning these models or exploring more complex algorithms to improve the identification of heroin users.

Next steps

- **Feature Engineering**
- **Ensemble Methods**: Implement ensemble models like Random Forest, Gradient Boosting
- **Hyperparameter Tuning**
- **Cross-Validation**
- **Feature Selection**