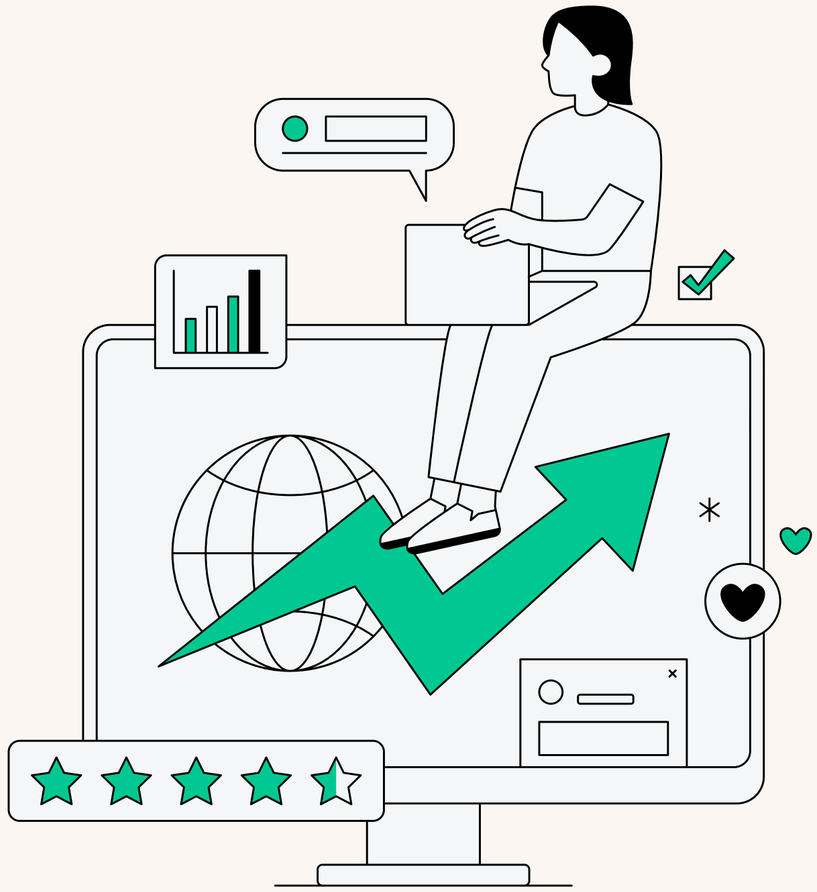
Presented by Alex, Olsi

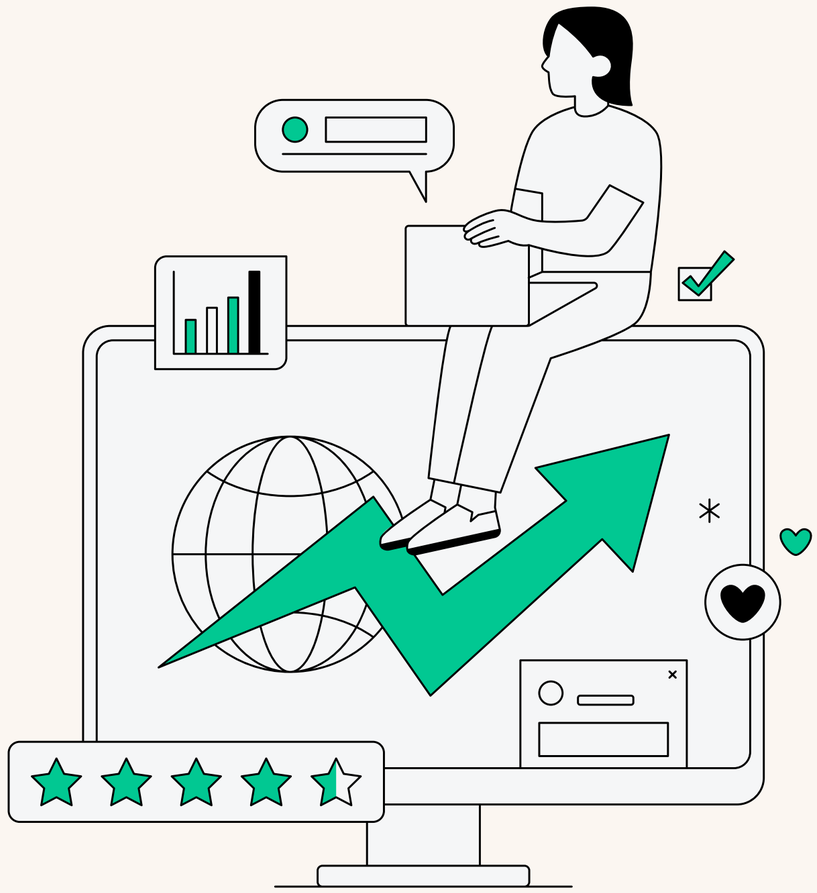
Welcome to Our Presentation

Building a machine learning module



Predicting Drug Consumption: A Machine Learning Approach

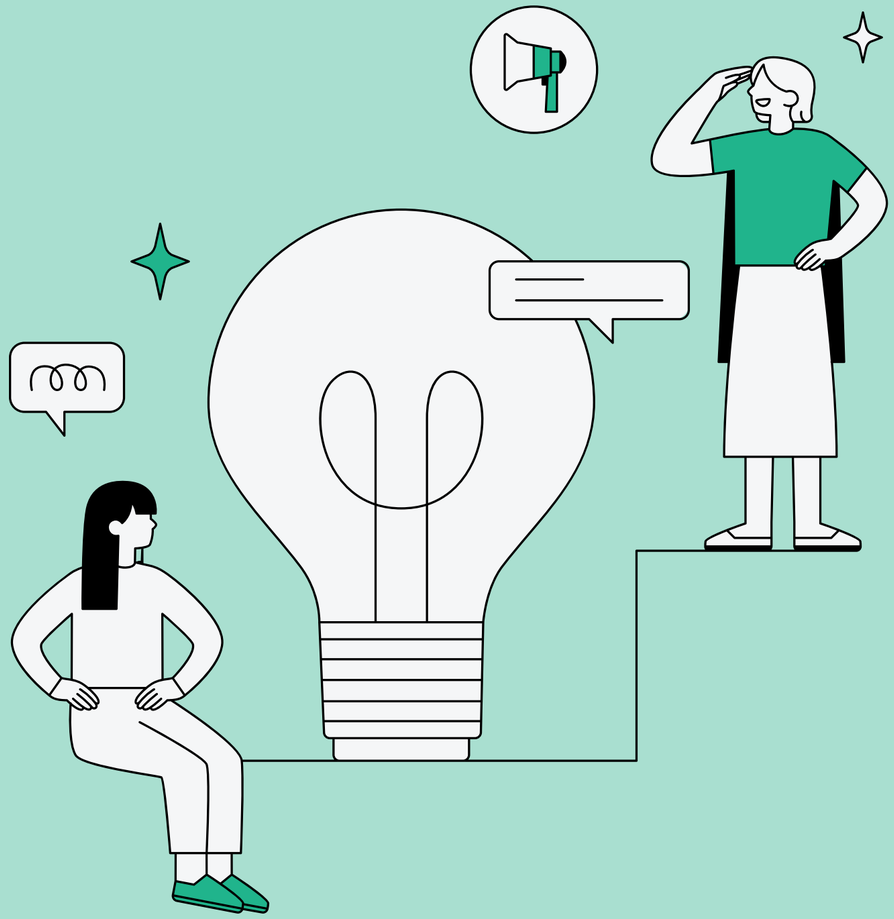
https://www.kaggle.com/datasets/mexwell/drug-consumption-classification



Introduction :

We are tackling the issue of predicting drug consumption per demographic using machine learning.

Accurately predicting drug use is essential for targeted interventions and preventive measures.



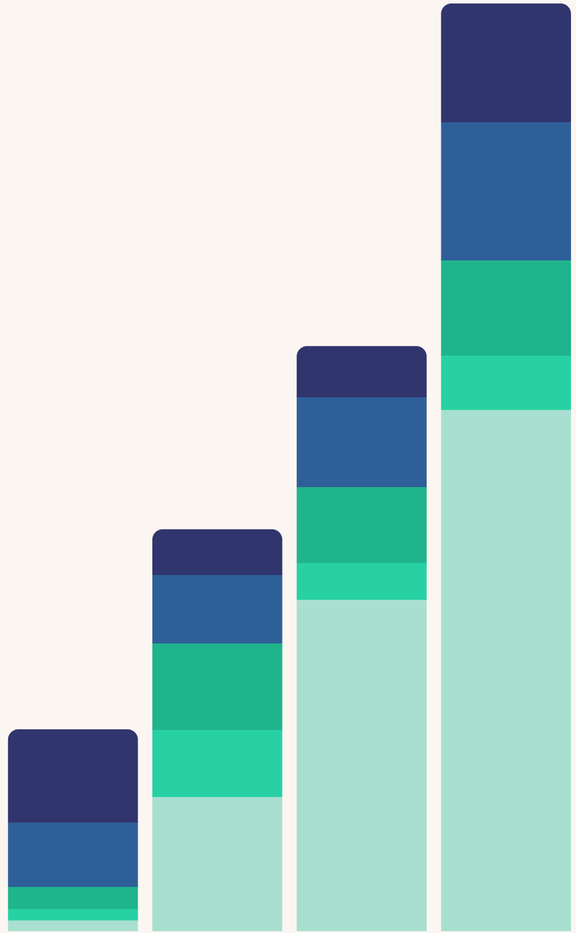
Data Selection:

Drug Consumption Classification

database from kaggle

shape : 1885 rows

32 columns



Data Cleaning

- Dropping all the rows with non-zero “Semer” values (fictitious drug meant to catch over-reporters)
- Grouping drugs by class for easier analysis
- Unstandardizing the data

Feature Engineering

1

Normalization

2

Label encoding
/
Hotcoding

3

Data
condensation

Model Building

01. Random Forest

02. KNN

03. Tuning

What we encountered in our machine learning model, was the complexity of predicting not just one, but a multitude of target variables, which posed difficulties in achieving higher accuracy.

Our approach involved grouping the targets and attempting to make predictions collectively, which added complexity to the task.

Our Average Accuracy is 0.57 between the groups because some target groups had the accuracy very slow. After we ungruped data we had an accuracy 0.63 (with RandomForestClassifier)

The Results:

For raw grouped data:

Output Accuracies: [0.8617021276595744, 0.4308510638297872, 0.7712765957446809, 0.43351063829787234, 0.3803191489361702, 0.4973404255319149, 0.5425531914893617, 0.8324468085106383, 0.44946808510638 3, 0.5611702127659575]

Corresponding to:

Soft Prescription Stims HardStims Downers HardDowners Psychs Legal Illegal Inhalants

Average Accuracy: 0.576063829787234

The Results:

For normalized grouped data:

Output Accuracies: [0.8643617021276596, 0.4148936170212766, 0.776595744680851, 0.4468085106382978, 0.35106382978723405, 0.5, 0.5159574468085106, 0.8351063829787234, 0.4308510638297872, 0.5664893617021277]

Corresponding to:

Soft Prescription Stims HardStims Downers HardDowners Psychs Legal Illegal Inhalants

Average Accuracy: 0.5702127659574467

Hyperparameter Tuning:

(for grouped data)

Best hyperparameters found:

Max estimator depth: 10

Minimum samples per leaf: 2

Minimum samples per split data: 10

Number of estimators: 100

Average Test Accuracy: 0.5819148936170213

The Results:

For raw ungrouped data:

Output Accuracies: [0.3696808510638298, 0.5425531914893617, 0.6835106382978723, 0.5398936170212766, 0.7553191489361702, 0.4069148936170212 7, 0.598404255319149, 0.8617021276595744, 0.5877659574468085, 0.8696808510638298, 0.8138297872340425, 0.648936170212766, 0.6382978723404256, 0.7686170212765957, 0.5585106382978723, 0.38563829787234044, 0.773936170212766]

Corresponding to:

Alcohol, Amphet, Amyl, Benzos, Caff, Cannabis, Coke, Crack, Ecstasy, Heroin, Ketamine, Legalh, LSD, Meth, Mushrooms, Nicotine, VSA

Average Accuracy: 0.6354818523153941

The Results:

For normalized ungrouped data:
Output Accuracies: [0.3856382978723404, 0.5531914893617021, 0.6941489361702128, 0.5664893617021277, 0.7659574468085106, 0.4015957446808511, 0.6090425531914894, 0.8696808510638298, 0.5797872340425532, 0.8643617021276596, 0.8164893617021277, 0.6303191489361702, 0.6409574468085106, 0.7632978723404256, 0.5425531914893617, 0.9361702127659576, 0.776595744680851]

Corresponding to:
Alcohol, Amphet, Amyl, Benzos, Caff, Cannabis, Coke, Crack, Ecstasy, Heroin, Ketamine, Legalh, LSD, Meth, Mushrooms, Nicotine, VSA
Average Accuracy: 0.6384543178973716

Hyperparameter Tuning:

(for ungrouped data)

Best hyperparameters found:

Max estimator depth: None

Minimum samples per leaf: 4

Minimum samples per split data: 2

Number of estimators: 100

Average Test Accuracy: 0.64

Random Tree Findings:

Works best for ungrouped data

Struggles to accurately predict most substances

Does a good job with:

Caffeine (0.76)

Crack (0.87)

Heroin (0.86)

Ketamine (0.82)

Meth (0.76)

VSA [Inhalants] (0.78)

DecisionTreeClassifier:

For raw ungrouped data:
Output Accuracies: [0.4300847457627119, 0.6440677966101694,0.7944915254237288
, 0.7436440677966102, 0.875, 0.5529661016949152, 0.8072033898305084
, 0.68855932033898 4, 0.579787234042553 2, 0.8643617021276596, 0.8164893617021277,
0.6303191489361702, 0.6610169491525424, 0.7632978723404256, 0.82553191489361 7,
0.9361702127659576, 0.5529661016949152
]

Corresponding to:
Alcohol, Amphet, Amyl, Benzos, Caff, Cannabis, Coke, Crack, Ecstasy, Heroin, Ketamine, Legalh, LSD, Meth, Mushrooms, Nicotine, VSA
Average Accuracy: 0.6384543178973716

KNN, split up individually:

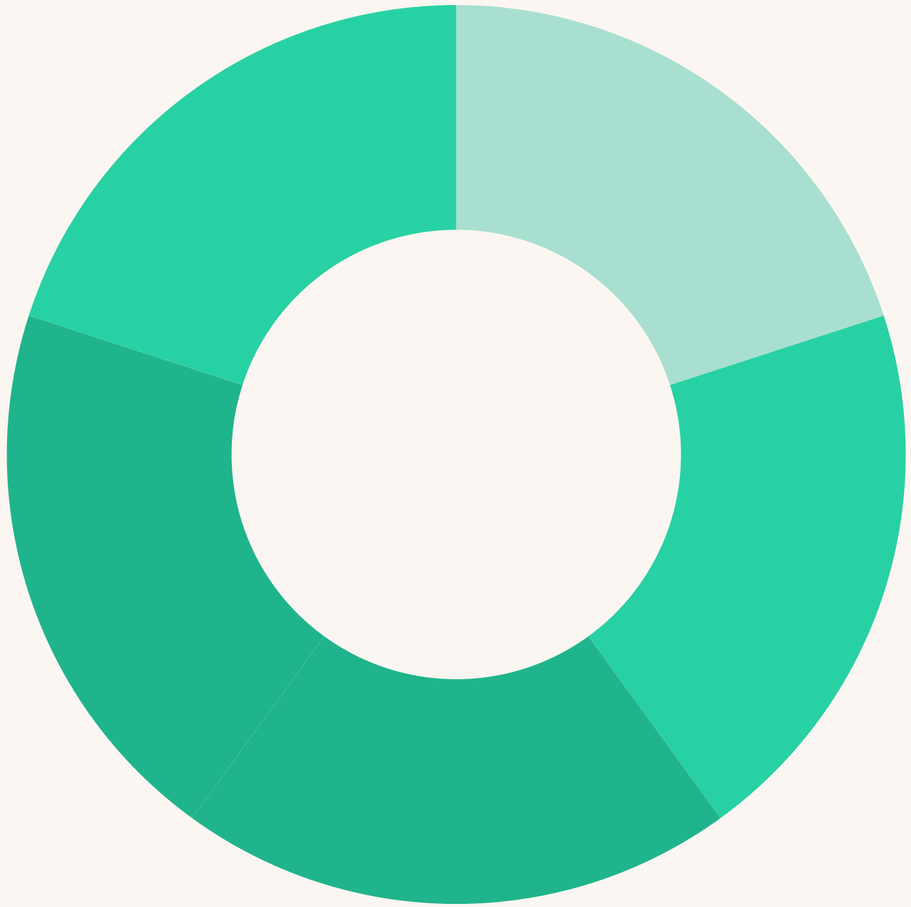
Alcohol: 0.389
Amphet: 0.633
Amyl: 0.787
Benzos: 0.400
Caff: 0.763
Cannabis:0.474
Coke:0.429
Crack:0.655
Ecstasy: 0.883
Heroin: 0.687
Ketamine: 0.848

Challenges and Learnings

- Choosing the right machine learning algorithms was a complex process. It required testing various algorithms and adjusting parameters to find the most effective model through experimentation and validation
- We encountered difficulties as we neared the project deadline, but at the end we succeeded

Future Work and Improvements

- Trying to bild new models with higher accuracy
- Adding more data points with a higher diversity to make the prediction model better



Any Questions ?

Presented by Alex & Olsi

Thank you very much!