




Drug Consumption Classification Checkpoint



Inteligência Artificial

Rita Cachaldora up202108798
Jorge Restivo up202108886
Samuel Maciel up202108697



Especificação do Trabalho

O objetivo deste trabalho é conceber um modelo capaz de prever o consumo de drogas com base na personalidade e na demografia de um indivíduo, com base no dataset descrito abaixo.

Dataset

- Contém registos de 1885 participantes.
- Cada registo com 12 atributos:
 - Atributos de Personalidade: NEO-FFI-R (neuroticismo, extroversão, abertura à experiência, amabilidade e conscienciosidade), BIS-11 (impulsividade) e ImpSS (busca de sensações).
 - Demográfico e Educacional: Inclui dados sobre nível de educação, idade, género, país de residência e etnia.
- Atributos de Uso de Drogas: Regista o uso de 18 drogas legais e ilegais, bem como uma droga fictícia (Semeron), quantificando o uso ao longo do tempo.
- Classificação dos Problemas: Contém 18 problemas de classificação, cada um com sete classes distintas, representando diferentes padrões de uso de drogas.



Referências Relacionadas

Machine Learning for Behavioral Prediction

https://www.academia.edu/75314154/Predicting_Real_World_Behaviors_from_Virtual_World_Data?hb-sb-sw=4577250

Personality Traits and Drug Consumption

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10548323/>



Ferramentas e Algoritmos

Ferramentas

1. Pandas e NumPy: Ótimas para manipulação e pré-processamento de dados, como codificação de variáveis categóricas, normalização de dados e tratamento de valores ausentes.
2. Matplotlib: Útil para visualização de dados, ajudando a entender padrões nos dados e a avaliar o desempenho dos modelos.

Algoritmos

1. Decision Trees:
 - Intuitivo e interpretável, adequado para capturar relações não lineares em dados de tipos mistos.
2. Logistic Regression:
 - Eficiente para tarefas de classificação binária, fornecendo interpretações probabilísticas das previsões.
3. Neural Networks (for Complex Modeling):
 - Destaca-se na captura de padrões complexos e não lineares e escala bem com conjuntos de dados grandes, embora seja menos interpretável do que modelos mais simples como árvores de decisão ou regressão logística.



Implementação

Linguagem

Python3.

Ambiente de Programação

PyCharm.

Estruturas de Dados

- Features (X):
 - Atributos Demográficos
 - Traços de Personalidade
- Target Variable (y):
 - Categorias de Comportamento de Uso de Drogas:
 - Rótulos codificados para diferentes frequências de uso (ex: 0 para "Nunca Usado", 1 para "Usado há mais de uma década", etc.)

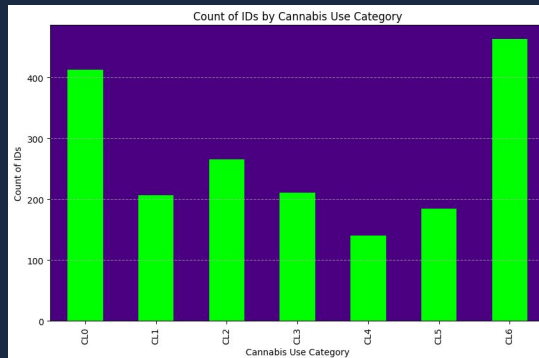


Data Pre-processing

Antes de começar, verificamos se tem duplicados e espaços vazios e retiramos todas as drogas exceto a “cannabis” porque é o objeto em estudo.

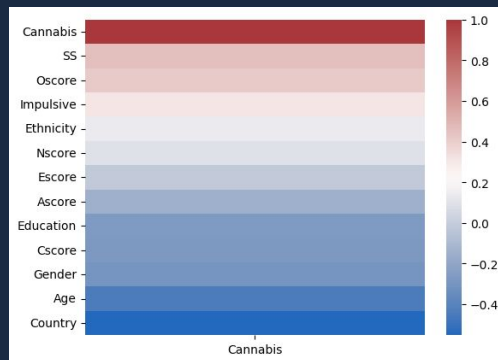
A cannabis estava separada por 6 categorias de consumo por cada pessoa que respondeu, por isso mudamos as categorias para só duas (‘Never Used’ e ‘Used Recently’) de forma a melhorar a accuracy.

Value	Description
CL0	Never Used
CL1	Used over a Decade Ago
CL2	Used in Last Decade
CL3	Used in Last Year
CL4	Used in Last Month
CL5	Used in Last Week
CL6	Used in Last Day



Data Pre-processing

Fazendo um correlação entre os dados e o consumo de cannabis, percebemos que quanto maior a abertura mental (**Oscore**) de uma pessoa e melhor as sensações (**SS**) produzidas pela maconha, maior a probabilidade de consumo.



Algoritmos Desenvolvidos

- Decision Tree - 74,54% accuracy
- Neural Network - 76,39% accuracy
- Naive Bayes - 78,25% accuracy
- Logistic Regression - 81,96% accuracy



Resultados

O algoritmo com melhor accuracy foi o Logistic Regression, contudo ainda tivemos:

- 88 True Negative
- 26 False Positive
- 42 False Negative
- 221 True Positive

Achamos que talvez os resultados estão um pouco desequilibrados pois há claramente mais “Used Recently” do “Never Used”.

