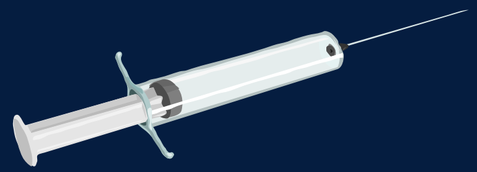
DATASET: DRUGS ANALYTICS

ANÁLISE E EXPLORAÇÃO DOS DADOS
APRESENTAÇÃO DO MODELO

Alunos: Iury Mota Santos, Thiago Rocha Amaral, Bernardo Costa Lima Bentes, Artur Braga Mota.
Curso: Ciência de Dados
Período: 1° Período
Universidade: Puc-MG

DrugAnalytics

O projeto visa desenvolver um sistema inteligente para analisar bancos de dados sobre o uso de substâncias na sociedade, fornecendo feedback customizado para entidades privadas ou públicas. O sistema será projetado para realizar análises de dados atuais, utilizando bancos de dados confiáveis. O processo incluirá extração completa, interpretação e análise dos dados, além de visualizações intuitivas e relatórios detalhados. O objetivo é auxiliar na tomada de decisões informadas, promovendo políticas de saúde pública e estratégias de prevenção eficazes. O sistema será desenvolvido para análises de dados atuais, tendo como base bancos de dados de credibilidade que a partir de um processo de completo extração, interpretação e análise dentre outras etapas para gerar bons resultados.





DrugAnalytics

Bibliotecas Utilizadas:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.metrics import accuracy_score, classification_report
```

- **Pandas: Ferramenta essencial para manipulação e análise de dados em Python, permitindo leitura e processamento de dados em estruturas como DataFrames.**
- **Seaborn: Biblioteca de visualização baseada no Matplotlib, otimizada para criar gráficos estatísticos e informativos.**
- **Matplotlib: Utilizada para criar gráficos e visualizações em 2D, oferecendo controle detalhado sobre o layout e a aparência dos gráficos.**
- **NumPy: Biblioteca fundamental para operações numéricas, facilitando o uso de arrays multidimensionais e funções matemáticas de alto desempenho.**
- **Scikit-learn: Biblioteca de aprendizado de máquina que fornece ferramentas para modelagem preditiva, incluindo regressão, classificação e validação cruzada.**
- **LogisticRegression: Modelo de classificação do Scikit-learn que prevê a probabilidade de classes binárias.**
- **LinearRegression: Modelo de regressão linear que se ajusta a dados contínuos, identificando a relação linear entre variáveis independentes e dependentes.**
- **GradientBoostingClassifier: Algoritmo de ensemble que combina múltiplos modelos fracos para criar um modelo forte, frequentemente usado para melhorar a precisão em tarefas de classificação.**
- **Cross-validation: Técnica para avaliar a eficácia de um modelo de aprendizado de máquina, dividindo os dados em subconjuntos e treinando o modelo em diferentes porções para evitar overfitting.**
- **Mean_squared_error: Métrica de avaliação de modelos de regressão, calculando o quadrado médio das diferenças entre valores previstos e reais.**

DrugAnalytics

Atributos:

#	Column	Non-Null Count	Dtype
0	ID	1885 non-null	int64
1	Age	1885 non-null	float64
2	Gender	1885 non-null	float64
3	Education	1885 non-null	float64
4	Country	1885 non-null	float64
5	Ethnicity	1885 non-null	float64
6	Nscore	1885 non-null	float64
7	Escore	1885 non-null	float64
8	Oscore	1885 non-null	float64
9	Ascore	1885 non-null	float64
10	Cscore	1885 non-null	float64
11	Impulsive	1885 non-null	float64
12	SS	1885 non-null	float64
13	Alcohol	1885 non-null	object
14	Amphet	1885 non-null	object
15	Amyl	1885 non-null	object
16	Benzos	1885 non-null	object
17	Caff	1885 non-null	object
18	Cannabis	1885 non-null	object
19	Choc	1885 non-null	object
20	Coke	1885 non-null	object
21	Crack	1885 non-null	object
22	Ecstasy	1885 non-null	object
23	Heroin	1885 non-null	object
24	Ketamine	1885 non-null	object
25	Legalh	1885 non-null	object
26	LSD	1885 non-null	object
27	Meth	1885 non-null	object
28	Mushrooms	1885 non-null	object
29	Nicotine	1885 non-null	object
30	Semer	1885 non-null	object
31	VSA	1885 non-null	object

Os dados consistem em várias características relacionadas a indivíduos, abrangendo aspectos demográficos, psicológicos e comportamentais:

- ID: Identificador único para cada entrada.
- Age: Idade dos participantes, representada como um número decimal
- Gender: Gênero dos participantes, codificado numericamente.
- Education: Nível educacional, também codificado numericamente.
- Country: País de origem, representado numericamente.
- Ethnicity: Etnia dos participantes, codificada numericamente.
- Nscore a Cscore: Pontuações em diferentes dimensões de personalidade (Neuroticismo, Extroversão, Abertura, Amabilidade, Conscienciosidade).
- Impulsive: Medida de impulsividade.
- SS: Sensação de busca.
- Alcohol a VSA: Uso de diversas substâncias (álcool, anfetaminas, cannabis, etc.), representado como categorias de dados.

DrugAnalytics

Preparação do modelo:

Durante a preparação dos dados, verificamos a presença de valores ausentes e substituímos as classificações categóricas por valores apropriados. Em seguida, realizamos o pré-processamento necessário para preparar o conjunto de dados para a inicialização do primeiro modelo.

```
drugs.isnull().sum()
#Observar se a dados omissos dentro do Dataset.
```

```
ID          0
Age         0
Gender      0
Education   0
Country     0
Ethnicity   0
Nscore      0
Escore      0
Oscore      0
Ascore      0
Cscore      0
Impulsive   0
SS          0
Alcohol     0
Amphet      0
Amyl        0
Benzos      0
Caff        0
Cannabis    0
Choc        0
Coke        0
Crack       0
Ecstasy     0
Heroin      0
Ketamine    0
Legalh      0
LSD         0
Meth        0
Mushrooms   0
Nicotine    0
Semer       0
VSA         0
dtype: int64
```

```
drugs['Age'] = drugs['Age'].apply(lambda x: str(x).replace( '-0.95197','18 - 24'))
drugs['Age'] = drugs['Age'].apply(lambda x: str(x).replace( '-0.07854','25 - 34'))
drugs['Age'] = drugs['Age'].apply(lambda x: str(x).replace( '0.49788','35 - 44'))
drugs['Age'] = drugs['Age'].apply(lambda x: str(x).replace( '1.09449','45 - 54'))
drugs['Age'] = drugs['Age'].apply(lambda x: str(x).replace( '1.82213','55 - 64'))
drugs['Age'] = drugs['Age'].apply(lambda x: str(x).replace( '2.59171','65+'))
```

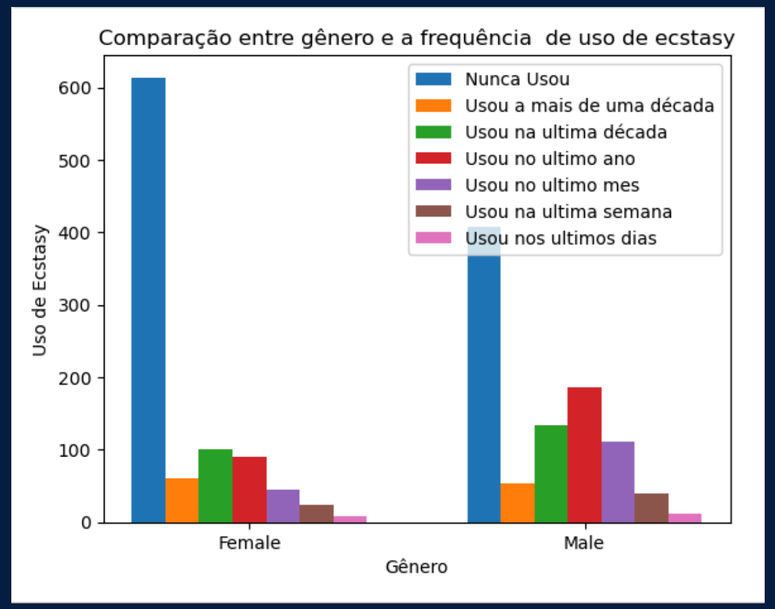
```
drugs['Alcohol'] = drugs['Alcohol'].apply(lambda x: str(x).replace( 'CL0','Nunca Usou'))
drugs['Alcohol'] = drugs['Alcohol'].apply(lambda x: str(x).replace( 'CL1','Usou a mais de uma década'))
drugs['Alcohol'] = drugs['Alcohol'].apply(lambda x: str(x).replace( 'CL2','Usou na ultima década'))
drugs['Alcohol'] = drugs['Alcohol'].apply(lambda x: str(x).replace( 'CL3','Usou no ultimo ano'))
drugs['Alcohol'] = drugs['Alcohol'].apply(lambda x: str(x).replace( 'CL4','Usou no ultimo mes'))
drugs['Alcohol'] = drugs['Alcohol'].apply(lambda x: str(x).replace( 'CL5','Usou na ultima semana'))
drugs['Alcohol'] = drugs['Alcohol'].apply(lambda x: str(x).replace( 'CL6','Usou nos ultimos dias'))
```

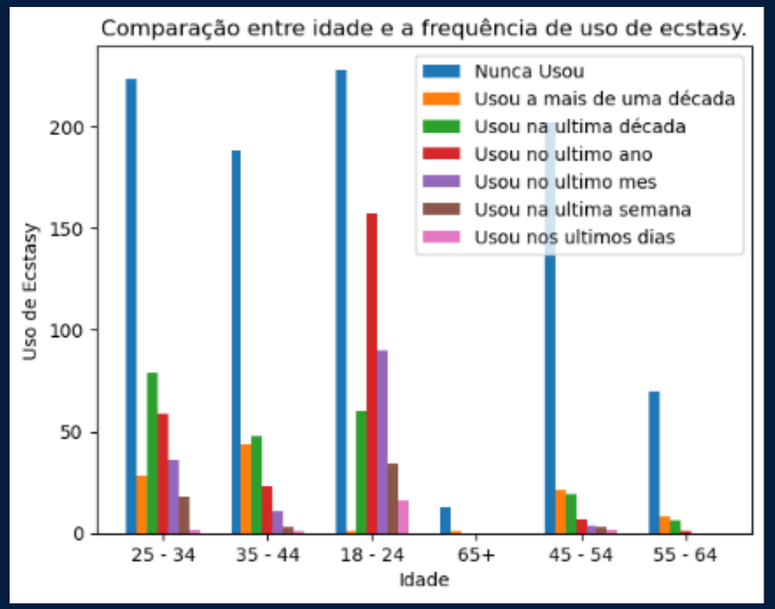
```
drugs['Education'] = drugs['Education'].apply(lambda x: str(x).replace( '-2.43591','Saiu da escola com menos de 16 anos'))
drugs['Education'] = drugs['Education'].apply(lambda x: str(x).replace( '-1.73790','Saiu da escola com 16 anos'))
drugs['Education'] = drugs['Education'].apply(lambda x: str(x).replace( '-1.43719','Saiu da escola com 17 anos'))
drugs['Education'] = drugs['Education'].apply(lambda x: str(x).replace( '-1.22751','Saiu da escola com 18 anos'))
drugs['Education'] = drugs['Education'].apply(lambda x: str(x).replace( '-0.61113','Ensino superior incompleto'))
drugs['Education'] = drugs['Education'].apply(lambda x: str(x).replace( '-0.05921','Profissional com certificado'))
drugs['Education'] = drugs['Education'].apply(lambda x: str(x).replace( '0.45468','Individuo Graduado'))
drugs['Education'] = drugs['Education'].apply(lambda x: str(x).replace( '1.16365','Individuo com Mestrado'))
drugs['Education'] = drugs['Education'].apply(lambda x: str(x).replace( '1.98437','Individuo com Doutorado'))
```

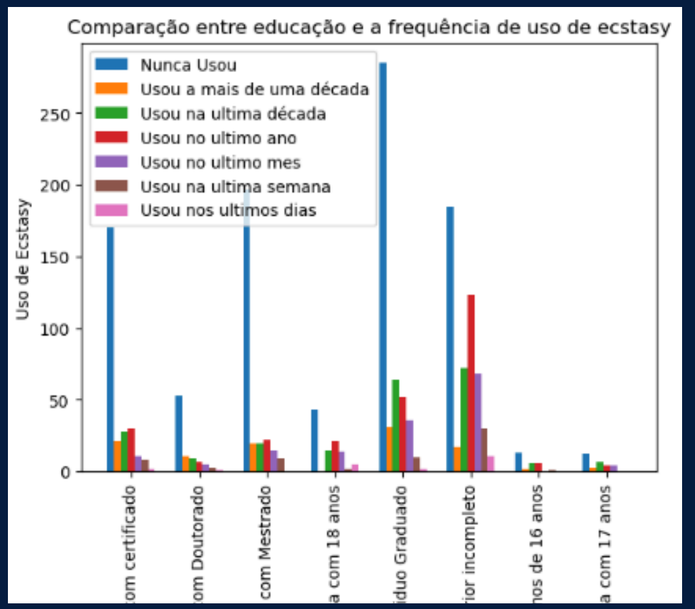
```
drugs['Gender'] = drugs['Gender'].replace({0.48246: 'Female', -0.48246: 'Male'})
drugs['Ethnicity'] = drugs['Ethnicity'].replace({-0.50212: 'Asian', -1.10702: 'Black', 1.90725:
drugs['Country'] = drugs['Country'].replace({-0.09765: 'Australia', 0.24923: 'Canada',-0.46841:
```

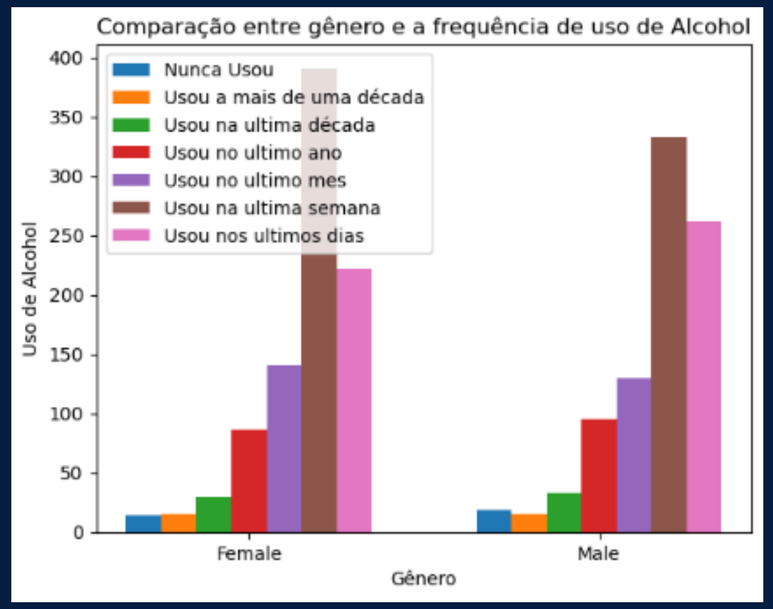
DrugAnalytics

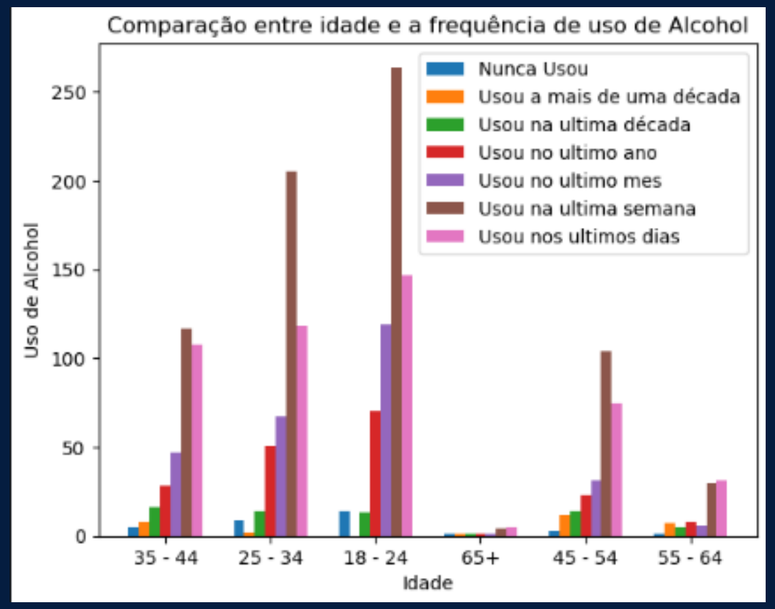
Gráficos Comparativos Gerados:

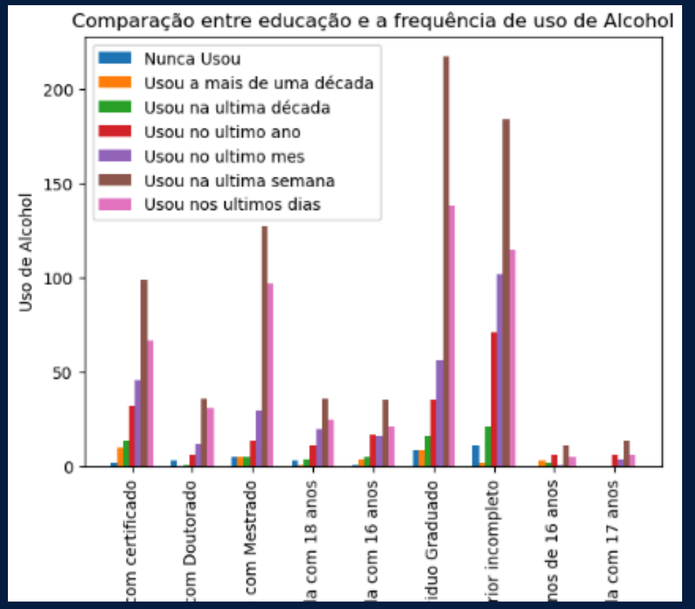












DrugAnalytics

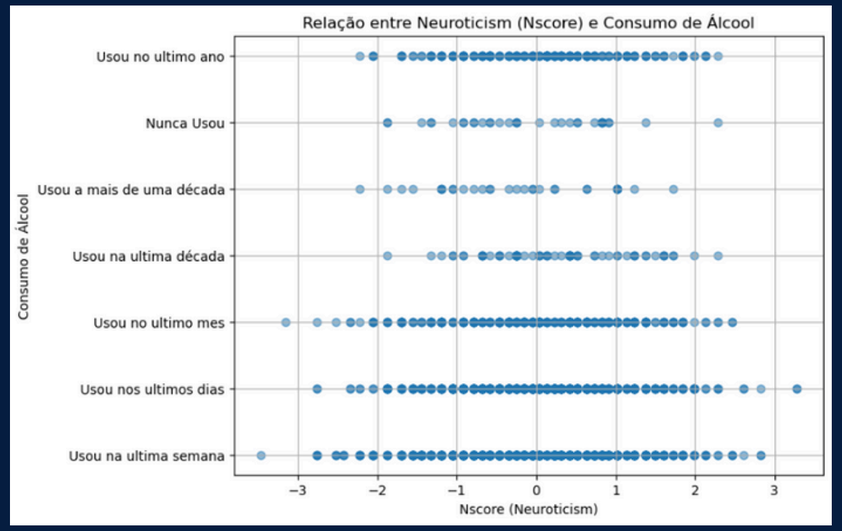
Descrição dos Dados das relações psicológicas e sensitivas dos entrevistados:

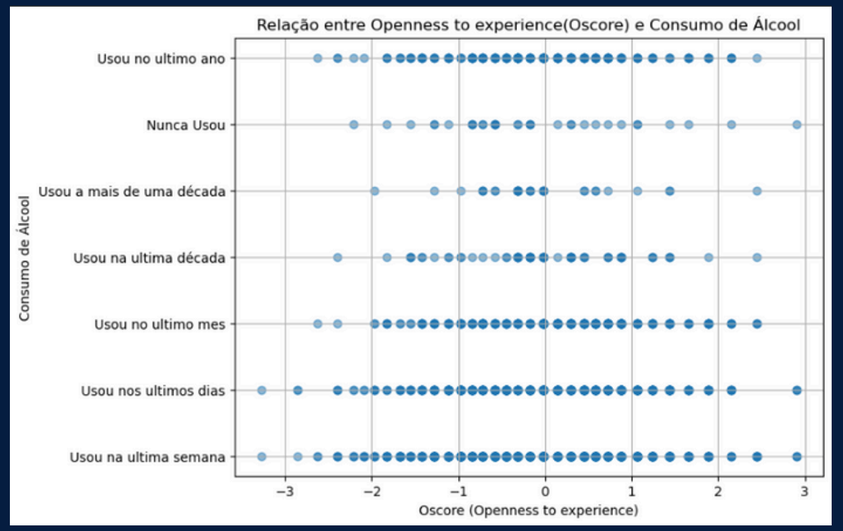
Atríbuto	Descrição	Tipo de dado
Nscore	Indivíduos com pontuação alta em neuroticismo têm maior probabilidade do que a média de serem mal-humorados e de experimentar sentimentos como ansiedade, preocupação, medo, raiva, frustração, inveja, ciúme, culpa, humor deprimido e solidão.	Númerico
Escore	Uma pessoa com pontuação alta em extroversão em um teste de personalidade é a vida da festa. Gostam de estar com as pessoas, de participar de reuniões sociais e são cheios de energia.	Númerico
Oscore	Uma pessoa com um alto nível de abertura para experiências em um teste de personalidade gosta de experimentar coisas novas. Eles são imaginativos, curiosos e de mente aberta. Indivíduos com baixa abertura à experiência preferem não tentar coisas novas. Eles têm a mente fechada, são literais e gostam de ter uma rotina.	Númerico
Ascore	Uma pessoa com alto nível de agradabilidade em um teste de personalidade geralmente é calorosa, amigável e diplomática. Geralmente têm uma visão otimista da natureza humana e se dão bem com os outros.	Númerico
Cscore	Uma pessoa com pontuação alta em consciência geralmente possui um alto nível de autodisciplina. Esses indivíduos preferem seguir um plano, em vez de agir espontaneamente.	Númerico
Impulsive	Na psicologia, a impulsividade (ou impulsividade) é uma tendência a agir por capricho, exibindo um comportamento caracterizado por pouca ou nenhuma premeditação, reflexão ou consideração das consequências.Se você descreve alguém como impulsivo, quer dizer que ele faz as coisas de repente, sem pensar primeiro nelas com cuidado.	Númerico

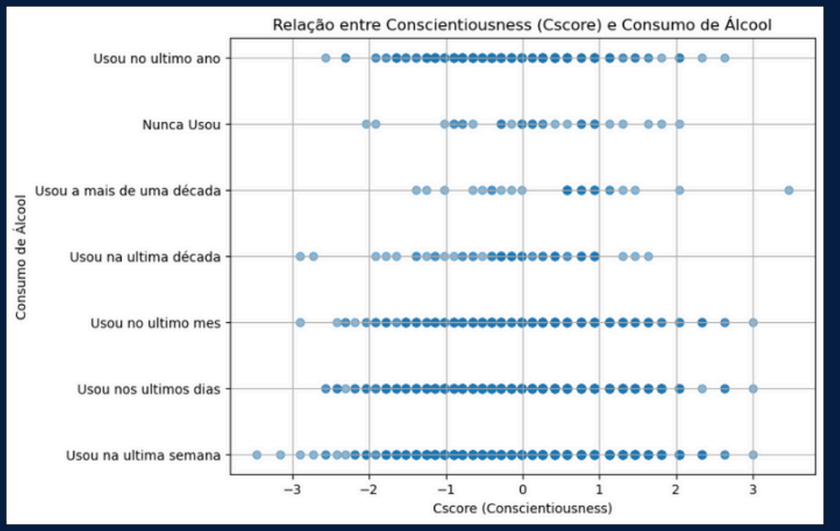
DrugAnalytics

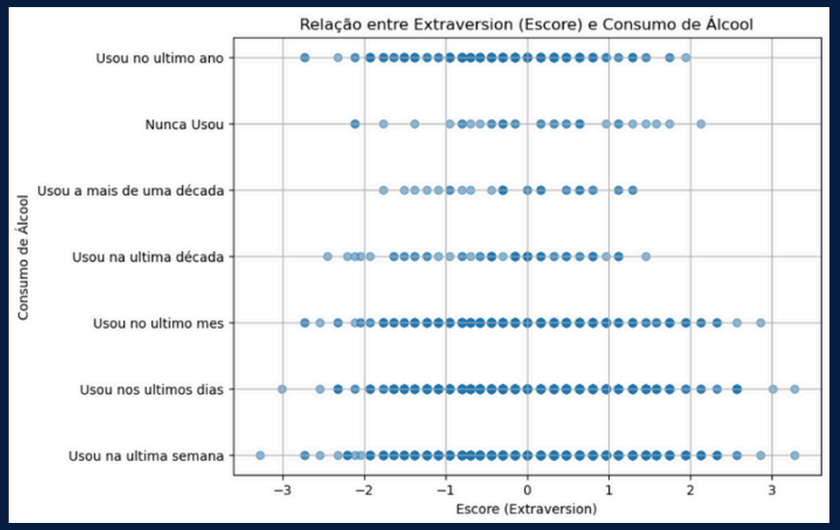
Gráficos Comparativos Gerados:

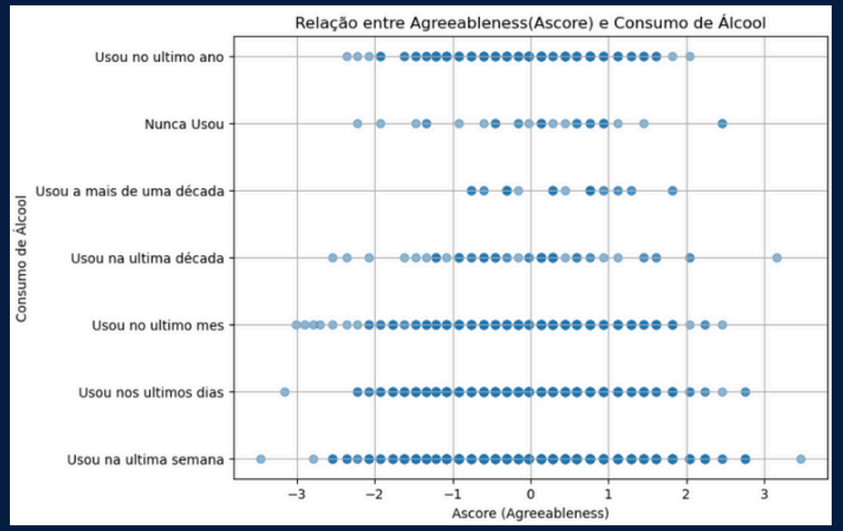
Relações Psicológicas e Sensitivas dos entrevistados.

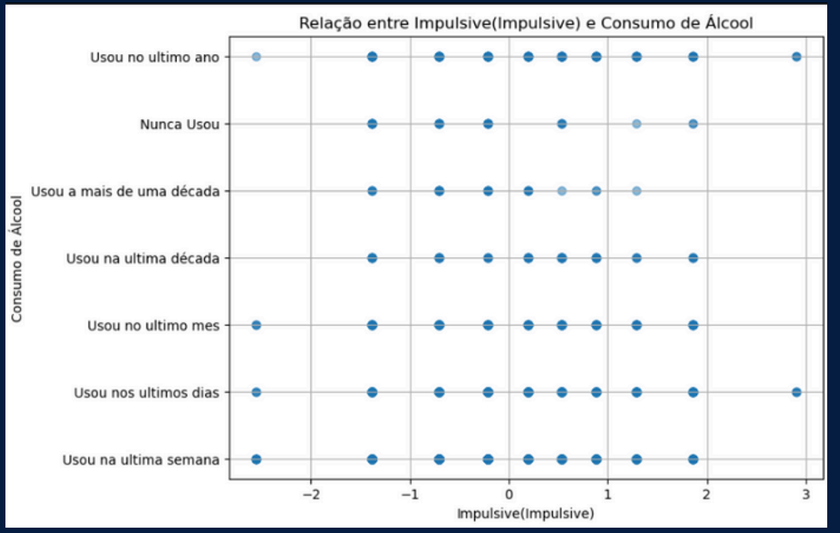












DrugAnalytics

Desenvolvimento do Primeiro Modelo:

Modelo 1: Regressão Linear.

Escolhemos a regressão linear para modelar a relação linear entre as variáveis. Nosso objetivo era prever o consumo de álcool com base em características individuais, tornando a regressão linear adequada por permitir uma interpretação direta do impacto de cada variável independente.
Dividimos os dados em 80% para treinamento e 20% para teste, garantindo uma avaliação eficaz do modelo.

```
X = drugs[['Nscore', 'Escore', 'Oscore', 'Ascore', 'Cscore']]
y = drugs['Alcohol']
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = LogisticRegression()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
print("Acurácia da Regressão Linear:", accuracy)

plt.figure(figsize=(8, 6))
plt.bar(['Acurácia'], [accuracy], color='Green')
plt.title('Acurácia da Regressão Linear')
plt.ylabel('Acurácia')
plt.ylim(0, 1)
plt.show()
```

Acurácia do modelo 1: 0.4%

```
Acurácia: 0.3994413407821229
Relatório de Classificação(Alcohol):
                          precision    recall  f1-score   support

              Nunca Usou       0.00      0.00      0.00         5
Usou a mais de uma década       0.00      0.00      0.00         7
    Usou na ultima década       0.00      0.00      0.00         8
    Usou na ultima semana       0.41      0.92      0.57       148
       Usou no ultimo ano       0.00      0.00      0.00        37
       Usou no ultimo mes       0.00      0.00      0.00        55
    Usou nos ultimos dias       0.23      0.07      0.11        98

                 accuracy                           0.40       358
                macro avg       0.09      0.14      0.10       358
             weighted avg       0.24      0.40      0.27       358
```

Matriz de Confusão.

```
Matriz de Confusão:
[[  0   0   0   5   0   0   0]
 [  0   0   0   7   0   0   0]
 [  0   0   0   5   0   0   3]
 [  0   0   0 136   0   0  12]
 [  0   0   0  34   0   0   3]
 [  0   0   0  50   0   0   5]
 [  0   0   0  91   0   0   7]]
```

Relatorio de Classificação.

Precisão: O relatorio junto ao treinamento trouxe uma porcentagem precisa de 41% das instâncias dentro de uma das classificaçoes.

Recall: O relatorio junto ao treinamento trouxe uma porcentagem de 0.92% de acertividade em relação ao conjunto presente verificado.

F-Score: Com 0.57% de acertividade o treinamento mostrou um equilibrio balanceado entre as métricas precisão e recall.

É necessario a maximização de precisão e recall para todas as classes.

DrugAnalytics

Desenvolvimento do Segundo Modelo:

Modelo 2: GradientBoosting.

Optamos pelo Gradient Boosting devido à sua capacidade de construir modelos a partir de árvores de decisão sequenciais. O objetivo permaneceu o mesmo: prever o consumo de determinada substância com base nas características individuais dos participantes do dataset.

1-) Localizar as strings classificatórias e substituí-las por valores numéricos.

```
cl = {'CL0': 0, 'CL1': 1, 'CL2': 2, 'CL3': 3, 'CL4': 4, 'CL5': 5, 'CL6': 6}
```

2-) Localizar as strings classificatórias e substituí-las por valores numéricos.

```
colunas = ['Alcohol', 'Amphet', 'Amyl', 'Benzos', 'Caff', 'Cannabis',
           'Choc', 'Coke', 'Crack', 'Ecstasy', 'Heroin', 'Ketamine',
           'Legalh', 'LSD', 'Meth', 'Mushrooms', 'Nicotine', 'Semer', 'VSA']
```

3-) Definindo as variaveis X e Y do modelo..

Heroin Target

```
X = drugs[['Age', 'Gender', 'Education', 'Country', 'Ethnicity',
           'Nscore', 'Escore', 'Oscore', 'Ascore', 'Cscore',
           'Impulsive']]
y = drugs['Heroin']
```

3-) Definindo as variaveis X e Y do modelo..

Alcohol Target

```
X = drugs[['Age', 'Gender', 'Education', 'Country', 'Ethnicity',
           'Nscore', 'Escore', 'Oscore', 'Ascore', 'Cscore',
           'Impulsive']]
y = drugs['Alcohol']
```

4-) Inicializar o modelo.

```
model = GradientBoostingClassifier(n_estimators=100, learning_rate=0.1, max_depth=3, random_state=42)
model.fit(X, y)
y_pred = model.predict(X)
```

Relatorio de Classificação.(Alcohol)

```
Acurácia média da validação cruzada (Gradient Boosting): 0.37
Avaliação do modelo final:
              precision    recall  f1-score   support

           0       1.00      0.47      0.64        34
           1       1.00      0.74      0.85        34
           2       0.97      0.46      0.62        68
           3       0.94      0.31      0.47       198
           4       0.95      0.28      0.44       287
           5       0.55      0.94      0.69       759
           6       0.70      0.49      0.58       505

    accuracy                           0.62      1885
   macro avg       0.87      0.53      0.61      1885
weighted avg       0.72      0.62      0.60      1885
```

Relatorio de Classificação.(Heroin)

```
Acurácia média da validação cruzada (Gradient Boosting): 0.83
Avaliação do modelo final:
              precision    recall  f1-score   support

           0       0.91      1.00      0.95      1605
           1       0.97      0.43      0.59        68
           2       1.00      0.32      0.48        94
           3       1.00      0.35      0.52        65
           4       1.00      0.75      0.86        24
           5       1.00      0.94      0.97        16
           6       1.00      1.00      1.00        13

    accuracy                           0.92      1885
   macro avg       0.98      0.68      0.77      1885
weighted avg       0.93      0.92      0.90      1885
```

GradientBoosting "Alcohol":
Acurácia: 0.37%

GradientBoosting "Heroin":
Acurácia: 0.83%

Durante a utilização dos modelos e obtendo seus resultados, é possível ver a disparidade relacionada ao processamento de informações ao lidar com algumas classes desbalanceadas como temos neste dataset. O modelo Gradient Boosting, por conseguir capturar diferentes aspectos dos dados, devido, por exemplo, à não linearidade das ações executadas no modelo, consegue modelar relações mais complexas, identificar os erros ao longo das iterações e, assim, otimizar o desempenho do modelo. Desta forma, o modelo de regressão linear é mais simples e direto, não capturando a complexidade do dataset como um todo, sendo inviável sua utilização em um conjunto de dados com tantas variações.

DrugAnalytics

Ameaças a Validade:

Validade Externa:

Generalização dos Resultados:

Os resultados obtidos pelo sistemas podem não ser generalizáveis para outras populações ou contextos além daqueles para os quais os dados foram coletados. Isso limita a aplicabilidade dos insights gerados em diferentes regiões geográficas ou em grupos demográficos distintos dos já análisados.

Interferência de Outras Variáveis:

A interação entre variáveis específicas no contexto do estudo e variáveis externas não controladas pode influenciar os resultados. Por exemplo, se mudanças demográficas ou econômicas afetarem os padrões de uso de drogas de maneira não prevista, podendo limitar a geração dos resultados obtidos.

DrugAnalytics

Ameaças a Validade:

Validade Interna:

Controle de Variáveis Externas:

Fatores externos que não são controlados podem influenciar os resultados. Por exemplo, mudanças nas políticas públicas relacionadas ao uso de drogas podem ocorrer durante o período de análise, afetando diretamente os resultados sem que isso seja considerado pelo sistema.

Viés de Seleção:

A seleção não aleatória de dados ou dos participantes pode introduzir viés nos resultados. Por exemplo, se determinados grupos populacionais são representados nos dados utilizados de forma equivocada pelo sistema, isso pode alterar as conclusões.

DrugAnalytics

Os resultados obtidos pelo sistema:

Vantagens:

Precisão na Análise:

Os modelos de Regressão Linear e Gradient Boosting permitiram processar eficientemente uma grande quantidade de dados. Além de prever o consumo de drogas, esses modelos forneceram análises consistentes para identificar padrões de comportamento e fatores de risco associados.

Visualização de Dados:

A utilização de gráficos e tabelas para apresentar os resultados facilitou a interpretação das análises, tornando as informações acessíveis e compreensíveis para diversos usuários.

Análises a Partir de Dados Diversificados:

Avaliamos a relação das substâncias com aspectos sociais como educação, idade e gênero utilizando scores. Isso enriqueceu as análises com informações relevantes e atuais, oferecendo insights valiosos sobre o consumo de substâncias na população estudada.

DrugAnalytics

Os resultados obtidos pelo sistema:

Desvantagens:

Necessidade de Constante Atualização:

A dinâmica do problema do uso de drogas exige que o sistema seja continuamente atualizado com novos dados e ajustado com novos algoritmos à medida que novas tendências e padrões forem surgindo.

Resultados Pouco Expressivos no 1° Modelo:

A implementação do primeiro modelo de Regressão Linear, gerou resultados pouco expressivos e ineficientes para que as análises sejam eficientes.

Limitações na Cobertura de Dados:

A abrangência dos dados disponíveis para a análise pode atrapalhar a ação dos modelos, uma vez que, dentre diversos dados alguns podem conter viés ideológicos e políticos.

DrugAnalytics

Conclusão:

Ao longo do trabalho, exploramos diversas perspectivas dos dados, desde características demográficas até traços de personalidade, utilizando modelos como regressão linear e GradientBoosting para prever comportamentos relacionados ao consumo de várias substâncias.

Os resultados obtidos, apesar dos desafios enfrentados ao longo do processo, revelaram avanços significativos na compreensão dos fatores que influenciam o uso de drogas e na capacidade preditiva dos modelos utilizados. Assim, o projeto DrugAnalytics não apenas oferece insights para melhorar políticas públicas ou estratégias no setor privado, mas também sublinha a importância das abordagens baseadas em dados para enfrentar um problema complexo e de grandes proporções como o uso de drogas. A análise realizada não se limita a estatísticas estáticas, mas propõe um sistema dinâmico e adaptável, essencial para lidar com os diversos cenários possíveis relacionados a um problema desafiador e duradouro como esse.

Como conclusão final, o projeto abre novas oportunidades de pesquisa e coleta de dados, destacando a necessidade de colaboração entre ciência de dados e saúde pública e privada para melhorar continuamente a disponibilidade de informações confiáveis e embasar decisões fundamentadas em benefício da sociedade.