O consumo de álcool no mundo e suas consequências

1st Alice Buarque Cadete
abc3@cin.ufpe.br

2nd Beatriz Galhardo Carneiro Leão
bgcl@cin.ufpe.br

3rd Danilo Lima de Carvalho
dlc3@cin.ufpe.br

4th Luisa Fonseca Leiria de Andrade
lfla@cin.ufpe.br

***Abstract*—Este projeto tem como objetivo realizar um estudo sobre o consumo de álcool na atualidade, utilizando informações relevantes sobre diversos países. Um modelo preditivo será desenvolvido em Python, incorporando um conjunto de técnicas de aprendizado de máquina, como o princípio de Naive Bayes. Esta análise visa proporcionar uma ferramenta significativa para compreender as causas subjacentes ao consumo de álcool além de suas consequências diretas no século XXI.**

Keywords— Álcool, Consumo, consequências, Naive Bayes, Modelo Preditor

I. INTRODUÇÃO

O perfil do consumidor de bebidas alcoólicas é complexo e multifatorial, sendo clara a relação entre o consumo de álcool e fatores como renda e taxa urbana, sendo esses influenciados pelas diversas dinâmicas sociais e econômicas de cada região do mundo. Além disso, também é possível observar a relação entre o consumo de bebidas alcoólicas e as variadas taxas de suicídio e de empregabilidade em diferentes países.

Em resumo, o consumo de álcool é regido por uma ampla gama de fatores. Compreender como esses fatores se entrelaçam pode fornecer insights valiosos para o desenvolvimento de políticas públicas direcionadas à redução do consumo de álcool e à mitigação de seus impactos sociais.

II. OBJETIVOS

Este projeto tem como principal foco compreender e antecipar os padrões de comportamento relacionados ao consumo excessivo de álcool, desenvolvendo um classificador ingênuo de Bayes para prever relações entre o consumo de álcool e fatos como taxas de empregos, suicídios, renda e urbana em um país. A abordagem se baseia nas características gerais de países de diversas regiões.

III. JUSTIFICATIVA

A realização de uma análise sobre o consumo de álcool ao redor do globo a fim de determinar a relação entre o consumo dessa droga com fatores sociais e econômicos. Por meio da investigação desses dados, torna-se viável a identificação da relação entre a ingestão de destilados e aspectos presentes na sociedade contemporânea. Além disso, essa ferramenta pode ser útil em pesquisas sobre o uso exacerbado de álcool e seus malefícios aos cidadãos, o que possibilita o desenvolvimento de políticas públicas que visam mitigar o consumo dessa droga.

IV. METODOLOGIA

A. Dataset

A fim de. Obter um embasamento de análises, serão utilizadas bases de dados disponíveis na Fundação Gapminder que possui informações sobre os seguintes tributos:

1. Consumo de Alcool: Numérico
2. Renda: Numérico
3. Taxa de Suicídio: Numérico
4. Taxa de Emprego: Numérico
5. Taxa Urbana: Numérico

B. Processamento do dataset

O processamento do dataset será aplicado em 3 etapas:

1. Filtro de instâncias: Nesta etapa, dados incompletos serão removidos ou preenchidos. Entretanto, como não há dados faltando no dataset, apenas instâncias irrelevantes e duplicadas serão removidas.
2. Seleção de atributos: Nessa etapa, serão escolhidos atributos do conjunto de dados usados como entrada no modelo. Serão necessários dados de grande dimensão, por isso, os atributos relevantes para a análise serão usados.
3. Engenharia de atributos: Nessa etapa, acontecerá a transformação de atributos para que se tornem entradas prontas para o algoritmo de aprendizagem. Como os atributos já foram tratados, essa etapa se torna desnecessária.

C. Teorema de Bayes

O Teorema de Bayes, tambem conhecido como o "cálculo que prova milagres", proposto por Pierre-Simon Laplace em 1812, é um conceito fundamental na teoria das probabilidades e estatística, tendo o propósito de calcular a probabilidade de um evento acontecer, dado que outro evento já ocorreu previamente, ou seja, calculando a probabilidade condicional.

Esse teorema recebe esse nome devido a seguinte formula desenvolvida no século XVIII pelo pastor matemático Thomas Bayes.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Onde:

- P(A|B): Probabilidade do evento A ocorrer, dado que o evento B já ocorreu.

- P(B|A): Probabilidade do evento B ocorrer, dado que o evento A já ocorreu.
- P(A): Probabilidade do evento A ocorrer.
- P(B): Probabilidade do evento B ocorrer.

Atualmente o Teorema de Bayes possui diversas aplicações na área de Machine Learning e Inteligência artificial, sendo utilizados em algoritmos, como o Classificador Naive Bayes.

D. Classificador Naive Bayes

O classificador Naive Bayes é um dos algoritmos mais importantes e tradicionais de aprendizagem de máquina, representando uma solução simples para problemas de classificação e sendo essencial para cientistas de dados. Este algoritmo se baseia no Teorema de Bayes para realizar predições em aprendizagem de máquina. Ele analisa as características de uma base de dados de maneira ingênua ("Naive"), desconsiderando completamente a correlação entre features (variáveis).

Para concluir as classificações com boa precisão ele precisa de um pequeno número de dados, sendo um bom modelo para a classificações de dados discretos. Uma das mais famosas aplicações desse método é na classificação de SPAM, em que ele analisa e-mails e com base nas suas informações e estrutura, o classifica como spam ou não spam.

O funcionamento desse algoritmo é relativamente simples, possuindo um desempenho maior que outros classificadores, sendo assim amplamente adotado pelos cientistas de dados. Tal método pode ser descrito em termos estatísticos: Inicialmente, o algoritmo estabelece uma tabela de probabilidades que contém a frequência dos preditores em relação às variáveis de saída. Em seguida, ocorre a análise dessa tabela, com a criação de classes, proporcionando assim uma resposta com base nos critérios estabelecidos.

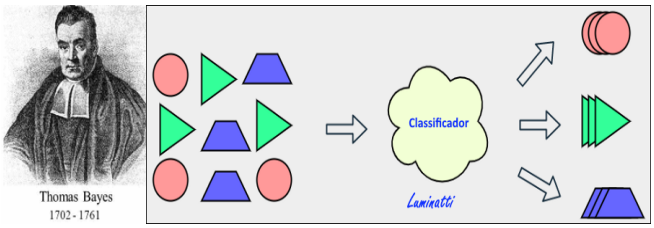


Fig 1. Visualização do funcionamento do classificador

E. Aplicação

A implementação do classificador Naive Bayes foi feita na plataforma Google Colab utilizando a linguagem de programação Python. Foi empregada a biblioteca Scikit-Learn, conhecida por sua documentação abrangente e implementações eficientes de algoritmos de aprendizado de máquina, incluindo o Naive Bayes .

Para realizar a análise de dados e operações matemáticas, são utilizadas as bibliotecas Pandas, NumPy e SciPy, pois se integram de forma harmoniosa com o Scikit-learn. O Pandas é essencial para a manipulação e organização de dados em formato de DataFrame, simplificando e tornando mais fluida a tarefa de pré-processamento. O NumPy é empregado para executar operações matemáticas e cálculos numéricos, fornecendo uma base sólida para uma variedade de funcionalidades. Além disso, o SciPy complementa o NumPy com funcionalidades adicionais para análise de dados e estatísticas, ampliando ainda mais as capacidades de análise e modelagem de dados.

Além das ferramentas essenciais mencionadas, é importante destacar o uso das bibliotecas Seaborn e Matplotlib. Matplotlib oferece recursos robustos para criar uma variedade de gráficos e plots, permitindo a representação visual das distribuições e tendências dos dados. Por sua vez, Seaborn complementa essa visualização com paletas de cores atraentes e funções específicas para gráficos estatísticos, enriquecendo a exploração de dados com detalhes informativos.

A integração das bibliotecas Scikit-learn, Pandas, NumPy, SciPy, Seaborn e Matplotlib, juntamente com a linguagem Python e o ambiente Google Colab, assegurou uma implementação eficiente e precisa do classificador Naive Bayes. Além disso, proporciona uma experiência de desenvolvimento mais fluida e facilitada, já que permite a exploração, bem como a visualização dos dados.

O projeto compreenderá uma etapa crucial de análise e tratamento dos dados, garantindo a qualidade e relevância das informações utilizadas. Posteriormente, o conjunto de dados será dividido em duas partes distintas: o conjunto de treinamento e o conjunto de teste. O conjunto de treinamento será empregado para alimentar o modelo de classificação, permitindo que ele aprenda e se ajuste aos padrões presentes nos dados. Por sua vez, o conjunto de teste será reservado exclusivamente para avaliar a eficiência do classificador Naive Bayes, sendo fundamental para medir sua capacidade de generalização em novos dados não observados durante o treinamento.

A abordagem de dividir os dados em treino e teste é essencial para garantir que ele aprenda padrões que possam ser generalizados para novas amostras. Isso proporciona uma estimativa realista do desempenho do classificador em situações práticas. Ao avaliar o classificador com o conjunto de teste, podemos medir sua acurácia, precisão, recall e outras métricas de desempenho, oferecendo detalhes sobre sua capacidade de fazer previsões precisas e auxiliando na tomada de decisões.

V. ANÁLISE EXPLORATÓRIA

A análise exploratória desempenha um papel fundamental na implementação de algoritmos preditivos, atuando como uma fase crucial para a compreensão aprofundada dos dados antes de nos envolvermos no processo de modelagem. Essa etapa é essencial para assegurar a qualidade e confiabilidade dos dados, além de fornecer insights valiosos sobre tendências, padrões e correlações ocultas.

O cerne dessa análise consiste em uma investigação clara e minuciosa da distribuição dos dados, frequentemente auxiliada por representações gráficas, com o objetivo de compreender como cada parâmetro pode influenciar a variável de interesse. Dada a natureza diversificada dos conjuntos de dados, várias abordagens podem ser adotadas na análise exploratória. Em muitos casos, opta-se por uma análise univariada, que examina individualmente cada variável. Essa análise permite extrair informações cruciais, como a média, a mediana, o desvio padrão, os valores mínimo e máximo, bem como a frequência dos valores associados a uma variável específica. Esses dados estatísticos revelam-se essenciais para obtermos uma compreensão profunda dos dados em questão.

No âmbito das variáveis analisadas, existiam apenas variáveis continuas, então se fez necessário categoriza-las em grupos, isso porque, no Dataset original, cada variável assume valores diferentes, tornando inviável realizar a analise dos seus valores. Como solução e forma de tornar

possível a analise dessas, elas foram categorizadas em diferentes faixas de acordo com seus valores:

- Valor de Consumo: [0, 4[: consumo baixo (1)
  [4, 8[: consumo moderado (2)
  [8, inf[: consumo alto (3)

- Intervalo de Renda: [0, 600[: muito baixa (1)
  [600, 2400[: baixa (2)
  [2400, 8600[: media (3)
  [8600, inf[: alta (4)

- Taxa de Suicídio: [0, 5.8[: muito baixa (1)
  [5.8, 9[: baixa (2)
  [9, 13[: media (3)
  [13, inf[: alta (4)

- Taxa de Emprego: [0, 51.5[: muito baixa (1)
  [51.5, 58.8[: baixa (2)
  [58.8, 65[: media (3)
  [65, inf[: alta (4)

- Taxa Urbana: [0, 37[: muito baixa (1)
  [37, 57.6[: baixa (2)
  [57.6, 73.4[: media (3)
  [73.4, inf[: alta (4)

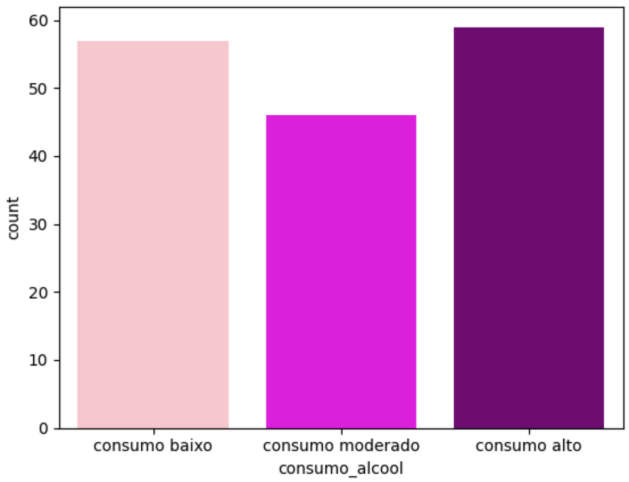


Fig2. Distribuição do target

Realizou-se uma contagem das linhas associadas a cada classe (1, 2 ou 3) da variável target, com o intuito de avaliar o equilíbrio do conjunto de dados. Observou-se que o dataset já estava balanceado, eliminando a necessidade de realizar quaisquer ajustes adicionais para balanceamento.

Com o propósito de aprofundar a compreensão, gerou-se um gráfico para cada variável, relacionando o número de ocorrências de cada classe do target com o atributo correspondente. Essa análise busca identificar a relevância do atributo e compreender de que forma ele efetivamente influencia o target.

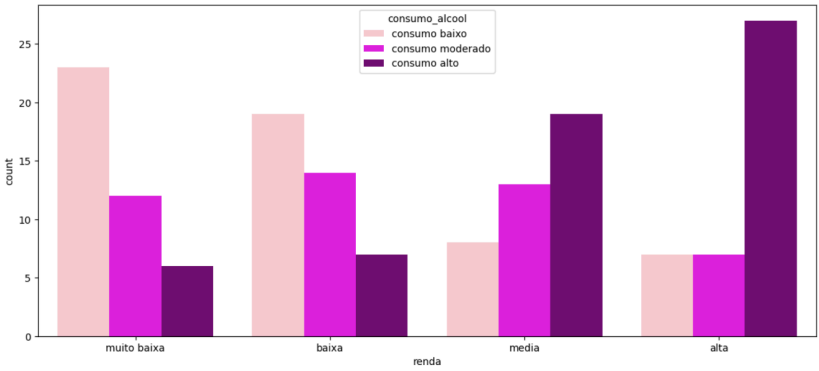


Fig3. Distribuição de renda

Como pode ser evidenciado no gráfico, em países com alta renda per capita, ocorre o maior consumo de bebidas alcoólicas, enquanto em países com baixa renda per capita observamos o menor consumo de tais destilados.

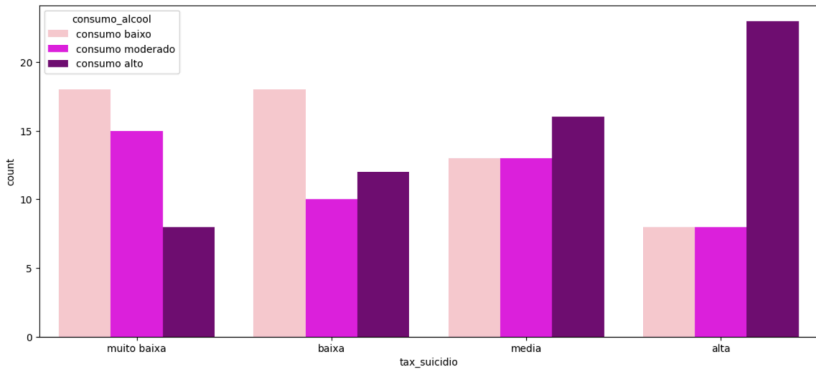


Fig 4. Taxa de Suicídio

É evidente ao analisar o gráfico, que em países com alta taxa de suicídio o consumo alcoólico se faz altamente presente, assim como em países onde a taxa de suicídio é considerada muito baixa o consumo do álcool não é elevado.

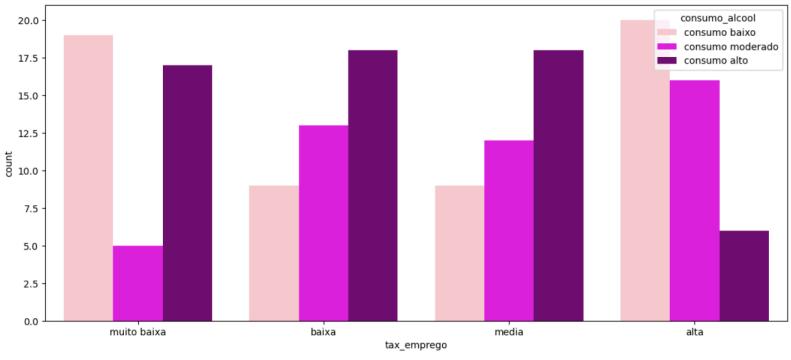


Fig 5. Taxa de Emprego

Analisando o gráfico da taxa de emprego em relação ao consumo alcoólico, é facilmente observado que lugares com alta taxa de emprego possuem o baixo consumo alcoólico predominante. Porém é notório que em países onde a taxa de emprego é muito reduzida a taxa de alto consumo apresenta semelhança com a de baixo consumo.

Tal fato se faz presente devido aos países onde a renda é baixíssima, assim como as taxas de emprego e nesses locais o consumo da bebida alcoólica também não ocorre de maneira extremamente acentuada.

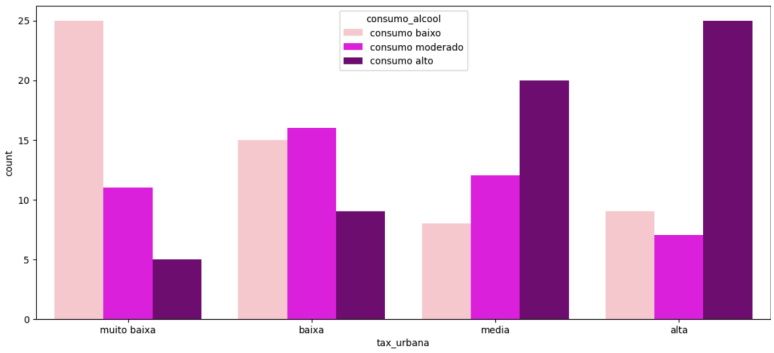


Fig 6. Taxa Urbana

Ao Analisar o gráfico é possível notar que em países com maior taxa urbana o consumo alcoólico se faz mais presente que em países com a taxa urbana extremamente reduzida.

VI. Resultados e Discussão

A função 'classification_report', disponível na biblioteca Scikit-Learn, foi a escolhida para a visualização dos resultados. Quando 30% da distribuição foi para teste a acurácia foi de 0.47, já utilizando 20% tivemos a seguinte acurácia média:

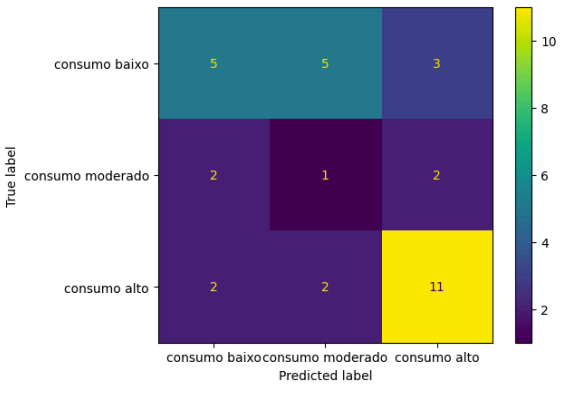
```
Accuracy: 0.5151515151515151
Classification Report:
                  precision    recall  f1-score   support

    consumo alto       0.69      0.73      0.71        15
   consumo baixo       0.56      0.38      0.45        13
consumo moderado       0.12      0.20      0.15         5

        accuracy                           0.52        33
       macro avg       0.46      0.44      0.44        33
    weighted avg       0.55      0.52      0.52        33
```

Para uma compreensão mais aprofundada dos resultados, foi elaborada uma Matriz de Confusão utilizando a biblioteca Scikit-Learn para uma visualização gráfica mais clara. Esta matriz permite avaliar o desempenho do modelo de machine learning de forma mais detalhada do que métricas simples, como a acurácia. Além disso, auxilia na identificação dos tipos de erros que o modelo está cometendo, revelando quantos falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos estão sendo produzidos. Isso é valioso para compreender quais classes o modelo tem dificuldade em prever corretamente e onde estão os problemas.

A Matriz de Confusão é usada para calcular várias métricas de desempenho, como precisão, recall, F1-score e a matriz de confusão balanceada (para classes desequilibradas). Essas métricas são mais informativas do que a acurácia, especialmente quando as classes possuem tamanhos diferentes ou quando os custos de falsos positivos e falsos negativos são distintos.



VII. Conclusão

A partir da análise dos dados evidenciada pela Matriz de Confusão, conclui-se que o modelo utilizado inicialmente não produziu resultados satisfatórios, uma vez que a acurácia encontrada foi relativamente baixa.

Isso pode ser explicado por dois fatores principais. Primeiramente, a grande complexidade dos dados utilizados, que extrapola a capacidade do modelo atual de análise. E, além disso, a ausência de informações críticas que contribuiriam na formação de um cenário mais confiável e que produziriam resultados mais detalhados.

Apesar dos desafios da metodologia inicial, a pesquisa apresentada apresenta uma promissora linha de investigação. Nesse sentido, melhorar a adequação do conjunto de dados com um modelo diferente do utilizado é uma possibilidade válida. Logo, a exploração de modelos alternativos e a incorporação de informações relevantes, mas que se fizeram ausentes nessa pesquisa, consistem em uma oportunidade real de aprimorar os resultados obtidos e atingir uma maior precisão em nossa análise.

O código pode ser acessado através do link: https://github.com/luisaleiria/Projeto-de-Estatistica/tree/main

Referencias

1. Classificador Naive bayes da biblioteca Scikit Learn [https://scikitlearn.org/stable/modules/naive bayes.html]
2. A.C. James “O consumo nocivo de alcool: dados epidemiológicos mundiais” [https://www.saudedireta.com.br/docsupload/1333061103alcoolesuasconsequencias-pt-cap1.pdf]
3. OPAS: Álcool [*https://www.paho.org/pt/topicos/alcool]*
4. O que é Naive Bayes e como funciona esse algoritmo de classificação. [https://rockcontent.com/br/blog/naive-bayes/]
5. Algoritmo de Classificação de Naive Bayes [https://www.organicadigital.com/blog/algoritmo-de-classificacao-naive-bayes/]
6. Entenda o Teorema de Bayes [https://didatica.tech/entenda-o-teorema-de-bayes/]
7. Teorema de Bayes; Entenda o que é e como calcular [https://www.suno.com.br/artigos/teorema-de-bayes/]