O consumo de álcool per capita dos Estados Unidos

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 per capita nos Estados Unidos da América, utilizando informações relevantes sobre diversos cidadãos americanos. 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 na população dos Estados Unidos.**

Keywords— Álcool, Consumo per capita, Naive Bayes, Modelo Preditor

I. Introdução

O perfil do consumidor de bebidas alcoólicas é complexo e multifatorial, sendo influenciado por diversos fatores sociais, culturais e religiosos. Dentre esses fatores, idade, nível de riqueza, situação profissional e religião influenciam diretamente no nível de consumo individual. Cada um desses aspectos modifica a relação pessoal com a bebida e, por consequência, seu uso.

Em resumo, o consumo de álcool é regido por uma ampla gama de fatores. Entender a interação desses aspectos pode contribuir para o desenvolvimento de políticas públicas que visam diminuir o uso ou seus efeitos 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 os fatores que contribuem para o alto consumo alcoólico por uma pessoa. A abordagem se baseia nas características individuais, tais como profissão, religião, idade e nível de riqueza.

III. Justificativa

A realização de uma análise sobre o consumo de álcool per capita nos Estados Unidos da América a fim de determinar os principais fatores que relacionam o consumo dessa droga com a renda dos indivíduos. Por meio da investigação desses dados, torna-se viável a identificação dos aspectos socioeconômicos que mais influenciam a ingestão de destilados pela população norte-americana. 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 em Pew Research Center, pela empresa de análise e consultoria Gallup Pool e pelo National Center for Biotechnology Information, que possui informações sobre os seguintes tributos:

1. Idade da pessoa consumidora: categórico [1, 3];
2. Religião do consumidor: categórico [1, 3], sendo 1 = Católico, 2 = Protestante e 3 = Não Filiado;
3. Etinia do consumidor: categórico [1, 5];
4. Sexo do consumidor: binário (Homem ou Mulher);
5. Nível de Escolaridade: binário (Ensino Superior Completo, Ensino Superior Incompleto);
6. Renda Familiar Anual: categórico [1, 3];
7. Frequência de Serviço Religioso: categórico [1, 3];
8. Preferência de Bebida: categórico [1, 3];]

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.

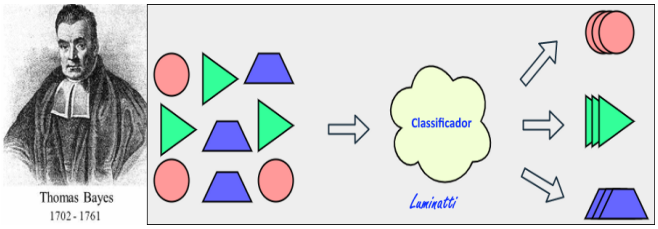

Fig 1. Visualização do funcionamento do classificador

E. Implementação

Para desenvolver o modelo destinado à análise de dados, optaremos por utilizar a linguagem Python no ambiente do Google Colaboratory. A nossa principal biblioteca será a Pandas, a qual se apoia em duas outras bibliotecas Python: Matplotlib e NumPy. Essa ferramenta é essencial para a manipulação e análise de dados, utilizando o Matplotlib para visualização gráfica e o NumPy para operações matemáticas.

Outra biblioteca que poderá ser utilizada é a Scikit-Learn, que possui algoritmos de classificação, regressão e agrupamento por ser uma biblioteca de *Machine Learning*, além de ter sido projetada para interagir com bibliotecas numéricas e científicas como as citadas acima.

Inicialmente, conduziremos a análise exploratória dos dados para adquirirmos uma compreensão sobre o seu comportamento. Isso incluirá a identificação das principais correlações, a avaliação de atributos que mais estão faltando e a detecção de atributos que possam conter valores incorretos, entre outros aspectos.

A segunda etapa será passar para o tratamento dos dados, de forma a retirar ou estimar instâncias com atributos faltantes ou valores incompatíveis com sua descrição. Posteriormente, os dados serão divididos entre dataset de treino e teste. O dataset de treino será empregado para treinar o modelo de Bayes, estimando as probabilidades necessárias. Já o dataset de teste será utilizado para avaliar a eficiência do classificador, utilizando dados que não participaram do treinamento a fim de evitar viés nos resultados.

V. Cronograma de Atividades

O grupo seguirá um plano de atividades de acordo com a tabela a seguir:

TABELA I

Cronograma de Atividades

Data*	***Atividade Planejada***	***Detalhes***
14/12/2023	Seleção do dataset	Em grupo
15/12/2023	Escrita da proposta	Em grupo
29/01/2024	Entrega da proposta	-
30/01/2024	Divisão de tarefas	Em grupo
02/01 – 20/02/2024	Desenvolvimento do projeto	Colab/Python
25/02/2024	Finalização do projeto	Em grupo
28/02/2024	Finalização do projeto	Em grupo
02/03/2024	Planejamento para apresentação	Em grupo
03/03 – 06/03/2024	Escrita do relatório	Individual
07/03 – 08/03/2024	Elaboração dos slides	Individual
11/03 – 18/03/2024	Entrega do projeto final	-

*As datas podem mudar em caso de imprevistos

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/1333061103alcoolesuas consequencias-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/]