**Previsão de Avaliação Geral de Cervejas com Base em Características Sensoriais Usando Machine Learning**

Alexandre Polizelli Marino¹\*; Wagner W. Bortoletto²

¹ Rua: Doutor Francisco José Longo, 191 – Apto 113 – Chácara Inglesa; 04140-060 São Paulo, SP, Brasil

**²** PECEGE. Mestre em Engenharia de Produção e Manufatura. Rua Antônio Menardi – Residencial Victor D’Andrea; 13481-344, Limeira, São Paulo, Brasil.

\*autor correspondente: ale\_marino1@hotmail.com

**Previsão de Avaliação Geral de Cervejas com Base em Características Sensoriais Usando Machine Learning**

**Resumo (ou Sumário Executivo)**

**Introdução**

O mercado de cervejas artesanais tem crescido exponencialmente ao longo dos últimos anos, levando a um aumento na diversidade de estilos, sabores e características sensoriais disponíveis para os consumidores. (MAPA, 2023)

Ao mesmo tempo, plataformas de avaliação de cervejas, como aplicativos e sites especializados, permitem que milhões de usuários compartilhem suas impressões e notas sobre diversos aspectos das bebidas que consomem.‌

A avaliação de qualidade de cervejas tem se tornado um aspecto crucial para a indústria, especialmente considerando o aumento significativo na diversidade de estilos e o crescente número de cervejarias artesanais no Brasil. Em 2022, o crescimento deste mercado no brasil foi de 11,6%, subindo para um total de 1.729 cervejarias registradas e é o terceiro maior produtor de cerveja no mundo, atrás apenas da China e dos Estados Unidos, conforme MAPA (2023). Dado tamanho e potencial crescimento do mercado, a compreensão de como as características sensoriais, como aroma, gosto, aparência e paladar, influenciam a avaliação geral dos consumidores pode fornecer insights valiosos para otimizar a produção e atender melhor às expectativas dos consumidores.

O aumento de dados disponibilizados por consumidores representa uma oportunidade valiosa para aplicar técnicas de ciência de dados com o objetivo de entender e prever como diferentes características de uma cerveja podem influenciar a avaliação geral que ela receberá. Estudos como o de ARAUJO (2003), procuram entender como as características sensoriais ajudam a traçar um perfil de consumo e este trabalho se propõe a utilizar modelos preditivos como regressão linear, árvores de decisão, e algoritmos de ensemble como Random Forest e XGBoost, para entender como estas variáveis impactam na percepção geral do consumidor.

Estudos anteriores sobre previsão de preferências de consumidores já demonstraram a eficácia de técnicas de machine learning em domínios como recomendação de filmes KNIAZIEVA (2022), e até mesmo previsão de sucesso de músicas, conforme MUKHOPADHYAY (2024). No entanto, a previsão da avaliação geral de cervejas com base em suas características sensoriais ainda é um campo pouco explorado, especialmente considerando a rica diversidade de variáveis envolvidas nesse processo de avaliação.

Este projeto busca explorar uma base de dados contendo mais de 1,5 milhões de avaliações de cervejas realizadas por consumidores. As avaliações abrangem múltiplos critérios sensoriais, como aroma, aparência, paladar e gosto, e permitem a análise de como essas variáveis podem prever a nota geral atribuída pelos consumidores. O dataset foi disponibilizado pela empresa BeerAdvocate, dos Estados Unidos e contém percepções de mais de 10 anos de consumidores finais que avaliaram as características sensoriais das cervejas e atribuíram uma nota geral.

A relevância deste estudo em demonstrar qual das características sensoriais entre paladar, gosto, aroma e aparência é a mais relevante na percepção geral do consumidor, contribui para o campo de previsão de preferências de consumidores e se o teor alcoólico da cerveja possui alguma influência na avaliação final, os resultados podem beneficiar diretamente cervejarias e distribuidores, que poderiam ajustar suas produções com base nas características sensoriais mais valorizadas pelos consumidores, aumentando suas chances de sucesso no mercado.

**Metodologia ou Material e Métodos**

O conjunto de dados utilizado contém 1.586.614 linhas e 14 colunas, onde cada linha representa a avaliação do consumidor em relação a cerveja e as colunas representam as variáveis como avaliações sensoriais (aroma, paladar, gosto e aparência), teor alcoólico, nome da cervejaria, estilo de cerveja, data da avaliação entre outros.

As variáveis deste banco de dados, são:

|  |  |  |
| --- | --- | --- |
| Variável (inglês) | Variável (português) | Observação |
| brewery\_id | id\_cervejaria | Identificador da cervejaria |
| brewery\_name | nome\_cervejaria | Nome da cervejaria |
| review\_time | data\_avaliacao | Data da avaliação |
| review\_overall | avaliação\_geral | Avaliação geral para a cerveja |
| **review\_aroma** | avaliação\_aroma | Avaliação para o aroma da cerveja |
| review\_appearance | avaliação\_aparencia | Avaliação para a aparência da cerveja |
| review\_palate | avaliação\_textura | Avaliação da textura da cerveja |
| review\_taste | avaliação\_gosto | Avaliação do gosto da cerveja |
| review\_profilename | nome\_avaliador | Nome do avaliador |
| beer\_style | estilo\_cerveja | Estilo da cerveja |
| beer\_name | nome\_cerveja | Nome da cerveja |
| beer\_abv | teor\_acoolico | Teor alcóolico da cerveja |
| beer\_beerid | id\_cerveja | Identificador da cerveja |

Fonte: <https://www.kaggle.com/datasets/thedevastator/1-5-million-beer-reviews-from-beer-advocate/data>.

. Durante a análise preliminar, foi constatada a necessidade de realizar diversas transformações e tratamentos, como conversão de variáveis e preenchimento de valores ausentes. Essas ações foram fundamentais para garantir a qualidade da base de dados e a integridade das variáveis utilizadas nos modelos de machine learning.

**Pacotes Utilizados**

Foram importadas bibliotecas essenciais para o desenvolvimento do modelo de machine learning, como “pandas”, “numpy” e “scikit-learn”. Essas bibliotecas desempenham papéis fundamentais na manipulação e análise de dados, além de fornecerem as ferramentas necessárias para a criação e avaliação dos modelos preditivos.

A biblioteca “pandas”, por exemplo, é utilizada para o carregamento e manipulação do conjunto de dados, permitindo organizar as variáveis de forma clara e acessível. Com o “numpy”, é possível realizar operações matemáticas e manipulações de “arrays” de maneira eficiente, essencial quando se trabalha com grandes volumes de dados. Já o “scikit-learn” é uma biblioteca poderosa para a implementação de algoritmos de machine learning, contendo diversas funcionalidades para realizar a divisão de dados, a construção de modelos e a avaliação de desempenho.

No âmbito de visualização de dados, pacotes como “matplotlib”, “seaborn” e “plotly” foram importados para a visualização dos dados e análise exploratória. Isso foi fundamental para entender melhor a distribuição das variáveis e possíveis correlações entre elas. Visualizações como histogramas, “boxplots” e gráficos de dispersão facilitaram a identificação de padrões e ajudaram na escolha das variáveis mais relevantes para o modelo.

**Tratamento de Valores Ausentes ou Inconsistentes**

Duas colunas apresentaram um número considerável de valores ausentes, o que exigiu um tratamento cuidadoso:

beer\_abv: Foi detectado 67.785 valores em branco, que foram preenchidos com a média do teor alcoólico por estilo de cerveja, assumindo que os mesmos tipos de cerveja possuem teores alcoólicos semelhantes. Esta abordagem foi adotada para evitar a perda de observações significativas, considerando que o teor alcoólico pode ser característica chave na análise da qualidade das cervejas.

review\_appearance: Esta variável, que avalia a aparência da cerveja, apresentou quinze observações com notas ausentes ou iguais a zero. Por não haver métodos de atribuição plausíveis que representem a nota do avaliador, as observações foram removidas.

brewery\_name: A variável tinha nove observações ausentes as quais foram removidas por não haver métodos de atribuição.

**Feature Engineering**:

Para melhorar a capacidade preditiva dos modelos, foram criadas novas variáveis a partir dos dados originais:

A variável “brewery\_popularity” foi criada visando refletir a popularidade das cervejarias, calculada com base no número total de avaliações recebidas por cada cervejaria. A hipótese a ser testada é que cervejarias mais populares podem receber avaliações mais consistentes ou tendenciosas, afetando as notas. ENCONTRAR ALGO QUE RELACIONE A INFLUENCIA DA MARCA DA CERVEJA NA NOTA DO CONSUMIDOR

Devido ao grande número de estilos de cerveja, a variável “beer\_style\_cluster” foi criada com o intuito de realizar uma clusterização dos cento e quatro (104) estilos de cerveja presentes na base de dados em sete grandes grupos, levando em consideração tanto o nome do estilo dada pela variável “beer\_style” quanto o teor alcoólico dado pela variável “beer\_abv”. Os agrupamentos criados foram: Pilsen, Lager, Bock, Malzbier, Ale, Stout, Porter e Weissbier, que são os estilos de cerveja reconhecidos pela ABRACERVA (Associação Brasileira de Cerveja Artesanal) que indica que mesmo as cervejas importadas devem ser caracterizadas nas opções acima com base no seu tipo. ABRACERVA (2019).

A criação destes grupos tem como finalidade a utilização do estilo da cerveja como variável explicativa, visando entender se o tipo de cerveja tem alguma influência na avaliação geral do consumidor.

**Treinamento de Modelos**

Os dados foram divididos em treino (80%) e teste (20%) utilizando a função train\_test\_split da biblioteca scikit learn. As variáveis preditoras foram normalizadas utilizando a função StandardScaler, também da biblioteca scikit learn.

Quatro modelos de machine learning foram treinados para prever a variável review\_overall, utilizando como variáveis explicativas as características sensoriais das cervejas e o teor alcoólico.

**Regressão Linear**

A Regressão Linear foi utilizada como modelo base para estabelecer um ponto de partida. Este modelo assume uma relação linear entre as variáveis explicativas e a variável alvo. O objetivo inicial era verificar se as características sensoriais e o teor alcoólico poderiam prever a avaliação geral de forma linear.

**Árvore de decisão**

A Árvore de Decisão foi selecionada por sua capacidade de capturar relações não lineares entre as variáveis. Este modelo constrói uma estrutura de árvore onde, a cada nó, uma condição binária é usada para dividir os dados em subconjuntos, tentando maximizar a pureza da variável alvo em cada divisão. Os principais hiperparâmetros adotados foram:

max\_depth:8 (define a profundidade máxima da árvore)

min\_samples\_split: 10 (número mínimo de amostrar para divisão de cada nó)

min\_samples\_leaf: 5 (número mínimo de amostras por folha)

random\_state: 36 (estado aleatório para garantir a reprodutibilidade do modelo)

**Random Forest**

O Random Forest, um modelo baseado em múltiplas árvores de decisão, foi treinado para melhorar a robustez das previsões e reduzir a variabilidade. O Random Forest cria várias árvores independentes e as combina para obter uma previsão média, reduzindo o risco de overfitting típico de uma única árvore. Os principais hiperparâmetros adotados foram:

n\_estimators: 100 (número de árvores na floresta)

max\_depth: 10 (define a profundidade máxima da árvore)

min\_samples\_split: 10 (número mínimo de amostrar para divisão de cada nó)

min\_samples\_leaf: 5 (número mínimo de amostras por folha)

random\_state: 36 (estado aleatório para garantir a reprodutibilidade do modelo)

**XGBoost**

O XGBoost (Extreme Gradient Boosting) foi escolhido como um dos modelos de aprendizado supervisionado devido à sua eficácia em lidar com dados complexos e não lineares. O XGBoost é um algoritmo baseado em árvores de decisão que utiliza o método de boosting, uma técnica de ensemble que combina várias árvores de decisão fracas para formar um modelo robusto. Ele é amplamente reconhecido por sua eficiência em problemas de regressão e classificação, além de ser escalável para grandes conjuntos de dados, o que se alinha bem com o tamanho da base de dados utilizada neste estudo, composto por mais de 1,5 milhão de avaliações de cervejas. Os principais hiperparâmetros adotados foram:

n\_estimators: 100 (número de árvores na floresta)

max\_depth: 10 (define a profundidade máxima da árvore)

learning\_rate:0,2 (taxa de aprendizado)

subsample: 0,8 (proporção de amostras usadas para treinar cada árvore)

colsample\_bytree: 0,8 (proporção de features usadas para cada árvore

random\_state: 36 (estado aleatório para garantir a reprodutibilidade do modelo)

**Validação Cruzada**

Para validar o desempenho dos modelos e garantir sua generalização, foi utilizada a técnica de validação cruzada k-fold, com k = 5. Isso significa que os dados de treinamento foram divididos em cinco subconjuntos (folds), e o modelo foi treinado e avaliado cinco vezes, utilizando cada fold uma vez como conjunto de teste e os outros quatro como conjunto de treinamento. Ao final, as métricas de avaliação foram calculadas pela média dos resultados obtidos em cada rodada.

**Métricas de Avaliação**

O desempenho dos foi avaliado utilizando quatro métricas principais: Erro Quadrático Médio (MSE), Erro Absoluto Médio (MAE), Raiz do Erro Quadrático Médio (RMSE) e o Coeficiente de Determinação (R²), onde cada métrica fornece uma perspectiva diferente sobre a precisão das previsões do modelo.

Erro Quadrático Médio (MSE): Mede a média dos quadrados das diferenças entre os valores previstos e os reais. Penaliza mais fortemente grandes erros, o que torna essa métrica sensível a outliers.

Erro Absoluto Médio (MAE): Mede a média das diferenças absolutas entre as previsões e os valores reais. O MAE é mais robusto contra outliers do que o MSE.

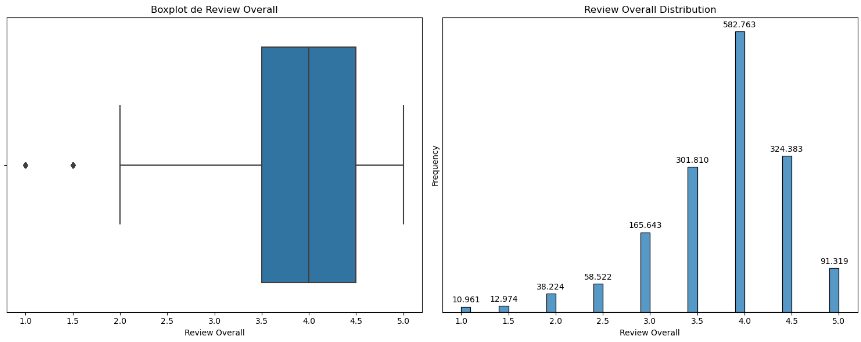
Raiz do Erro Quadrático Médio (RMSE): Representa a raiz quadrada do MSE, o que a torna uma métrica com a mesma unidade da variável de destino. Ela penaliza mais fortemente grandes erros, como o MSE, mas é mais interpretável.

Coeficiente de Determinação (R²): Avalia a proporção da variância dos dados que é explicada pelo modelo. Um valor de R² próximo de 1 indica que o modelo consegue explicar grande parte da variância da variável alvo.

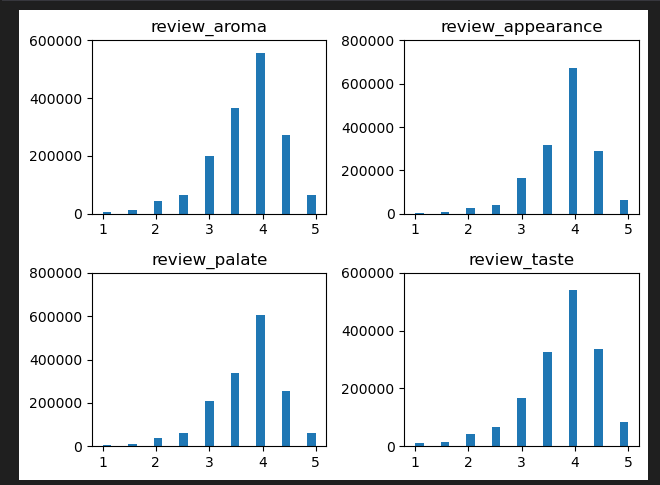
**Resultados Preliminares**

**Análise Exploratória de Dados**

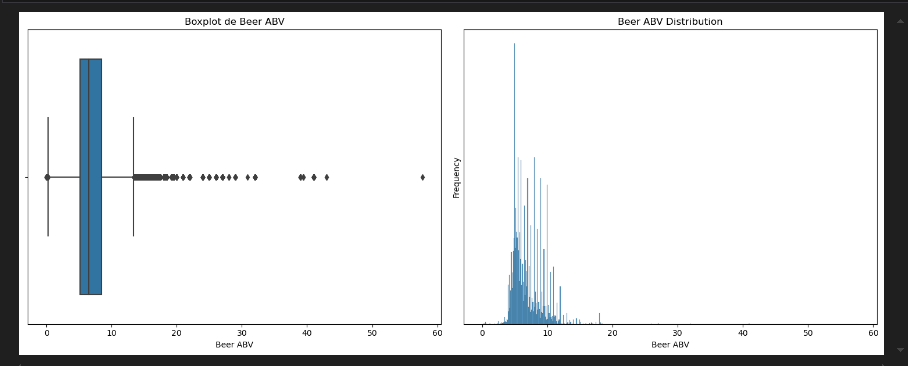
A análise uni variada das variáveis revelou algumas características importantes. Para a variável resposta “review\_overall”, que mede a nota geral da avaliação do consumidor, tida neste projeto como variável resposta, foram gerados histogramas e “boxplots”, que mostraram uma distribuição que, embora próxima da normal, apresentava uma cauda alongada à esquerda. Isso indica que a maioria das avaliações tende a ser positiva, mas há uma proporção menor de notas baixas.



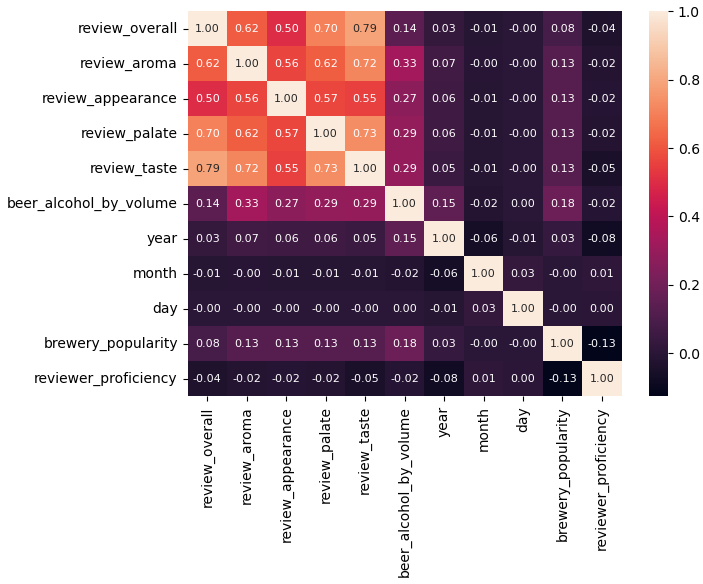
Para as variáveis sensoriais (review\_aroma, review\_taste, review\_palate, review\_appearance), foram observadas distribuições semelhantes, conforme figura abaixo:



Por outro lado, a variável beer\_abv mostrou uma distribuição bem alongada à direita, mostrando que o teor alcoólico tem um range maior do que as outras variáveis quantitativas.



Na análise multivariada, foi gerado um mapa de calor para analisar as correlações entre as variáveis numéricas. As variáveis review\_palate e review\_taste apresentaram as maiores correlações com a variável resposta review\_overall, com coeficientes de 0,79 e 0,72, respectivamente. Isso sugere que o paladar e o gosto da cerveja desempenham um papel crítico na percepção geral do consumidor.



**Seleção de Variáveis**

Para o desenvolvimento do primeiro modelo, as variáveis selecionadas foram: beer\_abv, review\_aroma, review\_taste, review\_palate e review\_appearance, por apresentarem maior relevância na análise exploratória.

Os primeiros resultados obtidos ao aplicar os modelos de machine learning no conjunto de dados de avaliações de cervejas indicam alguns padrões importantes e áreas de melhoria a serem trabalhadas nas próximas etapas. Estes resultados refletem o comportamento dos modelos em relação às variáveis explicativas e à previsão da variável target review\_overall, que varia de 1 a 5, em incrementos de 0,5. A distribuição da variável alvo revelou uma concentração significativa de avaliações em torno de valores mais elevados. especificamente, a distribuição é dada da seguinte maneira:

|  |  |
| --- | --- |
| Review Overall | Observações |
| 4,0 | 582.763 |
| 4,5 | 324.383 |
| 3,5 | 301.810 |
| 3,0 | 165.643 |
| 5,0 | 91.319 |
| 2,5 | 58.522 |
| 2,0 | 38.224 |
| 1,5 | 12.974 |
| 1,0 | 10.961 |

Essa distribuição desbalanceada nos dados sugere que a maioria das cervejas avaliadas são vistas de forma bastante positiva pelos consumidores. Esse fator deve ser levado em consideração na avaliação de métricas de erro dos modelos, pois os modelos podem tender a prever valores próximos de 4.0 com maior frequência, dada a prevalência dessas notas na base de dados.

Os modelos de regressão testados, incluindo Regressão Linear, Árvore de Decisão, Random Forest e XGBoost foram avaliados inicialmente com validação cruzada em 5 k-folds. O desempenho foi avaliado usando as métricas descritas na seção pertinente e as primeiras análises trouxeram insights importantes:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Modelo | MAE | MSE | RMSE | R² |
| Regressão Linear | 0.31 | 0,17 | 0,41 | 0,67 |
| Árvore de Decisão | 0,31 | 0,17 | 0,41 | 0,68 |
| Random Forest | 0,31 | 0,17 | 0,41 | 0,68 |
| XGBoost | 0,31 | 0,16 | 0,40 | 0,68 |

Esses resultados mostram uma consistência nas métricas, especialmente após o processo de validação cruzada, com uma leve melhora no desempenho geral dos modelos. O XGBoost foi o que apresentou o melhor ajuste com menor MSE e RMSE, indicando que o modelo conseguiu capturar melhor os padrões dos dados, mesmo com a complexidade e volume do banco de dados.

As variáveis explicativas utilizadas até o momento, incluindo review\_aroma, review\_taste, review\_palate, review\_appearance e beer\_abv, mostraram desempenhos distintos em termos de contribuição para a previsão da avaliação geral. Análises preliminares de importância de variáveis indicam que review\_taste e review\_aroma têm maior impacto na previsão da review\_overall, sugerindo que os consumidores valorizam fortemente essas características sensoriais na hora de avaliar a cerveja.

No entanto, outras variáveis como review\_palate e beer\_abv mostraram uma correlação mais fraca com a variável target, podendo ser menos influentes no modelo final. Ainda assim, a interação entre essas variáveis e o desempenho geral dos modelos requer mais investigação.

Para melhor compreender a relação entre as variáveis, foram gerados gráficos como scatter plots e boxplots, permitindo visualizar as correlações entre as características sensoriais e a avaliação geral das cervejas. Foi possível observar que cervejas com avaliações mais altas, geralmente, têm notas muito consistentes em todas as variáveis sensoriais, ou seja, há uma alta correlação positiva entre as notas de aroma, sabor, paladar e aparência para as melhores cervejas. Nas cervejas com notas mais baixas, existe uma discrepância maior entre as avaliações de diferentes características sensoriais, indicando que consumidores tendem a penalizar fortemente uma única característica ruim, mesmo que outras sejam avaliadas de forma mais positiva.

Até o momento, um dos principais desafios encontrados é o impacto do desbalanceamento da variável target. Modelos como a Árvore de Decisão tendem a favorecer as notas mais frequentes, resultando em previsões que frequentemente se aproximam de 4, o que leva a um viés positivo.

Os próximos passos envolvem ajustes nos modelos para tentar melhorar a previsão, especialmente para as notas mais baixas. Algumas estratégias planejadas incluem técnicas de balanceamento de dados, como SMOTE (Synthetic Minority Over-sampling Technique) ou undersampling, serão testadas para mitigar o problema do desbalanceamento na variável target. Será efetuada a análise de hiperparâmetros através de Grid Search para otimizar o desempenho dos modelos, principalmente na árvore de decisão, que apresentou variações significativas em suas previsões.

Em suma, os resultados preliminares forneceram uma base sólida para o entendimento do comportamento do modelo, mas evidenciaram desafios claros relacionados ao desbalanceamento da variável target e à necessidade de ajustes nos parâmetros dos modelos para melhorar a generalização das previsões. O processo de otimização e experimentação com diferentes técnicas seguirá nas próximas etapas.

**REFERENCIA**

Ministério da Agricultura, Pecuária e Abastecimento [MAPA], 2022. Setor cervejeiro segue crescendo a cada ano, aponta anuário. Disponível em: https://www.gov.br/agricultura/pt-br/assuntos/noticias/setor-cervejeiro-segue-crescendo-a-cada-ano-aponta-anuario..Acesso em: 17 de junho de 2024

CARVALHO,. Impacto do aumento do número de cervejarias artesanais sobre o mercado de cervejas comerciais nacional. **Repositorio.ufu.br**, 2022

Descobrir LIMA (2023) - <https://repositorio.ufu.br/bitstream/123456789/39812/1/CiênciaDadosAprendizado.pdf>

ARAÚJO, F. B.; SILVA, P. H. A.; MINIM, V. P. R. Perfil sensorial e composição físico-química de cervejas provenientes de dois segmentos do mercado brasileiro. **Ciência e Tecnologia de Alimentos**, v. 23, n. 2, p. 121–128, ago. 2003.

KNIAZIEVA, Y. **Guide to Movie Recommendation Systems Using Machine Learning**. Disponível em: <https://labelyourdata.com/articles/movie-recommendation-with-machine-learning>.