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

Alexandre Polizelli Marino, Wagner W. Bortoletto

**Previsão de Avaliação Geral de Cervejas com Base em Características Sensoriais Usando Técnicas de 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. 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.

Este 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. O uso de modelos preditivos, como regressão linear, árvores de decisão, e algoritmos de ensemble como Random Forest e XGBoost, oferece uma abordagem robusta para investigar esse fenômeno.

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. 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. Vários estudos na área de análise de sentimentos e machine learning têm demonstrado a importância de utilizar dados sensoriais para modelar a percepção de qualidade. No entanto, a complexidade inerente à interação entre esses fatores sensoriais e variáveis adicionais, como o teor alcoólico e o tempo de avaliação, torna esse um problema ideal para a aplicação de técnicas preditivas.

Estudos anteriores sobre previsão de preferências de consumidores já demonstraram a eficácia de técnicas de machine learning em domínios como recomendação de filmes, previsão de sucesso de músicas e até mesmo avaliações de vinhos. 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 preencher essa lacuna ao 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.

A relevância deste estudo está ancorada em duas frentes principais. Primeiro, do ponto de vista acadêmico, ele contribui para o campo de previsão de preferências de consumidores, com foco em bebidas, um setor que apresenta variações complexas de qualidade percebida. Segundo, do ponto de vista prático, os resultados podem beneficiar diretamente cervejarias e distribuidores, que poderiam ajustar suas produções com base nas características 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, com variáveis como avaliação sensorial (aroma, paladar, gosto, aparência), teor alcoólico, nome da cervejaria e estilo de cerveja. 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 utilizado 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\_alcohol\_by\_volume*: Foi detectado 67.785 valores em branco, que foram preenchidos com a média do teor alcoólico por estilo de cerveja. 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. Foi optada a atribuição da nota mínima permitida para essas observações a fim de manter a integridade dos dados e refletir a experiência do consumidor ao avaliar visualmente a cerveja.

brewery\_name: A variável tinha nove observações ausentes as quais foram removidas.

***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.

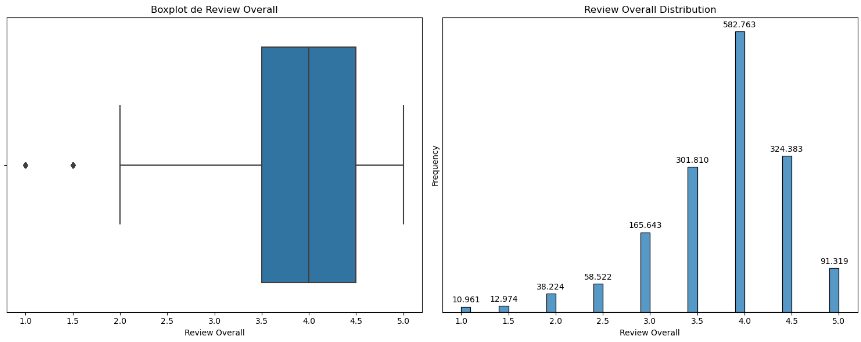
Uma variável que mede a experiência do avaliador chamada de *“reviewer\_proficiency”* foi obtida com base no número de avaliações que algum consumidor fez. A expectativa é que consumidores mais experientes sejam mais críticos e forneçam notas diferentes dos novatos.

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\_alcohol\_by\_volume*. Os agrupamentos criados foram*: Pale Ale, IPA, Stout, Weizenbier, Pilsner, Helles* e *Dunkel*. A criação destes grupos tem como finalidade a utilização do estilo da cerveja como variável explicativa.

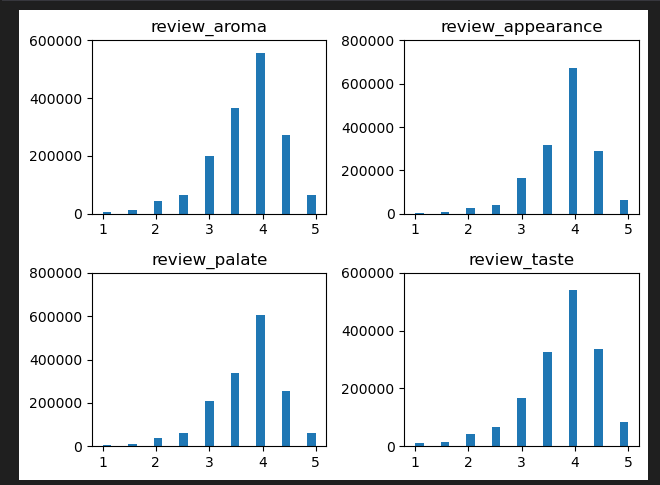
A variável *review\_time* foi convertida em múltiplos componentes temporais, como ano, mês, trimestre e semestre. Isso permitiu uma análise temporal mais granular, identificando tendências e sazonalidades nas avaliações, que podem influenciar a variável resposta.

**Análise Exploratória de Dados**

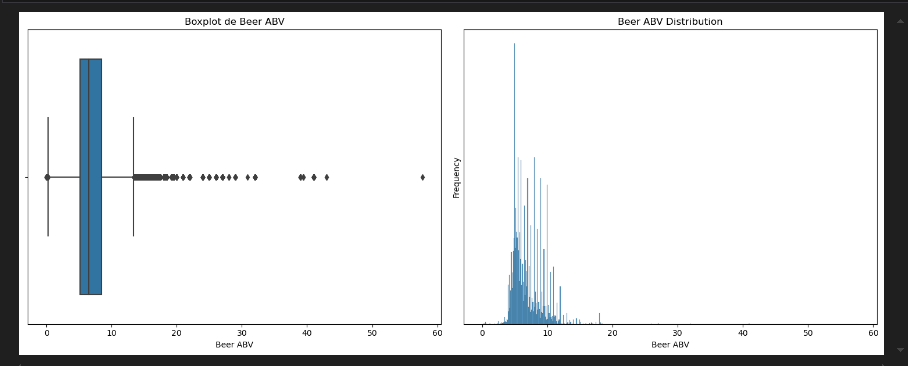
A análise uni variada das variáveis revelou algumas características importantes. Para a variável resposta *review\_overall*, foram gerados histogramas e boxplots, que mostraram uma distribuição que, embora próxima da normal, apresentava uma cauda alongada à esquerda. Isso indica que a maioria das avaliações tende a ser positiva, mas há uma proporção menor de notas baixas.



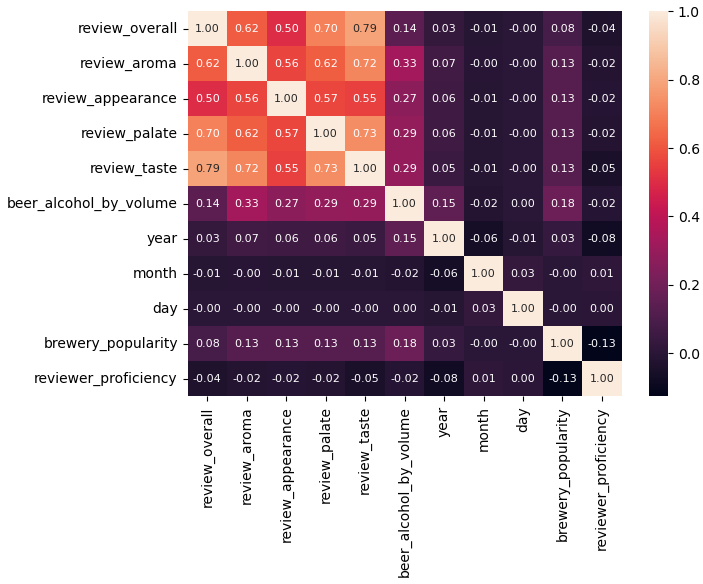
Para as variáveis sensoriais (*review\_aroma, review\_taste, review\_palate, review\_appearance*), foram observadas distribuições semelhantes, reforçando que essas variáveis têm um comportamento consistente e podem ser preditoras importantes da nota geral da cerveja.



Por outro lado, a variável *beer\_alcohol\_by\_volume* 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\_alcohol\_by\_volume, review\_aroma, review\_taste, review\_palate e review\_appearance, por apresentarem maior relevância na análise exploratória.

**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**

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\_alcohol\_by\_volume*, 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\_alcohol\_by\_volume* 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.