**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)**

**Título em inglês** **ou espanhol** (opcional)

**Abstract** ou **Resumen** (opcional)

**Keywords** ou **Palabras Clave:** (opcional)

**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).

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 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 sensoriais envolvidas nesse processo de avaliação.

Artigos 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 a regressão logística multinominal para entender como algumas variáveis sensoriais, como textura, gosto, aroma, aparência e se teor alcoólico e tipo de cerveja influenciam na nota geral de uma cerveja do ponto de vista do consumidor.

Dado tamanho e potencial crescimento do mercado, MAPA (2023) 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.

A relevância deste estudo está em demonstrar qual das características é a mais relevante na percepção geral do consumidor do ponto de vista da influência na avaliação final. 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.

Este projeto busca explorar uma base de dados contendo mais de 1,5 milhões de avaliações de cervejas realizadas por consumidores. O dataset foi disponibilizado pela empresa BeerAdvocate, dos Estados Unidos e contém percepções de mais de 10 anos no que tange à avaliação de cervejas produzidas em mais 5.800 cervejarias diferentes, trazendo uma ampla variedade de observações.

**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\_style | grupo\_estilo\_cerveja | Agrupamento dos estilos de 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>.

O desenvolvimento deste projeto evidenciou a importância de uma abordagem meticulosa e sistemática ao se trabalhar com modelos multiclasse, assegurando que cada etapa, desde o pré-processamento até a avaliação do modelo, fosse realizada com rigor e atenção aos detalhes, contribuindo assim para a eficácia do modelo preditivo.

**Pacotes Utilizados**

Foram utilizadas 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.

**Pré-processamento dos dados**

Para garantir a robustez e a validade dos resultados, foram adotados cuidados específicos durante o processo de modelagem. Primeiramente, o pré-processamento dos dados se tornou uma etapa fundamental. Tratou-se de garantir que não houvesse valores ausentes ou inconsistentes nas variáveis, uma vez que isso poderia afetar a qualidade do modelo preditivo. Segundo Little e Rubin (2002), a manipulação adequada de dados ausentes é essencial para evitar viés nas análises e nas inferências feitas a partir dos modelos

**Transformação de variáveis explicativas de contínuas para discretas**

A transformação de variáveis preditoras e dependentes em uma escala numérica discreta foi fundamental para o treinamento de modelos de regressão logística ordinal, pois essa abordagem permitiu que as relações entre variáveis fossem capturadas e analisadas de maneira mais eficaz. As variáveis categóricas, quando convertidas em valores numéricos, como variáveis dummy ou utilizando métodos de codificação ordinais, possibilitaram a inclusão adequada de informações sobre a ordem das categorias na modelagem. De acordo com Agresti (2010), a conversão de variáveis categóricas em formas numéricas facilitou a estimativa de probabilidades acumuladas e a interpretação dos parâmetros do modelo. Além disso, McCullagh (1980) destacou que essa transformação foi essencial para garantir que os modelos refletissem corretamente a estrutura ordinal dos dados, permitindo que as transições entre categorias fossem analisadas de maneira significativa. A utilização de escalas numéricas discretas não apenas melhorou a precisão da modelagem, mas também assegurou que os pressupostos estatísticos subjacentes da regressão logística ordinal fossem atendidos, conforme discutido por Long (1997). Assim, a transformação adequada das variáveis explicativas foi uma etapa crucial para garantir a validade e a interpretabilidade dos resultados obtidos por meio de modelos de regressão logística ordinal. O método utilizado neste projeto foi de multiplicação por 2 que fez com que as variáveis mantivessem sua ordem original e perdessem a condição as caracterizava como contínua.

**Separação Treino e Teste**

A divisão dos dados em conjuntos de treino e teste também foi uma prática recomendada. Essa separação é fundamental para garantir que o modelo seja avaliado em dados não vistos, evitando o “overfitting” que é descrito como um sobre ajuste do modelo aos dados de treinamento, que é quando modelo lida muito bem com os dados de teste, mas não nos dados de treino, indicando a falta de capacidade de generalização do modelo, evidenciada pela falta de capacidade de efetuar predição em novos dados, conforme mencionado por Gulob (2017). Kohavi (1995) destaca que a validação cruzada é uma técnica eficaz para obter estimativas mais confiáveis da performance do modelo, permitindo uma avaliação mais rigorosa do seu desempenho, por isso os dados foram divididos em treino (80%) e teste (20%) utilizando a função “train\_test\_split” da biblioteca “scikit-learn”.

**Validação Cruzada**

A validação cruzada é uma técnica essencial para garantir que o modelo preditivo não apenas se ajusta bem aos dados de treino, mas também generaliza adequadamente para dados não vistos. A prática de dividir os dados em conjuntos de treino e teste é crucial para evitar o overfitting e fornecer estimativas mais confiáveis sobre o desempenho do modelo. Kohavi (1995) discute como a validação cruzada pode ser usada para obter uma avaliação rigorosa da performance do modelo, permitindo que o mesmo seja testado em múltiplas subdivisões dos dados. A utilização dessa técnica ajuda a garantir que as conclusões tiradas a partir dos resultados do modelo sejam robustas e confiáveis, refletindo o desempenho esperado em situações do mundo real.

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.

**Regressão Logística Ordinal**

No desenvolvimento de modelos preditivos, a escolha das variáveis e o tratamento adequado dos dados desempenham um papel crucial na eficácia do modelo. Neste contexto, foram utilizadas quatro variáveis sensoriais, a saber, review\_palate, review\_taste, review\_appearance e review\_aroma caracterizadas como variáveis discretas para prever a variável de interesse review\_overall, que se caracteriza como uma variável categórica multiclasse, que atua de forma ordenada, caracterizando a utilização de modelos de regressão logística ordinal.

A regressão logística ordinal é uma extensão da regressão logística que lida com variáveis dependentes categóricas que possuem uma ordem natural. O modelo é particularmente útil quando a variável resposta tem mais de duas categorias ordenadas e calcula a probabilidade cumulativa de uma observação pertencer a uma determinada categoria ou a uma categoria inferior, ou seja, o modelo estima a probabilidade de uma observação pertencer a uma categoria menor ou igual a um determinado valor. Segundo McCullagh (1980), o modelo ordinal trabalha com o pressuposto de que as categorias podem ser ordenadas e que os efeitos das variáveis preditoras são constantes para essas categorias, em um processo conhecido como assunção de “odds” proporcionais, desde que teste de Brant, valide esta hipótese.

A formula da regressão logística ordinal é dada por:

em que: P: é a probabilidade de ocorrência da categoria; y: é a variável dependente ordinal; X: são as variáveis independentes; α: é o intercepto; β1, β2, ..., βk são os coeficientes para as variáveis independentes.

McCullagh (1980) indica que os α’s são chamados de pontos de corte ou interceptos, e definem as fronteiras entre as categorias. Eles representam o logito das probabilidades acumuladas de passar de uma categoria para outra. Cada α especifica o ponto em que a probabilidade de pertencer a uma categoria ou inferior se altera. Em termos de interpretação, um maior o valor de α implica que é necessário ter um valor maior na combinação das variáveis explicativas para que a probabilidade de estar na categoria superior aumente. Isso sugere que a transição para categorias mais altas é mais difícil, ou seja, requer condições mais favoráveis nas variáveis preditoras para ser alcançada, conforme elencado por Williams, R. (2006).

O coeficiente β representa o efeito multiplicativo das variáveis explicativas nas “odds” cumulativas, conforme elencado por McCullagh (1980). Um β positivo aumenta a chance de a resposta estar em uma categoria superior, enquanto um β negativo aumenta a chance de a resposta estar em uma categoria inferior.

O logito que é a transformação logarítmica da razão de chances (odds) que descreve a relação entre as variáveis independentes X e probabilidade cumulativa de y, que está expresso no termo que dentro da elevação exponencial.

**O problema de multicolinearidade**

As variáveis explicativas podem ser contínuas ou categóricas. No caso de variáveis categóricas, como sugerido por Hosmer (2013), estas são geralmente transformadas em variáveis dummy para que possam ser inseridas no modelo, processo que foi feito com as variáveis sensoriais review\_taste, review\_palate, review\_aroma e review\_appearence. Todavia, a conversão das variáveis explicativas em dummies fez com que 36 colunas fossem adicionadas ao dataset o que pode trazer um problema de multicolinearidade, que segundo Akinyemi, Ogunyemi e Adeniran (2019) modelos treinados sem a verificação da multicolinearidade pode afetar a precisão dos estimadores e a interpretação dos resultados.

A multicolinearidade pode afetar a interpretação e a precisão dos coeficientes em modelos de regressão logística, uma vez que a presença de multicolinearidade pode levar a estimativas instáveis dos parâmetros, tornando difícil determinar a verdadeira relação entre as variáveis independentes e a variável dependente, conforme Wiggins & Hurst (2019). Os autores recomendaram o uso de diagnósticos estatísticos, como o VIF (Variance Inflation Factor), para identificar e quantificar a multicolinearidade e sugeriram também técnicas de regularização, como Ridge e Lasso, como formas eficazes de lidar com a multicolinearidade, melhorando a robustez e a interpretabilidade dos modelos.

Por outro lado, Faraway (2006) destacou que a multicolinearidade não afeta a previsão do modelo, mas complica a interpretação dos coeficientes, podendo levar a conclusões errôneas sobre a significância das variáveis. Faraway enfatizou a importância de identificar variáveis altamente correlacionadas e sugeriu que a exclusão de algumas variáveis ou a aplicação de métodos de redução de dimensionalidade, como a análise de componentes principais (PCA), pode ser útil para mitigar os problemas associados à multicolinearidade. O artigo também mencionou que a multicolinearidade é mais comum em conjuntos de dados que incluem variáveis dummies.

Finalmente, Jöreskog (2005) apresenta um estudo de caso que ilustra o impacto da multicolinearidade em modelos de regressão logística ordinal. O autor concluiu que a presença de multicolinearidade pode distorcer as estimativas dos parâmetros, dificultando a avaliação da significância estatística das variáveis independentes. O estudo destacou a necessidade de realizar uma análise cuidadosa da correlação entre variáveis explicativas antes da modelagem. O autor sugeriu que a escolha de uma categoria de referência apropriada e a aplicação de métodos de regularização podem ajudar a reduzir os efeitos da multicolinearidade, permitindo uma melhor interpretação dos resultados em contextos de análise de dados categóricos.

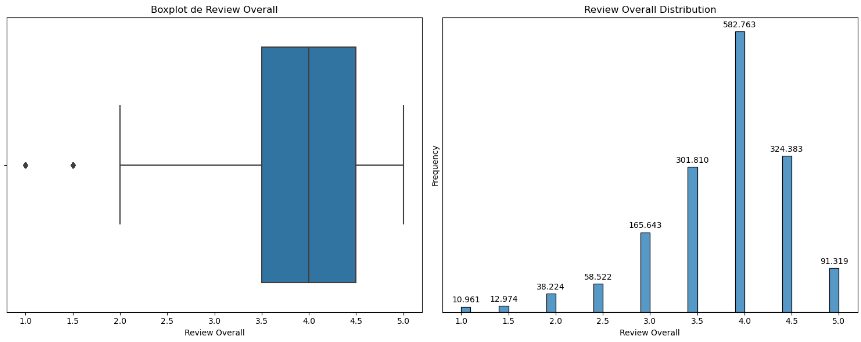
**Métricas de Avaliação**

Finalmente, a interpretação dos resultados requer atenção especial. Menard (2002) enfatiza a importância de compreender os coeficientes do modelo, pois eles fornecem insights sobre a relação entre as variáveis preditoras e a variável de resposta. O uso de métricas de avaliação adequadas, como precisão, F1-score e acurácia, é essencial para mensurar o desempenho do modelo de forma eficaz, conforme discutido por Sokolova e Lapalme (2009).

**Resultados e Discussão**

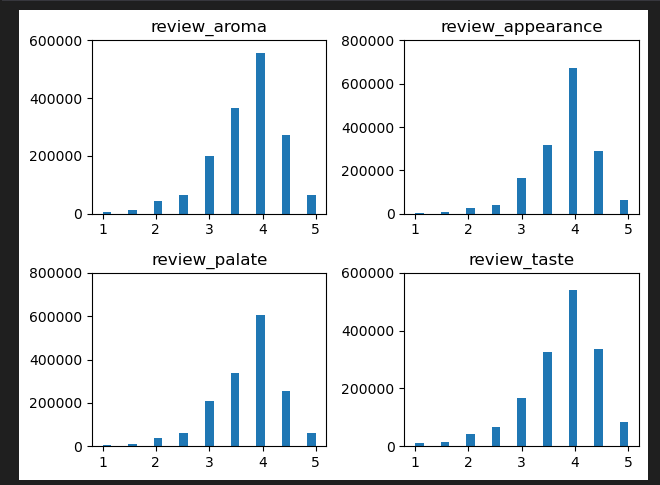
**Análise Exploratória de Dados**

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.



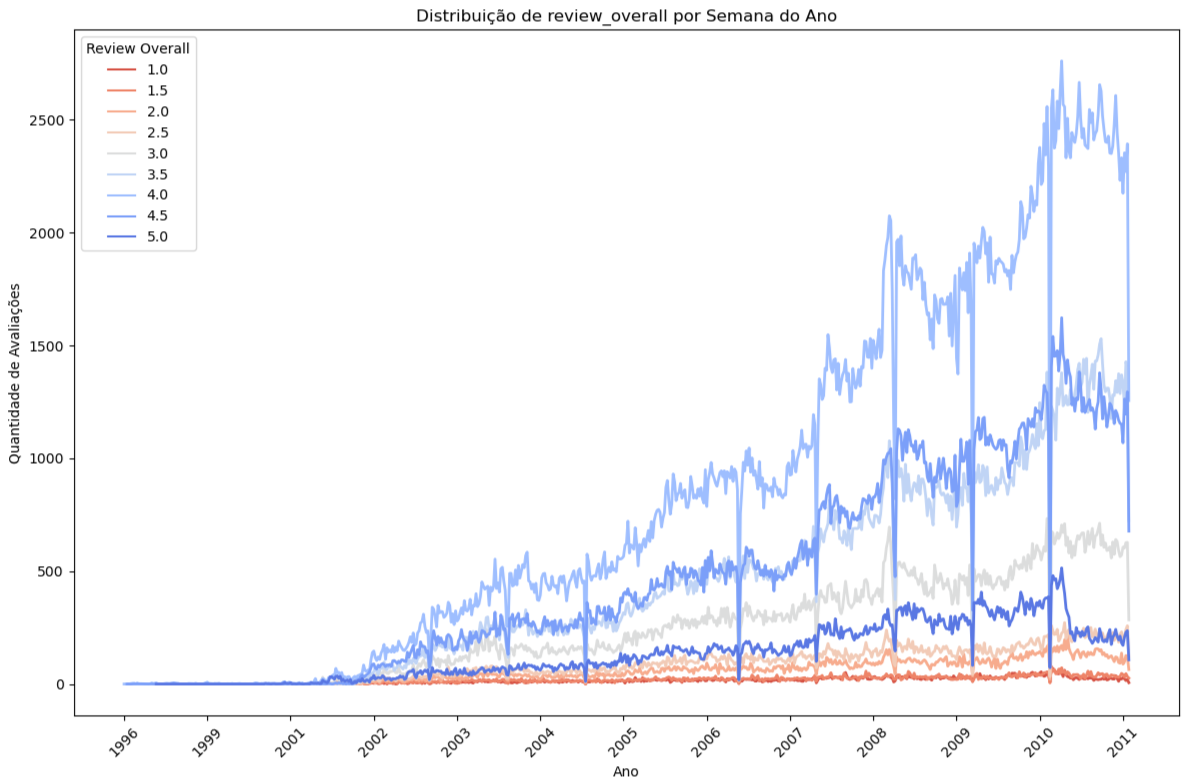
Fonte: Resultados originais da pesquisa

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



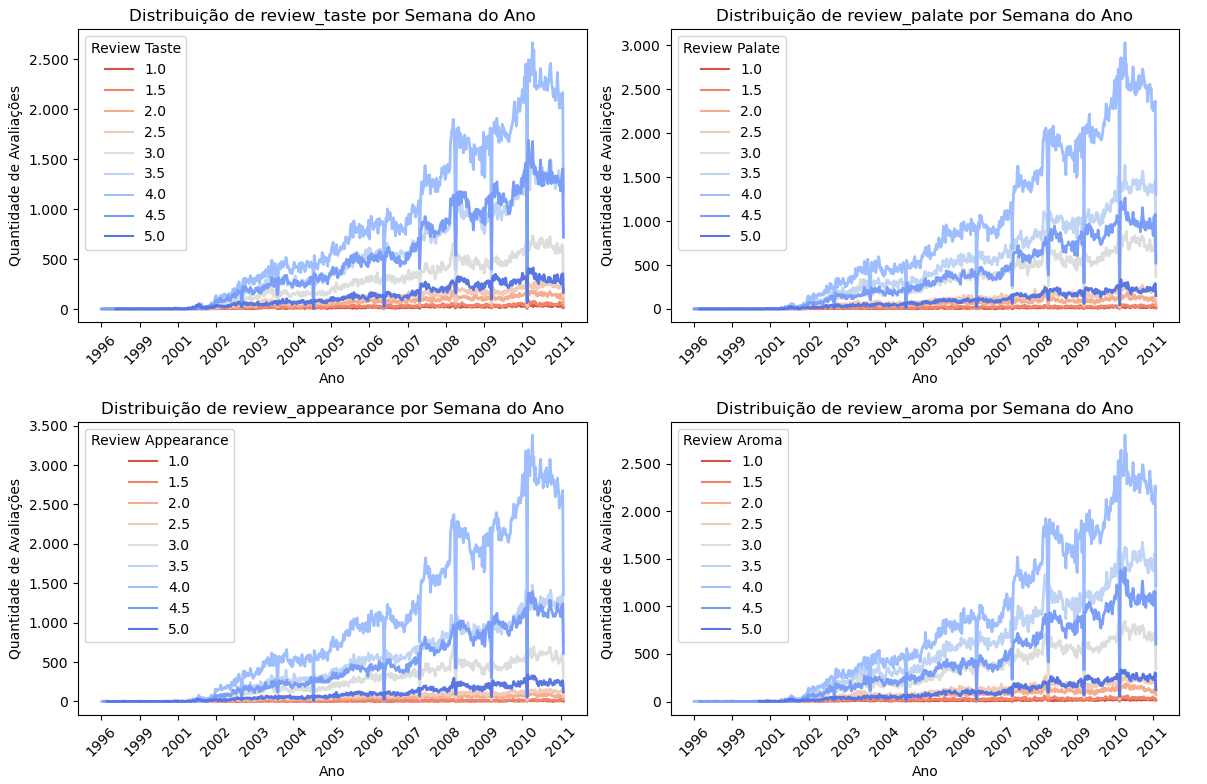
Fonte: Resultados originais da pesquisa

A análise bivariada foi elaborada a fim de entender a variação das variáveis sensoriais e da variável reposta ao longo do tempo. Os gráficos abaixo representam como cada uma das categorias de cada variável se distribui ao longo de cada ano e mês da amostra:



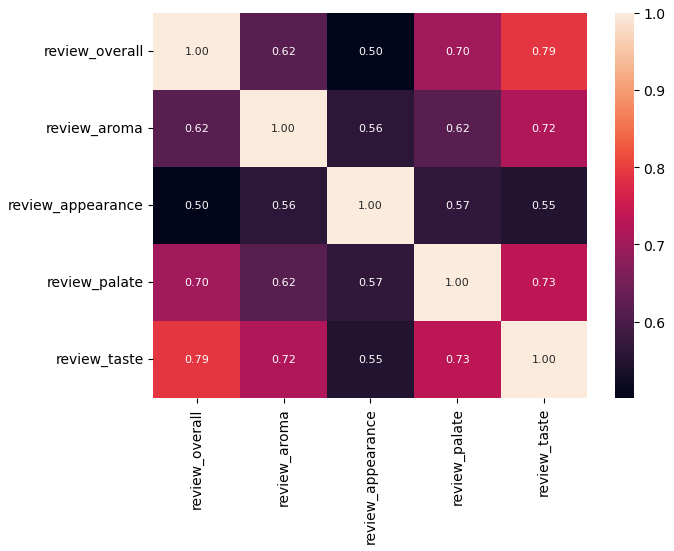
Fonte: Resultados originais da pesquisa

O gráfico acima indica que há uma certa proporção ao ser mantida em relação às notas, é possível perceber que cada uma das notas segue um certo padrão de distribuição, onde é mantida uma ordem de aparições para cada uma das notas independente da data de avaliação e que isso foi mantido por cerca de dez anos, o que pode resultar em uma predisposição de avaliação que se sobressai as características da cerveja em questão. Isso é ainda mais evidenciado quando olhamos a distribuição das variáveis explicativas, conforme figura abaixo:



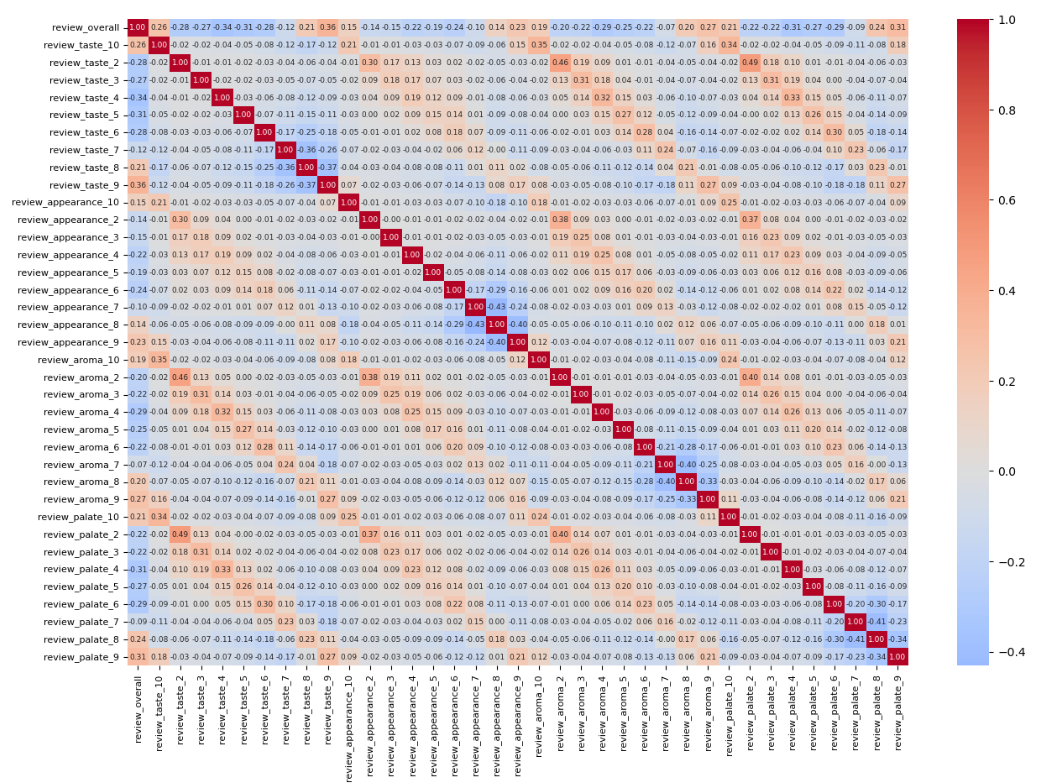
Fonte: Resultados originais da pesquisa

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.



Fonte: Resultados originais da pesquisa

Após a “dummização” das variáveis explicativas, um novo diagrama de correlação foi gerado conforme figura abaixo:



Fonte: Resultados originais da pesquisa

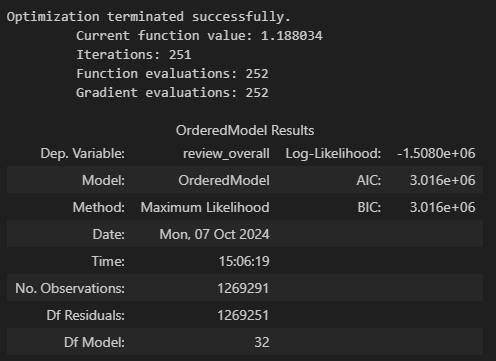
O gráfico acima sugere que não há muticolinearidade pois a maior correlação indicada na figura é de 0,43. Ao efetuar o cálculo do VIF, os valores são mostrados abaixo, o que sugere que não existe uma multicolinearidade entre as variáveis explicativas.

Tabela X. Cálculo do VIF para as quatro variáveis sensoriais

|  |  |  |
| --- | --- | --- |
|  | (continua) | |
| variável explicativa | VIF |
| review\_aroma\_2.0 | 0.06 | |
| review\_aroma\_3.0 | 0.02 | |
| review\_aroma\_4.0 | 0.01 | |
|  | (conclusão) | |
| variável explicativa | VIF | |
| review\_aroma\_5.0 | 0.02 | |
| review\_aroma\_6.0 | 0.00 | |
| review\_aroma\_7.0 | 0.02 | |
| review\_aroma\_8.0 | 0.02 | |
| review\_aroma\_9.0 | 0.03 | |
| review\_appearance\_2.0 | 0.00 | |
| review\_appearance\_3.0 | 0.02 | |
| review\_appearance\_4.0 | 0.02 | |
| review\_appearance\_5.0 | 0.03 | |
| review\_appearance\_6.0 | 0.02 | |
| review\_appearance\_7.0 | 0.03 | |
| review\_appearance\_8.0 | 0.03 | |
| review\_appearance\_9.0 | 0.03 | |
| review\_palate\_2.0 | 0.01 | |
| review\_palate\_3.0 | 0.04 | |
| review\_palate\_4.0 | 0.03 | |
| review\_palate\_5.0 | 0.02 | |
| review\_palate\_6.0 | 0.04 | |
| review\_palate\_7.0 | 0.06 | |
| review\_palate\_8.0 | 0.04 | |
| review\_palate\_9.0 | 0.06 | |
| review\_taste\_2.0 | 0.00 | |
| review\_taste\_3.0 | 0.00 | |
| review\_taste\_4.0 | 0.04 | |
| review\_taste\_5.0 | 0.04 | |
| review\_taste\_6.0 | 0.03 | |
| review\_taste\_7.0 | 0.04 | |
| review\_taste\_8.0 | 0.01 | |
| review\_taste\_9.0 | 0.00 | |

Fonte: Resultados originais da pesquisa

Após confirmação de não presença de multicolinearidade, o modelo de regressão logística ordinal, que utiliza a função “OrderedModel” da biblioteca “statsmodels.miscmodels.ordinal\_model”, os resultados foram os seguintes:



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| var. expl. / pto. corte | coef | std err | z | P>|z| | [0.025 | 0.975] |
| review\_aroma\_2.0 | 0.4316 | 0.037 | 11.561 | 0.000 | 0.358 | 0.505 |
| review\_aroma\_3.0 | 0.5130 | 0.035 | 14.527 | 0.000 | 0.444 | 0.582 |
| review\_aroma\_4.0 | 0.5411 | 0.036 | 15.147 | 0.000 | 0.471 | 0.611 |
| review\_aroma\_5.0 | 0.6366 | 0.035 | 17.942 | 0.000 | 0.567 | 0.706 |
| review\_aroma\_6.0 | 0.6777 | 0.036 | 19.062 | 0.000 | 0.608 | 0.747 |
| review\_aroma\_7.0 | 0.8106 | 0.036 | 22.767 | 0.000 | 0.741 | 0.880 |
| review\_aroma\_8.0 | 10.095 | 0.036 | 28.177 | 0.000 | 0.939 | 1.080 |
| review\_aroma\_9.0 | 13.436 | 0.037 | 36.460 | 0.000 | 1.271 | 1.416 |
| review\_appearance\_2.0 | 0.2549 | 0.055 | 4.616 | 0.000 | 0.147 | 0.363 |
| review\_appearance\_3.0 | 0.0306 | 0.051 | 0.599 | 0.549 | -0.070 | 0.131 |
| review\_appearance\_4.0 | -0.0748 | 0.051 | -1.461 | 0.144 | -0.175 | 0.026 |
| review\_appearance\_5.0 | -0.1025 | 0.051 | -2.027 | 0.043 | -0.202 | -0.003 |
| review\_appearance\_6.0 | -0.0945 | 0.051 | -1.868 | 0.062 | -0.194 | 0.005 |
| review\_appearance\_7.0 | 0.0345 | 0.051 | 0.683 | 0.495 | -0.065 | 0.134 |
| review\_appearance\_8.0 | 0.2284 | 0.051 | 4.503 | 0.000 | 0.129 | 0.328 |
| review\_appearance\_9.0 | 0.5182 | 0.051 | 10.087 | 0.000 | 0.418 | 0.619 |
| review\_palate\_2.0 | 0.5728 | 0.038 | 14.936 | 0.000 | 0.498 | 0.648 |
| review\_palate\_3.0 | 0.9800 | 0.036 | 27.247 | 0.000 | 0.910 | 1.051 |
| review\_palate\_4.0 | 14.042 | 0.036 | 38.491 | 0.000 | 1.333 | 1.476 |
| review\_palate\_5.0 | 18.649 | 0.036 | 51.507 | 0.000 | 1.794 | 1.936 |
| review\_palate\_6.0 | 24.375 | 0.036 | 66.974 | 0.000 | 2.366 | 2.509 |
| review\_palate\_7.0 | 30.918 | 0.036 | 84.800 | 0.000 | 3.020 | 3.163 |
| review\_palate\_8.0 | 37.466 | 0.037 | 101.969 | 0.000 | 3.675 | 3.819 |
| review\_palate\_9.0 | 43.734 | 0.038 | 115.631 | 0.000 | 4.299 | 4.448 |
| review\_taste\_2.0 | 16.163 | 0.032 | 50.344 | 0.000 | 1.553 | 1.679 |
| review\_taste\_3.0 | 31.159 | 0.032 | 96.261 | 0.000 | 3.052 | 3.179 |
| review\_taste\_4.0 | 44.572 | 0.033 | 133.358 | 0.000 | 4.392 | 4.523 |
| review\_taste\_5.0 | 59.202 | 0.034 | 175.816 | 0.000 | 5.854 | 5.986 |
| review\_taste\_6.0 | 72.222 | 0.034 | 212.865 | 0.000 | 7.156 | 7.289 |
| review\_taste\_7.0 | 85.309 | 0.034 | 250.102 | 0.000 | 8.464 | 8.598 |
| review\_taste\_8.0 | 95.145 | 0.034 | 276.743 | 0.000 | 9.447 | 9.582 |
| review\_taste\_9.0 | 105.872 | 0.035 | 298.967 | 0.000 | 10.518 | 10.657 |
| 1/2 | 15.805 | 0.051 | 30.793 | 0.000 | 1.480 | 1.681 |
| 2/3 | 0.4078 | 0.010 | 42.865 | 0.000 | 0.389 | 0.426 |
| 3/4 | 0.7234 | 0.006 | 129.114 | 0.000 | 0.712 | 0.734 |
| 4/5 | 0.4341 | 0.004 | 97.858 | 0.000 | 0.425 | 0.443 |
| 5/6 | 0.7650 | 0.003 | 285.304 | 0.000 | 0.760 | 0.770 |
| 6/7 | 0.7017 | 0.002 | 362.602 | 0.000 | 0.698 | 0.706 |
| 7/8 | 10.017 | 0.001 | 729.065 | 0.000 | 0.999 | 1.004 |
| 8/9 | 0.8983 | 0.002 | 467.523 | 0.000 | 0.895 | 0.902 |

Fonte: Resultados Originais da Pesquisa

A acurácia do modelo acima resultou em 51,1% nos dados de teste, outras métricas serão avaliadas posteriormente.

**Próximos Passos**

IANNARIO e PICCOLO (2016) discutem o desbalanceamento da variável resposta e sugerem que modelos desbalanceados, ou seja, com observações presentes em determinadas categorias são maiores do que as demais e sugerem como alternativa a utilização do SMOTE (Synthetic Minority Oversampling Technique) que é uma técnica de oversampling que fabrica dados para as classes minoritárias, ajudando a equilibrar a distribuição da variável resposta.

Além disso, será efetuada a etapa de validação cruzada, para garantir que o modelo não esteja enviesado, conforme seção mostrada em materiais e métodos.

Outro tópico a ser verificado, são as variáveis dummy que de acordo com o teste Z, não são estatisticamente significantes para a variável resposta, cuja probabilidade é superior a 0,05, será testado

Além disso outras métricas serão avaliadas, como F1 Score, área sob curva ROC e recall.

**Conclusão(ões) ou Considerações Finais**

**Agradecimento** (opcional, 1 parágrafo, bem sucinto)

**Referências**

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. Food Science and Technology, v. 23, n. 2, p. 121–128, 2003.

BENDER, R.; GROUVEN, U. Ordinal logistic regression in medical research. Journal of the Royal College of Physicians of London, v. 31, n. 5, p. 546–551, 1997.

FARAWAY, J. J. Multicollinearity and Logistic Regression. In Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models. [s.l.] CRC Press, 2006.

FERNANDES, A. A. T. et al. Read this paper if you want to learn logistic regression. Revista de Sociologia e Política, v. 28, n. 74, 2020.

GOLUB, G. H.; HEATH, M. A.; WAHBA, G. Overfitting in Machine Learning: A Review. IEEE Transactions on Neural Networks and Learning Systems, v. 28, n. 8, p. 1811–1820, 2017.

HARRELL, F. E. Ordinal logistic regression. Em: Springer Series in Statistics. Em: Springer Series in Statistics, p. 311–325, 2015.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. The elements of statistical learning: Data mining, inference, and prediction. 2. ed. Nova Iorque, NY, USA: Springer, 2009.

HERNÁN, M. A.; ROBINS, J. M. Causal Inference: What If. [s.l.] Chapman & Hall/CRC, 2010.

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. Generalized ordered logit/partial proportional odds models for ordinal dependent variables. The Stata Journal, v. 6, n. 1, p. 58–82, 2006.

IANNARIO, M.; PICCOLO, D. A comprehensive framework of regression models for ordinal data. Metron, v. 74, n. 2, p. 233–252, 2016.

JÖRESKOG, K. G. Ordinal Logistic Regression and Multicollinearity: A Case Study. International Journal of Quantitative Research in Education, v. 1, n. 2, p. 123–136, 2005.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Em: International Joint Conference on Artificial Intelligence. [s.l: s.n.]. p. 1137–1143.

LITTLE, R. J. A.; RUBIN, D. B. Statistical Analysis with Missing Data. [s.l.] Wiley, 2002.

MCCULLAGH, P. Regression models for ordinal data. Journal of the Royal Statistical Society. Series B, Statistical methodology, v. 42, n. 2, p. 109–127, 1980.

MTW; LONG, J. S. Regression models for categorical and limited dependent variables. Journal of the American Statistical Association, v. 92, n. 440, p. 1655, 1997.

OBRIEN, R. M. Detecting Multicollinearity: A Review of Statistical Approaches. Sociological Methods & Research, v. 36, n. 4, p. 479–507, 2007a.

OBRIEN, R. M. The Impact of Multicollinearity on Multiple Linear Regression Analysis. Quality & Quantity, v. 41, n. 5, p. 673–690, 2007b.

O’BRIEN, R. M. A caution regarding rules of thumb for variance inflation factors. Quality & quantity, v. 41, n. 5, p. 673–690, 2007.

SMEETON, N. A Review of: “Analysis of Ordinal Categorical Data, Second Edition, by A. Agresti,”: Hoboken, NJ: Wiley, 2010, ISBN 978-0-470-08289-8, xi + 396 pp., $110. Journal of biopharmaceutical statistics, v. 20, n. 5, p. 1080–1082, 2010.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. Information processing & management, v. 45, n. 4, p. 427–437, 2009.

WIGGINS, N. L.; HURST, H. S. Analysis of Multicollinearity and Its Effect on Logistic Regression Models. International Journal of Business and Social Science, v. 10, n. 6, p. 1–12, 2019.

WILLIAMS, R. Generalized ordered logit/partial proportional odds models for ordinal dependent variables. The stata journal, v. 6, n. 1, p. 58–82, 2006.

ZIEGEL, E. R.; MENARD, S. Applied Logistic Regression Analysis. Technometrics: a journal of statistics for the physical, chemical, and engineering sciences, v. 38, n. 2, p. 192, 1996.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), v, v. 67, n. 2, p. 301–320, 2005.

Disponível em: <https://www.gov.br/agricultura/pt-br/assuntos/noticias/setor-cervejeiro-segue-crescendo-a-cada-ano-aponta-anuario..>. Acesso em: 6 out. 2024a.

Disponível em: <https://repositorio.ufu.br/bitstream/123456789/39812/1/CiênciaDadosAprendizado.pdf>. Acesso em: 27 set. 2024b.

Disponível em: <https://www.sagepub.com/sites/default/files/upm-binaries/36660\_8.pdf>. Acesso em: 29 set. 2024c.

Disponível em: <https://repositorio.ufu.br/handle/123456789/35736>. Acesso em: 6 out. 2024d.

Disponível em: <https://labelyourdata.com/articles/movie-recommendation-with-machine-learning>.>. Acesso em: 6 out. 2024e.

**Apêndice ou Anexo** (opcional)