Ciência de Dados e Inteligência Artificial
Escola de Matemática Aplicada
Fundação Getúlio Vargas

Projeto Final

Modelagem Estatística

Iago Riveiro Santos Dutra

Rio de Janeiro, 2024

1 Introdução

O vinho é um produto de grande importância econômica e social, e sua história se confunde com a história da humanidade. Dos banquetes socráticos às festas contemporâneas, o vinho sempre esteve presente. Notoriamente, a tradição cristã conta na Bíblia que o primeiro milagre de Jesus teria sido transformar água em vinho para um casamento (BÍBLIA, 1999, João 2:1-11). Em particular, no seu diálogo *As Leis*, Platão discute extensamente sobre o vinho e a embriaguez, concluindo que os idosos, menos inclinados ao desvario, se beneficiam de bebê-lo, enquanto que o vinho deve ser vetado aos jovens, pois a bebida pode levar à perda do autocontrole e à desordem:

> "Não deveremos promulgar uma lei segundo a qual, em primeiro lugar, nenhuma criança abaixo de dezoito anos pode tocar de modo algum o vinho, ensinando que é errado verter fogo sobre fogo no corpo e na alma antes que comecem a se ocupar de seus efetivos labores, e assim nos guardando da disposição excitável dos jovens? E em seguida regulamentaremos para que o jovem de menos de trinta anos possa tomar vinho com moderação, abstendo-se inteiramente da intoxicação e da embriaguez. Mas quando um indivíduo atingir quarenta anos, poderá participar das reuniões festivas e invocar os deuses, particularmente Dionísio, convidando este deus para o que é simultaneamente cerimônia religiosa e recreação dos mais velhos, à qual ele concedeu à humanidade como um potente medicamento contra a rigidez da velhice, para que através deste possamos reavivar nossa juventude e que, olvidando seu zelo, a tempera de nossas almas possa perder sua dureza e se tornar mais macia e mais dúctil, tal como o ferro que foi forjado no fogo e passou a ser mais maleável" (PLATÃO, 2021).

A produção da bebida é uma arte, e a qualidade do vinho depende de muitos fatores, como o clima, o solo, a uva, o processo de produção, o armazenamento e o envelhecimento. Rios de tinta já foram dedicados a discutir a qualidade do vinho, derivando-se até mesmo uma ciência específica, a enologia.

Os especialistas possuem variados critérios para avaliar a qualidade do produto, e as diferentes características organolépticas são centrais para se decidir o valor de um lote. Nesse sentido, entender como as variáveis físico-químicas influenciam a qualidade do vinho é um passo importante para aprimorar a produção e a avaliação do produto.

O presente trabalho visa analisar um conjunto de dados sobre vinhos tintos portugueses, disponível no repositório da University of California, Irvine (CORTEZ *et al*., 2009), e propor um modelo de classificação para a qualidade do vinho, com base em suas características físico-químicas. A questão central é compreender quais fatores influenciam na avaliação de um lote de vinho e de que maneira.

O dataset é composto por 1599 observações e 12 variáveis físico-químicas, além da variável de qualidade, que é uma nota ordinal de 0 a 10. Dentre as covariáveis, estão o teor alcoólico, o pH, a acidez fixa, a acidez volátil, o ácido cítrico, o açúcar residual, os níveis de cloreto, o dióxido de enxofre livre e total, a densidade e os níveis de sulfatos. Um repositório com o código utilizado pode ser encontrado em: https://github.com/DutraIRS/Wine-Quality-Characteristics.

2 Metodologia

Para a modelagem do problema, utilizou-se um Modelo de Chances Proporcionais (MCCULLAGH, 1980). Esse modelo faz parte da família de modelos de regressão ordinal, que, por sua vez, é uma extensão da regressão multinomial e, por fim, da regressão logística. Essa classe de GLMs é utilizada para modelar variáveis dependentes ordinais, ou seja, variáveis que possuem um certo grau de ordenação. No caso do problema em questão, a variável dependente é a qualidade do vinho.

A interpretação do mecanismo da regressão ordinal depende da função de ligação escolhida. Para a função probit, o preditor linear é visto como uma variável latente contínua em cujo espaço deseja-se encontrar um ponto de corte que divide as categorias da variável dependente. Para a função logit, o preditor linear é visto como a log-odds de uma categoria da variável dependente ser igual ou maior que uma categoria de referência. A função de ligação log-log é uma generalização da função logit, que permite que a relação entre as categorias da variável dependente seja não-linear. Por possuir uma interpretação mais imediata, a função probit foi usada ao longo deste estudo.

Uma suposição importante que deve ser satisfeita é a proporcionalidade das chances. Isso significa que a razão entre as chances de um vinho ser de qualidade i ou maior e de qualidade i-1 ou menor é constante para todos os valores de i. A violação dessa suposição pode ser verificada por meio de testes de proporcionalidade das chances, como o teste de Brant. Mais concretamente, um incremento de uma unidade no preditor linear deve aumentar as chances de um vinho ser de qualidade i ou maior em um fator constante, independentemente do valor de i, pois definiu-se que os pontos de corte estão sobre o mesmo domínio da variável latente. Uma consequência importante desta definição é que os coeficientes angulares da regressão logística ordinal são iguais para todas as categorias da variável dependente, e cada coeficiente linear é responsável por deslocar a curva de probabilidade de uma categoria para a próxima. Além disso, os interceptos são diferentes para cada categoria, pois representam o ponto de corte entre as categorias da variável dependente.

A Figura 1 ilustra o modelo. Nela, C_i é o ponto de corte entre as categorias i e $i+1$. π_i é a probabilidade da categoria i. Se uma covariável X_i está associada a um coeficiente positivo, então um incremento em X_i desloca a curva para a direita e aumenta as probabilidades das categorias maiores.

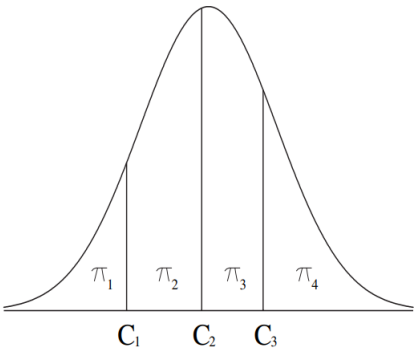

Figura 1: Distribuição da variável latente. Fonte: DOBSON, 2002.

2.1 Definição de Odds

As chances de um evento são definidas como a razão entre a probabilidade de o evento ocorrer e a probabilidade de o evento não ocorrer. Matematicamente, se $P(E)$ é a probabilidade de um evento E, então as chances de E são dadas por:

$$\text{Odds}(E) = \frac{P(E)}{1 - P(E)}$$

2.2 Modelo de Chances Proporcionais com Link Probit

No modelo de chances proporcionais com link probit, a relação entre as chances logarítmicas e as covariáveis X é definida pela função de distribuição acumulada (CDF) da distribuição normal padrão, conhecida como função probit:

$$\Phi^{-1}(P(Y \leq j)) = \alpha_j - \beta^T X, \tag{1}$$

onde:

- $\Phi^{-1}(\cdot)$ é a função inversa da CDF da distribuição normal padrão,
- $P(Y \leq j)$ é a probabilidade de Y ser menor ou igual a j,
- α_j é o intercepto específico para a categoria j de Y,
- β é um vetor de coeficientes associados às covariáveis X.

A função probit relaciona linearmente a variável latente $\eta = \alpha_j - \beta^T X$ com as covariáveis explicativas. A interpretação da variável latente η é fundamental no modelo probit: representa a medida de desvio das covariáveis explicativas da média para a categoria j da variável resposta Y.

2.3 Aplicação e Interpretação

O modelo de chances proporcionais com link probit é aplicado em situações onde se deseja modelar a relação entre variáveis explicativas e uma variável resposta ordinal, considerando a distribuição normal para a relação entre a variável latente η e as categorias ordinais de Y.

Interpretar os resultados do modelo probit envolve examinar como as covariáveis influenciam a variável latente η e, consequentemente, as probabilidades das categorias ordinais de Y. Coeficientes positivos indicam um aumento na variável latente η, o que pode levar a uma maior probabilidade de pertencer a uma categoria superior de Y, enquanto coeficientes negativos indicam o oposto.

2.4 Implementação

O modelo foi desenvolvido utilizando as bibliotecas *rcompanion*, *ordinal*, *MASS*, *readr*, *brant* e *AER* da linguagem de programação *R*. A análise exploratória foi feita em *Python* e utilizou os pacotes *Pandas*, *Numpy*, *Seaborn* e *Matplotlib*. A construção de modelos se deu de forma aditiva: inicialmente, se ajustou um modelo com todas as covariáveis e todas as interações; da observação dos efeitos desse modelo se ajustou uma série de modelos, a cada passo incluindo a próxima covariável/interação com maior capacidade explicativa até que o AIC começasse a cair. O ajuste dos modelos utilizou Fisher Scoring. A avaliação dos resultados se baseou em AIC e na acurácia de validação cruzada. Pormenores metodológicos podem ser encontrados no arquivo *analysis.R*, presente no repositório do projeto.

3 Resultados e Discussão

Dentre os padrões descobertos durante a análise exploratória, notou-se que, fora as variáveis pH, densidade e qualidade, todas as covariáveis apresentaram distribuição assimétrica, com

muitos casos com baixo valor e cauda pesada à direita. Como indicado na Figura 2, embora certos pares de covariáveis possuam comportamento linearmente correlacionado, esses efeitos não parecem ser de grande relevância.

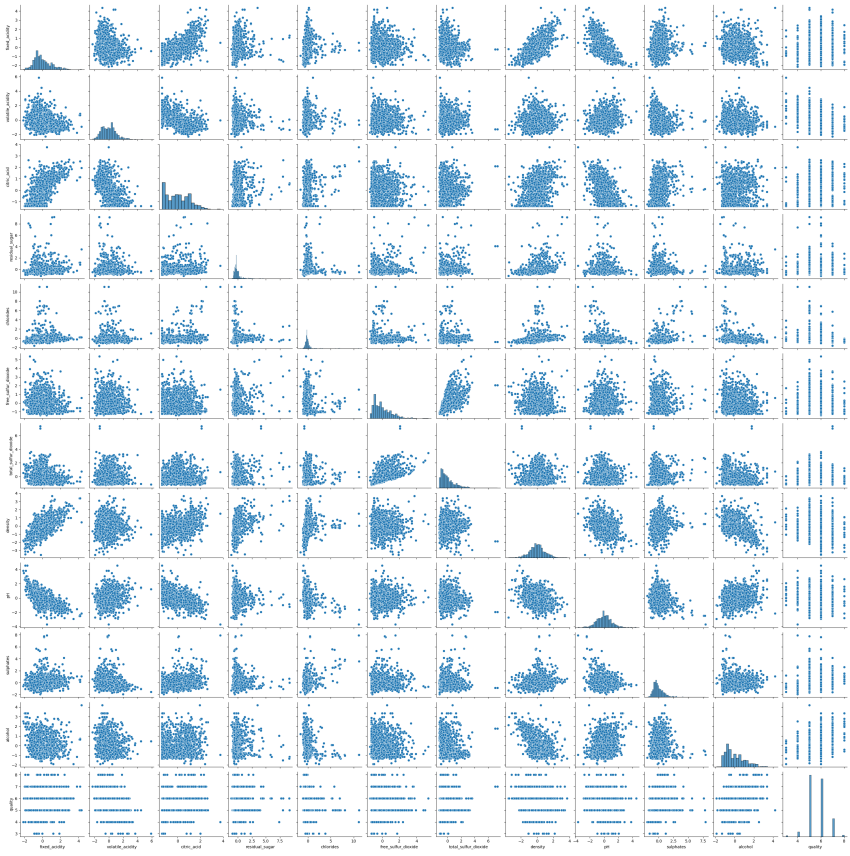

Figura 2: Distribuição conjunta das variáveis tomadas par a par.

A Figura 3 traz a matriz de covariâncias das variáveis. Todas, exceto qualidade, foram normalizadas. Nota-se que não há correlação linear superior a 0.7 para nenhum par de variáveis, o que é importante para contornar o problema de multicolinearidade.

A Tabela 1 apresenta os resultados dos diferentes modelos ajustados. O melhor modelo encontrado é apresentado na Figura 4. A Figura 5 traz a matriz de confusão do modelo. Nota-se que a maior parte dos erros do modelo são categorias vizinhas.

Tabela 1: Modelos ajustados

Modelo	**AIC**	**Acurácia (%)**
0	3220	57,03
1	3402	55,47
2	3209	58,16
3	3166	57,35
4	3151	58,04
5	3130	58,91
6	3131	58,85
7	3123	58,41

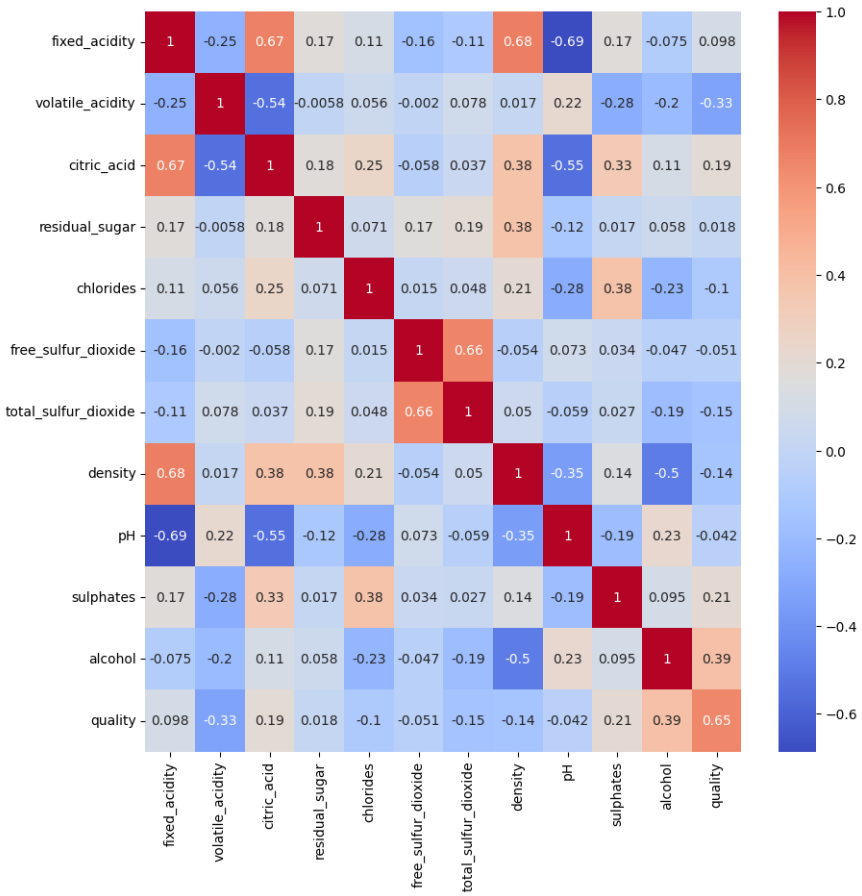

Figura 3: Matriz de covariância.

```
z test of coefficients:

                                Estimate Std. Error  z value  Pr(>|z|)
alcohol                        0.5294491  0.0340851  15.5332 < 2.2e-16 ***
volatile_acidity              -0.3423080  0.0373721  -9.1595 < 2.2e-16 ***
sulphates                      0.2673250  0.0329566   8.1114 5.003e-16 ***
chlorides                     -0.1559158  0.0336972  -4.6270 3.711e-06 ***
total_sulfur_dioxide          -0.1567804  0.0313111  -5.0072 5.523e-07 ***
residual_sugar                 0.0083629  0.0325215   0.2571 0.7970638
pH                            -0.1350297  0.0358180  -3.7699 0.0001633 ***
citric_acid                   -0.0583905  0.0417987  -1.3969 0.1624298
total_sulfur_dioxide:residual_sugar  0.0243123  0.0188161   1.2921 0.1963210
pH:citric_acid                 0.0023288  0.0277693   0.0839 0.9331662
3|4                           -3.0149496  0.1297392 -23.2385 < 2.2e-16 ***
4|5                           -2.1789557  0.0679291 -32.0769 < 2.2e-16 ***
5|6                           -0.1508340  0.0382381  -3.9446 7.993e-05 ***
6|7                            1.4685858  0.0523234  28.0675 < 2.2e-16 ***
7|8                            3.0255781  0.1140963  26.5178 < 2.2e-16 ***
---
```

Figura 4: Coeficientes do melhor modelo.

Como pode-se observar, o teor alcólico e os níveis de sulfatos contribuem positivamente para a avaliação do vinho, enquanto o teor de acidez volátil, de cloretos, de dióxido de enxofre total e o pH contribuem negativamente para a avaliação. As demais variáveis não possuem

influência ou sua influência é indistinguível de nula.

Como diagnóstico de satisfação dos pressupostos do método, aplicou-se o teste de Brant sobre os coeficientes do modelo para avaliar se a propocionalidade é atendida. A premissa de paralelismo não pôde ser rejeitada, de forma que a regressão utilizada é válida.

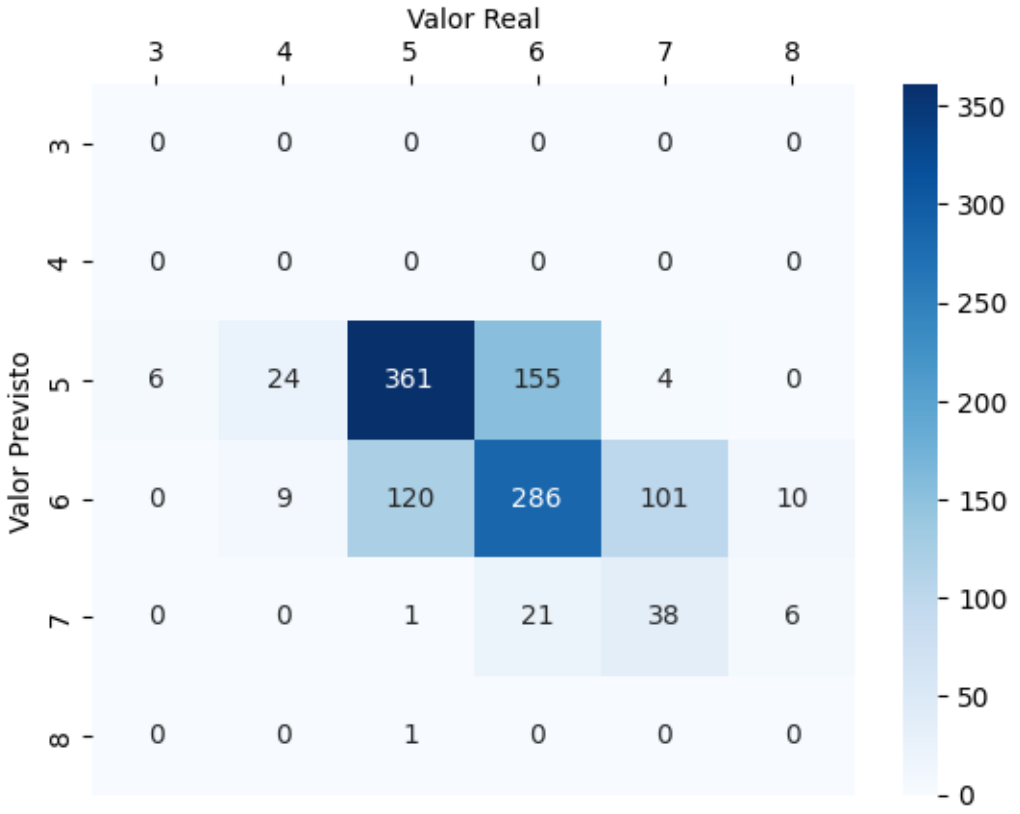

Figura 5: Matriz de confusão.

4 Conclusão

Foi possível chegar a um modelo que prevê corretamente a qualidade do vinho na maioria dos casos, identificando, com isso, os fatores que determinam a avaliação do vinho. Haja vista que se tratam de seis categorias, houve um aprendizado de padrões significativo, com taxa de acerto substancialmente maior que a aleatória. Ao avaliar os erros cometidos pelo modelo, é possível perceber que a maioria dos erros ocorre entre categorias vizinhas, o que é esperado, uma vez que a diferença entre as categorias é sutil e avaliadores humanos podem ser inconsistentes.

Dentre as fragilidades do método, destaca-se a necessidade de um número maior de observações para as categorias menos frequentes, uma vez que o modelo tende a classificar erroneamente essas observações como pertencentes a categorias mais frequentes. Além disso, o modelo assume que as chances são proporcionais entre as categorias. Embora o teste de Brant tenha comprovado que não é possível rejeitar a hipótese de proporcionalidade, não é inconsequente pensar que certas sutilezas e podem ser perdidas por tal simplificação. Em particular, não é difícil imaginar que o incremento no teor alcólico possa ter um impacto mais significativo na qualidade de vinhos de qualidade inferior do que na qualidade de vinhos de qualidade superior.

Trabalhos futuros poderiam incorporar informações para além das químicas, como a região de origem do vinho, o produtor, o tipo de uva etc. Seria interessante também avaliar outras funções de ligação, como a logística. Por fim, um tratamento bayesiano poderia ser interessante para lidar com a questão da proporcionalidade entre as categorias, definindo uma distribuição a priori para os coeficientes das covariáveis que dependa da categoria.

Referências Bibliográficas

BÍBLIA. João, 2:1-11. In: **BÍBLIA SAGRADA**. Tradução de João Ferreira de Almeida. 2. ed. Barueri: Sociedade Bíblica do Brasil, 1999.

Cortez, Paulo, Cerdeira, A., Almeida, F., Matos, T., Reis, J.. (2009). Wine Quality. **UCI Machine Learning Repository**. DOI: 10.24432/C56S3T.

Dobson, A. J. **An Introduction to Generalized Linear Models**. Chapman Hall/CRC, 2002

McCullagh, P. Regression Models for Ordinal Data. **Journal of the Royal Statistical Society**. Series B (Methodological), vol. 42, no. 2, 1980, pp. 109-142.

Platão. **As Leis (ou Da Legislação): incluindo Epinomis**. Tradução de Edson Bini. 3. ed. São Paulo: Edipro, 2021.