

Aplicação de modelos de aprendizado de máquina para classificação de vinhos

Arthur Vasconcelos de Oliveira Román Porciúncula

Departamento de Estatística
Centro de Ciências Exatas e da Natureza - CCEN
Universidade Federal de Pernambuco - UFPE

Recife, 21 de março de 2025



Roteiro

Introdução

Objetivos

Método e dados

Análise Exploratória

Modelos de ML

Conclusão

Referências

Sobre o vinho

- ▶ O vinho é uma iguaria apreciada há muito tempo pelo homem, com registros históricos apontando que essa relação já existe há mais de 2000 anos.
- ▶ Alguns estudos comprovam que o consumo em quantidades moderadas promove benefícios à saúde, como na prevenção de doenças cardiovasculares e alguns tipos de tumores.



Sobre o vinho

- ▶ Estima-se que em 2023, esse produto movimentou 36 bilhões de euros globalmente, com a produção girando em torno de 237 milhões de hectolitros, ou 23.7 bilhões de litros.
- ▶ Há pesquisas e concursos mundo afora em busca de tornar o vinho ainda mais palatável, analisando aromas e sabores que o compõem. Estes componentes dependem das uvas, da fermentação alcoólica na produção, do amadurecimento e envelhecimento do vinho, entre outros fatores.



Objetivos

Esse estudo tem como objetivos principais:

- ▶ Classificar, através de modelos de aprendizado de máquina, os vinhos segundo a qualidade, baseado em fatores físicos e químicos da bebida;
- ▶ Verificar quais as principais características que tornam os vinhos mais agradáveis, comparando com outro estudo similar.

Método

- ▶ São utilizados modelos de árvore de decisão e de floresta aleatória, pois são os que apresentaram melhores resultados.
- ▶ Enquanto a árvore de decisão é mais intuitiva e bem interpretável, a floresta aleatória apresenta menos risco de *overfitting* e de variância alta nos resultados.
- ▶ Inicialmente, os dados são divididos em conjuntos de treinamento (70%) e teste (30%).
- ▶ Para avaliar os modelos, o conjunto de teste é utilizado para construir algumas métricas apresentadas durante o curso.

Dados

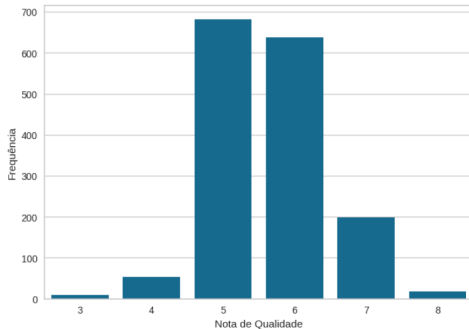
O conjunto de dados consiste em 1599 observações de vinhos tintos, em que as seguintes características são avaliadas:

- ▶ ácido tartárico (fixed acidity), em g/dm^3 ;
- ▶ ácido acético (volatile acidity), em g/dm^3 ;
- ▶ ácido cítrico, em g/dm^3 ;
- ▶ açúcar residual, em g/dm^3 ;
- ▶ cloreto de sódio, em g/dm^3 ;
- ▶ sulfato de potássio, em g/dm^3 ;
- ▶ dióxido de enxofre livre, em mg/dm^3 ;
- ▶ dióxido de enxofre total, em mg/dm^3 ;
- ▶ densidade do vinho, em g/dm^3 ;
- ▶ pH;
- ▶ a porcentagem de álcool no vinho.

Dados

- ▶ A variável principal (rótulo) é a nota de 0 (muito ruim) a 10 (excelente) dada ao vinho por meio de um teste sensorial às cegas. Para obtê-la, três pessoas avaliam cada vinho, e o resultado é a mediana das avaliações.

Figura: Distribuição de frequência da nota dos vinhos

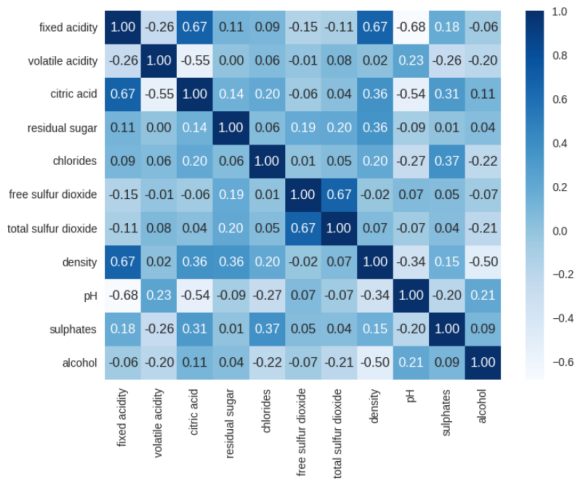


- Para simplificar o método de classificação e buscar separar os melhores vinhos dos demais, as observações são reduzidas a duas classes: “Qualidade ótima”, com nota igual a 7 ou 8, ou “Sem qualidade ótima”, com nota igual a 6 ou inferior.

Tabela: Principais medidas descritivas das características da amostra

Característica	Mínimo	Média	Mediana	Máximo
Ác. tartárico	4.60	8.32	7.90	15.90
Ác. acético	0.12	0.53	0.52	1.58
Ác. cítrico	0.00	0.27	0.26	1.00
Açúcar residual	0.90	2.54	2.20	15.50
Cloreto de sódio	0.01	0.09	0.08	0.61
Sulf. de potássio	0.33	0.66	0.62	2.00
Di. enx. livre	1.00	15.88	14.00	72.00
Di. enx. total	6.00	46.47	38.00	289.00
Densidade	0.9900	0.9967	0.9968	1.0037
pH	2.74	3.31	3.31	4.01
Álcool	8.40	10.42	10.20	14.90

Figura: Correlação entre as características da amostra



Árvore de decisão

- ▶ Tem como critério de impureza o Índice de Gini, profundidade igual a 3 e cada nó se bifurca.
- ▶ Para uma observação ser classificada como de “Qualidade ótima”, há dois caminhos:
 1. o volume de álcool é maior que 11.55; a concentração de sulfato de potássio é menor ou igual a 0.685; o dióxido de enxofre total é menor ou igual a 15.5.
 2. O volume de álcool é maior que 11.55; a concentração de sulfato de potássio excede 0.685, o dióxido de enxofre total deve ser menor ou igual a 18.5.
- ▶ Para qualquer outro caso, a observação é classificada como de “Qualidade não ótima”.

Árvore de decisão

Tabela: Matriz de confusão

Valores reais	Valores preditos	
	Qual. não ótima	Qual. ótima
Qual. não ótima	404	16
Qual. ótima	35	25

- ▶ Acurácia = 0.8983;
- ▶ MCC = 0.4479;
- ▶ Precisão = 0.6098.

Floresta Aleatória

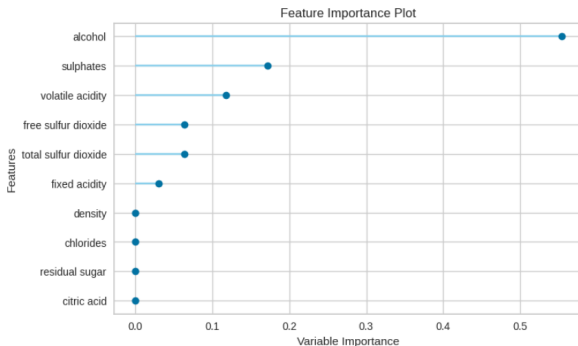
- ▶ A floresta aleatória utiliza os mesmos critério de Gini e consiste em 100 árvores diferentes.

Tabela: Matriz de confusão

Valores reais	Valores preditos	
	Qual. não ótima	Qual. ótima
Qual. não ótima	409	11
Qual. ótima	35	25

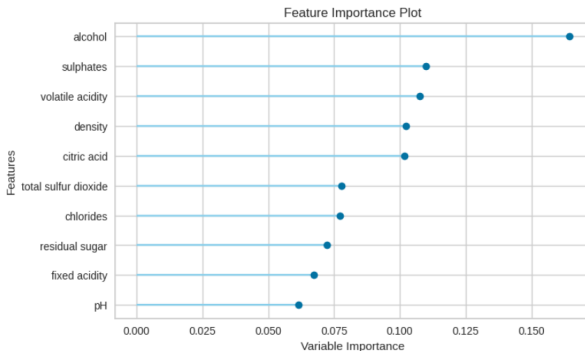
- ▶ Acurácia = 0.9042;
- ▶ MCC = 0.4903;
- ▶ Precisão = 0.6944.

Figura: Gráfico de importância das características da árvore de decisão



- O volume de álcool apresenta importância de mais de 50%, muito à frente das demais e seguida pela concentração de sulfato e de ácido acético, ambas com importância na faixa de 10% a 20%.

Figura: Gráfico de importância das características da floresta aleatória







- Aqui, a importância está muito mais bem distribuída entre as variáveis, inclusive de forma similar ao estudo feito com vinhos brancos, usando SVM. Em ambos, volume de álcool e concentração de sulfatos são as principais características.

Principais conclusões

- ▶ Os modelos performam de forma semelhante, com a floresta aleatória apresentando métricas ligeiramente superiores;
- ▶ Ambos os modelos apresentaram a mesma tendência de erro: classificar muitos vinhos de qualidade ótima como sendo de qualidade não ótima;
- ▶ Mesmo com acurácia alta, o fato de haver desbalanceamento nos dados, com mais de 80% sendo vinhos de qualidade não ótima, pode ter causado esse problema;
- ▶ Tanto a árvore de decisão quanto a floresta aleatória deram ao volume de álcool na bebida a maior importância entre todas as variáveis;
- ▶ A importância bem distribuída das variáveis da floresta aleatória é condizente com o estudo de outros tipos de vinhos realizado anteriormente.

Referências

-  P. Cortez, J. Teixeira, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, “Using data mining for wine quality assessment,” in *International Conference on Discovery Science*. Springer, 2009, pp. 66–79.
-  I. O. of Vine and M. Wine, “State of the world vine and wine sector in 2022,” 2023.
-  V. d. MORAES and C. Locatelli, “Vinho: uma revisão sobre a composição química e benefícios à saúde,” *Evidência*, vol. 10, no. 1-2, pp. 57–68, 2010.
-  N. G. Penna and L. H. R. Hecktheuer, “Vinho e saúde: uma revisão,” *Infarma-Ciências Farmacêuticas*, vol. 16, no. 1/2, pp. 64–67, 2004.