Universidade Federal de Pernambuco

Centro de Ciências Exatas e da Natureza

Departamento de Estatística

APLICAÇÃO DE MODELOS DE APRENDIZADO DE MÁQUINA PARA CLASSIFICAÇÃO DE VINHOS SEGUNDO A QUALIDADE E CARACTERÍSTICAS FÍSICAS E QUÍMICAS

Introdução

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 [1]. Além de estudos comprovando que o consumo em quantidades moderadas promove benefícios à saúde, principalmente na prevenção de doenças cardiovasculares e alguns tipos de tumores [2], essa bebida também tem um grande impacto na indústria alimentícia. Estima-se que em 2023, por exemplo, 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 [3].

Por conta dessa importância, surgem 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 [4], e existem para todos os gostos, tamanha a diversidade dessa indústria.

Visando esse tema, o presente estudo busca utilizar métodos de aprendizado de máquina para classificar alguns vinhos segundo a qualidade, baseado em fatores físicos e químicos da bebida. Assim, é possível comparar os resultados com um estudo semelhante, em que se realizou um modelo de SVM para classificar vinhos brancos [5], observar se tais métodos produzem bons resultados e verificar quais são as principais características que tornam os vinhos mais agradáveis.

Fundamentos Teóricos e Metodológicos

Após avaliação de cinco diferentes modelos para os dados, incluindo *SVM* radial e não radial e *kNN*, os que obtiveram melhores resultados foram a árvore de decisão e a floresta aleatória, que são brevemente descritas a seguir.

A árvore de decisão é um modelo de aprendizado supervisionado utilizado tanto para problemas de classificação quanto de regressão, e tem como principal vantagem a capacidade de representar "caminhos" de maneira intuitiva e interpretável. A estrutura da árvore consiste em nós, onde cada nó interno representa uma decisão com base em uma característica específica do conjunto de dados, e os nós finais (ou folhas) representam a previsão ou a classe final [6].

O processo de construção de uma árvore de decisão envolve a divisão recursiva do conjunto de dados em subconjuntos mais homogêneos, com base em um critério de impureza (nesse caso, o Índice de Gini). A árvore é construída até que um critério de parada seja atendido, como a profundidade máxima da árvore ou a pureza dos nós alcançada.

Já a floresta aleatória, também sendo um modelo de aprendizado supervisionado usado para classificação e regressão, utiliza diferentes árvores de decisão treinadas por diferentes subconjuntos de observações e caractéristicas [7]. Aqui, para cada observação, a moda das classes preditas por todas as árvores de decisão é a predita pela floresta aleatória. Desse modo, o risco

de *overfitting* e uma possível alta variância nos resultados diminui, porém a interpretabilidade não é tão fácil quanto o método anterior.

Algumas métricas importantes para avaliar o desempenho desses modelos são a acurácia (total de acertos dividido pelo total de observações), a precisão, o *recall*, o coeficiente *F1* e o coeficiente *MCC*, que é o mais completo, avaliando todos os cenários de acertos e erros de predição.

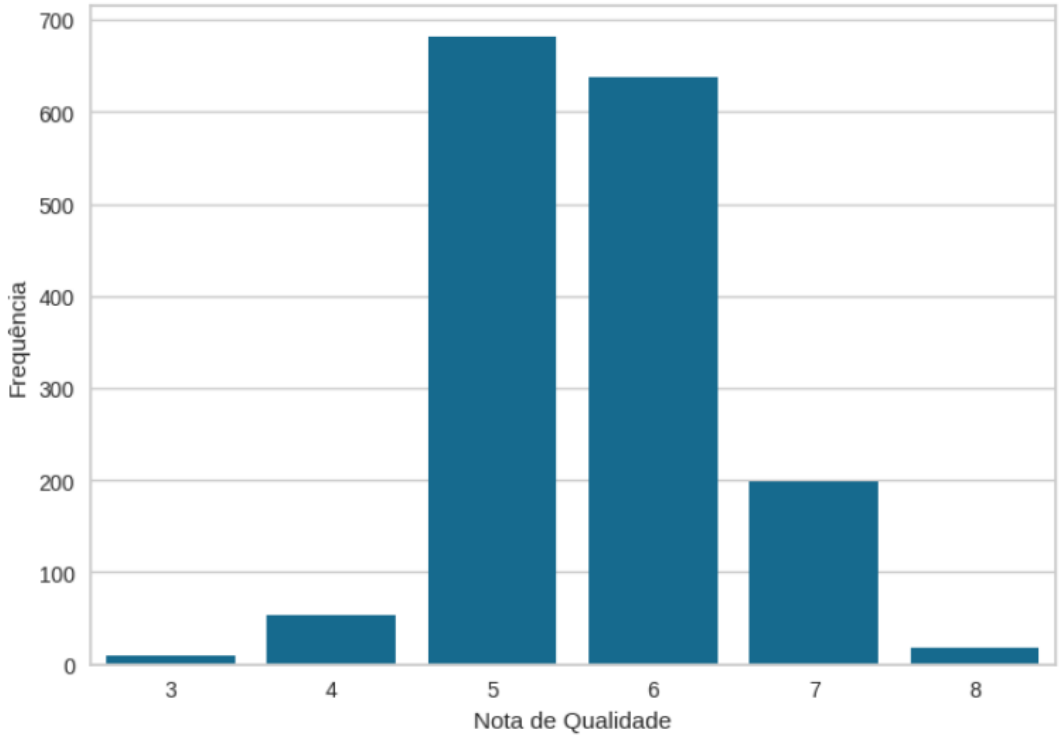
Aplicação

Análise Exploratória

O banco de dados utilizado consiste em 1599 observações de vinhos tintos, em que são avaliadas as concentrações de ácido tartárico (fixed acidity), ácido acético (volatile acidity), ácido cítrico, açúcar residual, cloreto de sódio e sulfato de potássio, em g/dm^3; dióxido de enxofre livre e dióxido de enxofre total, em mg/dm^3, a densidade do vinho, em g/dm^3, o pH e a porcentagem de álcool no vinho. A variável principal é a nota de 0 (muito ruim) a 10 (excelente) dada ao vinho por meio de um teste sensorial às cegas, realizado por três pessoas para cada vinho, e a nota em questão é a mediana dessas avaliações.

As notas apresentadas vão de 3 a 8, com mais de 80% delas sendo 5 ou 6, como pode se observar na Figura 1. Por conta disso, as observações são separadas entre "Qualidade ótima", com nota igual a 7 ou 8, ou "Sem qualidade ótima", com nota igual a 6 ou inferior. O objetivo com essa transformação é simplificar o método de classificação, reduzindo a apenas duas classes e buscando segmentar os melhores vinhos dos demais.

Figura 1: Distribuição de frequência da Nota de Qualidade dos vinhos



Com exceção da concentração de dióxido de enxofre total, as demais características aparentam distribuição comportada, com média e mediana muito próximas, conforme o Quadro 1.

Quadro 1: 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

Além disso, calculadas correlações de Pearson entre as características, duas a duas, tem-se que apenas três são relevantes, com 0.67 de correlação: densidade e ác. tartárico, ác. tartárico e ác. cítrico, e dióxidos de enxofre livre e total. Demais correlações são fracas e não passam de 0.4.

Modelos de Aprendizado de Máquina

Primeiramente, as 1599 instâncias são divididas aleatoriamente, com 1119 (70%) sendo alocadas para o conjunto de treinamento dos dados e 480 (30%) para o conjunto de teste. Dessa forma, é possível calcular métricas que estimem se os modelos criados conseguem generalizar bem para novos dados, isto é, apresentarem bons resultados para dados além dos que servem como treino.

O modelo de árvore de decisão criado 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, e em ambos, o volume de álcool deve ser maior que 11.55. Em seguida, se a concentração de sulfato de potássio for menor ou igual a 0.685, o dióxido de enxofre total deve ser menor ou igual a 15.5. Já no caso em que a concentração de sulfato de potássio excede 0.685, a concentração de dióxido de enxofre total deve ser menor ou igual a 18.5. Caso a observação esteja disposta de qualquer outra maneira, é classificada como de "Qualidade não ótima".

Com base no conjunto de treino, as principais métricas calculadas são: *MCC* de 0.4479, acurácia de 0.8983, precisão de 0.6098, *recall* de 0.4167 e *F1* de 0.4950. Analisando esses dados e a matriz de confusão disposta no Quadro 2, tem-se que a grande maioria dos vinhos de qualidade não ótima são certamente preditos como tal, mas o mesmo não pode ser afirmado para os de qualidade ótima, pois muitos estão erroneamente classificados como não ótimos.

Quadro 2: Matriz de confusão da árvore de decisão para o conjunto de teste

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

Já a floresta aleatória utiliza os mesmos critério de Gini e consiste em 100 árvores diferentes, obtendo os seguintes resultados: *MCC* igual a 0.4903, acurácia igual a 0.9042, precisão igual a 0.6944, *recall* igual a 0.4167 e *F1* igual a 0.5208. Enquanto as métricas são melhores que as da árvore, principalmente a precisão, a matriz de confusão é muito semelhante à anterior, apresentando a mesma dificuldade, conforme observada no Quadro 3.

Quadro 3: Matriz de confusão da floresta aleatória para o conjunto de teste

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

Quanto à importância das características para os modelos, os resultados diferem significativamente. Para a 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%, como constatado na Figura 2. Já para a floresta aleatória, a importância está muito mais bem distribuída entre as variáveis, inclusive de forma similar ao estudo feito com vinhos brancos, com volume de álcool e concentração de sulfatos sendo as principais características de ambos. Isto é visível pelas Figuras 3 e 4.

Figura 2: Gráfico de importância das características da árvore de decisão

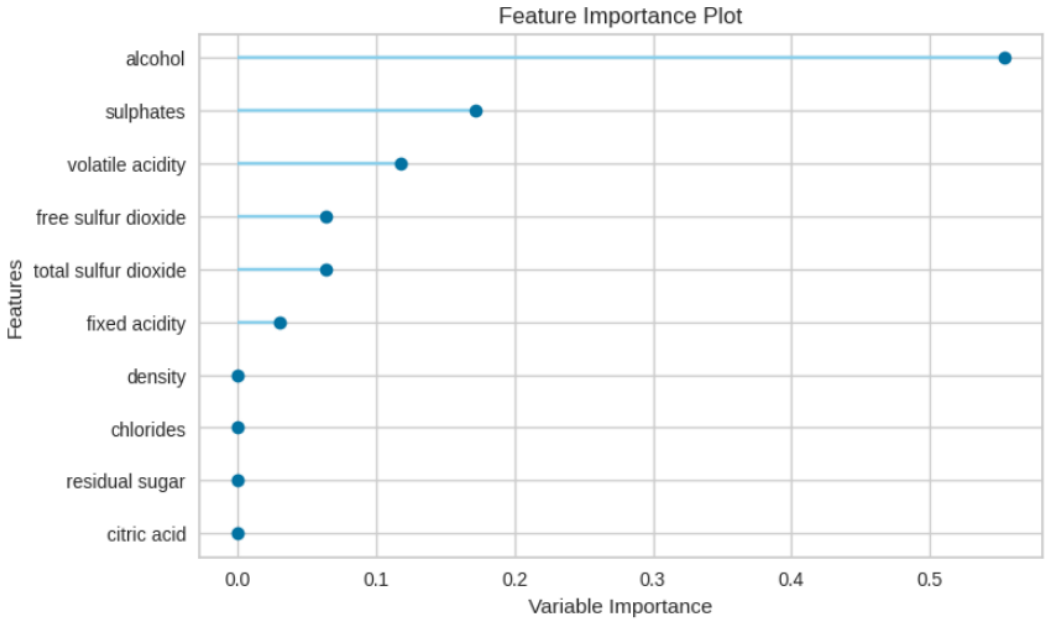


Figura 3: Gráfico de importância das características da floresta aleatória

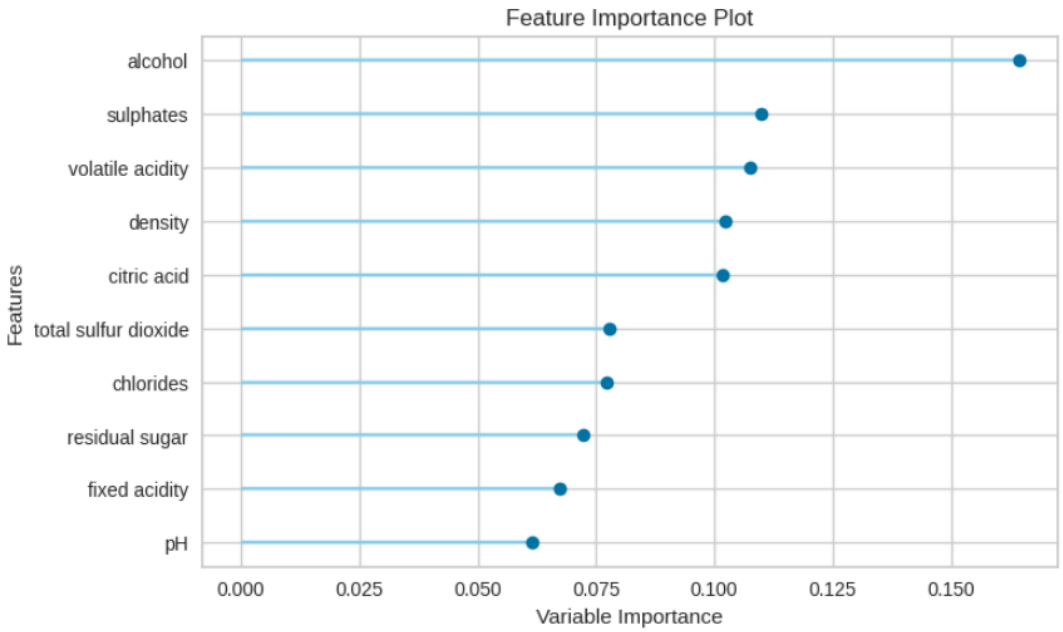
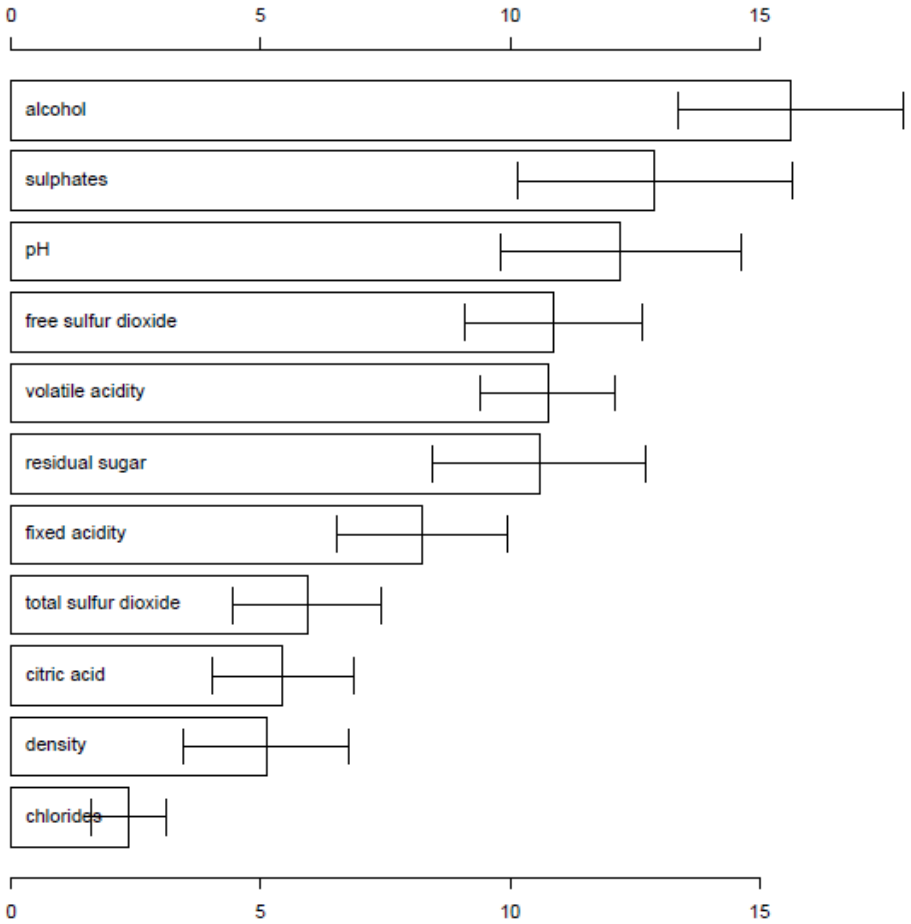


Figura 4: Gráfico de importância das características dos vinhos brancos



Fonte: The relative input importances for the SVM model (in %; bars denote the average value while the whiskers show the 95% con

dence intervals). (Cortez et al, 2009,p.12).

Conclusão

Ambos os modelos apresentam performances semelhantes, embora a floresta aleatória apresente métricas ligeiramente superiores a custo de uma menor interpretabilidade. Inclusive, os dois 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.

Por outro lado, a importância bem distribuída das variáveis da floresta aleatória é relativamente condizente com o estudo de outros tipos de vinhos realizado anteriormente. Além disso, tanto a árvore de decisão quanto a floresta aleatória deram ao volume de álcool na bebida a maior importância entre todas as variáveis.

Referências

[1] PENNA, Neidi Garcia; HECKTHEUER, Luísa Helena Rychecki. Vinho e saúde: uma revisão. Infarma-Ciências Farmacêuticas, v. 16, n. 1/2, p. 64-67, 2004. Disponível em: https://cff.emnuvens.com.br/infarma/article/view/332/321. Acesso em: 21/03/2025.

[2] MORAES, V. de; LOCATELLI, Claudriana. Vinho: uma revisão sobre a composição química e benefícios à saúde. Evidência, v. 10, n. 1-2, p. 57-68, 2010. Disponível em: https://periodicos.unoesc.edu.br/evidencia/article/view/1159/pdf_255. Acesso em: 21/03/2025.

[3] INTERNATIONAL ORGANISATION OF VINE AND WINE, M. State of the world vine and wine sector in 2023. 2024. Disponível em: https://www.oiv.int/sites/default/files/2024-04/OIV_STATE_OF_THE_WORLD_VINE_AND_WINE_SECTOR_IN_2023.pdf e http://oiv.int/sites/default/files/2024-04/2024_OIV_April_PressConference_PPT.pdf. Acesso em: 21/03/2025.

[4] ASSOCIAÇÃO BRASILEIRA DE SOMMELIERS - RS. Os aromas e sabores do vinho. 2021. Disponível em: https://www.absrs.com.br/post/os-aromas-e-sabores-do-vinho#:~:text=Os%20aromas%20e%20sabores%20prim%C3%A1rios,de%20aromas%20de%20frutas%20frescas. Acesso em: 21/03/2025.

[5] CORTEZ, Paulo et al. Using data mining for wine quality assessment. In: International Conference on Discovery Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. p. 66-79. Disponível em: https://www.researchgate.net/publication/221612614_Using_Data_Mining_for_Wine_Quality_Assessment. Acesso em: 21/03/2025.

[6] QUINLAN, J.. Ross . Learning decision tree classifiers. ACM Computing Surveys (CSUR), v. 28, n. 1, p. 71-72, 1996. Disponível em: https://meridian.allenpress.com/jim/article/47/1/31/131479/Random-Forest. Acesso em: 21/03/25

[7] RIGATTI, Steven J. Random forest. Journal of Insurance Medicine, v. 47, n. 1, p. 31-39, 2017. Disponível em: https://dl.acm.org/doi/pdf/10.1145/234313.234346. Acesso em: 21/03/25.