Wine Quality features

Lista **completa detalhada** das variáveis do famoso dataset “Wine Quality” (UCI Machine Learning Repository), que contém vinhos brancos e tintos portugueses (região dos Vinhos Verdes), muito utilizado em cursos de ciência de dados e machine learning para regressão e classificação da qualidade do vinho.

Acknowledgments to: P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Variáveis (features) do dataset:

Variável	Descrição detalhada	Unidade típica	Valores típicos (tintos)	Valores típicos (brancos)	Observações importantes
fixed acidity	Acidez fixa ou não volátil – ácidos orgânicos (principalmente tartárico, málico e succínico) que não evaporam facilmente.	g/dm³ (ácido tartárico)	6–16	5.5–14	Principal contribuinte da acidez total
volatile acidity	Acidez volátil – principalmente ácido acético e, em menor grau, ácido propiônico. Níveis altos causam sabor de vinagre.	g/dm³ (ácido acético)	0.1–1.1	0.1–0.9	Defeito muito penalizado na pontuação
citric acid	Ácido cítrico – presente em pequenas quantidades; dá frescor e “vivo” ao vinho.	g/dm³	0.0–1.7	0.0–1.7	Usado também para corrigir acidez
residual sugar	Açúcar residual – quantidade de açúcar que sobra após a fermentação. Vinhos < 1 g/L são secos; > 45 g/L são doces.	g/dm³	1–16	0.9–70 (muitos meio-doces)	Muito mais variável nos brancos

chlorides	Quantidade de cloretos (sais) no vinho – principalmente cloreto de sódio e potássio.	g/dm³ (NaCl)	0.01–0.6	0.01–0.4	Altos níveis dão sabor salgado
free sulfur dioxide	Forma livre de SO_2 (molecular + bisulfito) – parte ativa na proteção contra oxidação e microrganismos.	mg/dm³	1–70	3–290	Brancos geralmente têm mais
total sulfur dioxide	Soma da forma livre + forma ligada (bound) de SO_2.	mg/dm³	6–289	9–440	Limite legal na UE: ~150–200 mg/L dependendo do tipo
density	Densidade do vinho – próxima da água (1.000 g/cm³), varia com teor alcoólico (↓) e açúcar (↑).	g/cm³	0.990–1.003	0.987–1.003	Forte correlação negativa com álcool
pH	Medida da acidez/básicidade – vinhos geralmente entre 2.8 e 4.0. Valores mais baixos = mais ácidos.	escala 0–14	2.8–4.0	2.7–3.8	Tintos costumam ter pH ligeiramente maior
sulphates	Sulfatos (principalmente sulfato de potássio) – aditivo que aumenta os níveis de SO_2 livre; atua como antioxidante e antimicrobiano.	g/dm³ (sulfato de potássio)	0.3–2.0	0.2–1.2	Níveis muito altos podem ser perceptíveis
alcohol	Teor alcoólico (percentagem em volume).	% vol	8–15%	8–14%	Forte correlação positiva com a qualidade no dataset
quality (target)	Nota de qualidade atribuída por especialistas (escala ordinal de 0 a 10; no dataset real só aparecem notas de 3 a 9).	pontos (3–9)	3–8	3–9	Variável que normalmente se tenta prever

---

Notas adicionais importantes sobre o dataset:

- Existem duas versões separadas:
    - winequality-red.csv (1.599 amostras) –
    - winequality-white.csv (4.898 amostras)

  Muitas vezes são combinadas adicionando uma coluna “type” (red/white).
- A variável **quality** é o target típico para tarefas de classificação (classificar vinho como “bom” ou “ruim”) ou regressão (prever a nota exata).
- 12 variáveis físico-químicas + qualidade (inteiro de 3 a 9)
- Apesar de o dataset ser muito popular para ensino, tem algumas críticas acadêmicas:
    - As notas de qualidade são baseadas em apenas 3 provadores (mediana).
    - Distribuição da qualidade é assimétrica (muitos vinhos nota 5 e 6).
    - Algumas variáveis têm distribuições diferentes entre tintos e brancos.

Com estas 11 features + a variável alvo (quality), o dataset é perfeito para praticar regressão linear, árvores de decisão, random forest, SVM, redes neurais, etc.

O dataset único combinado a partir dos dois datasets iniciais:

- 1 599 vinhos tintos → coluna **type = red**
- 4 898 vinhos brancos → coluna **type = white**
- Total: **6 497 linhas**
- Coluna adicional **type** no início (para facilitar a diferenciação)
- Todas as 12 variáveis químicas + qualidade mantidas exatamente como nos originais
- Separador: ponto e vírgula (;)
- Primeira linha com cabeçalhos