AP1 – Projeto de Machine Learning

Dataset: Red Wine Quality | Dupla: Rafael Lima e André Silveira

1. **Escolha e apresentação do dataset**

- **Dataset escolhido:** Red Wine Quality;
- **Acesso para o dataset (Kaggle):** https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009;
- **Número de registros e variáveis:** 1599 observações e 12 variáveis, além da criação de uma variável extra (quality_bin);
- **Justificativa da escolha:** O dataset apresenta características de vinhos tintos e suas respectivas notas de qualidade, tornando-se ideal para:
    - **Regressão linear simples:** Predição da qualidade do vinho, que pode ser uma nota entre 3 e 8;
    - **Regressão logística:** Classificação da qualidade em categorias binárias (ex: qualidade inferior vs. Superior), através da aplicação de feature engineering na variável quality.
- **Objetivo da análise:** Entender a influência do teor alcoólico do vinho em sua qualidade.

2. **Pré-processamento e análise exploratória**

- **Tratamento de dados:**
    - O dataset **não apresenta valores ausentes**;
    - O dataset apresenta 240 valores duplicados – e que foram mantidos porque o modelo precisa aprender inclusive com repetições, e que elas representam variações reais na produção ou amostragem;
    - Criamos a variável binária quality_bin para classificação:
        - Qualidade superior (nota ≥ 6);
        - Qualidade inferior (nota < 6);

Observação: A separação do tipo de qualidade no valor 6 se deve ao fato de criar uma classe balanceada, sendo o ponto de equilíbrio no conjunto de dados e uma separação lógica entre a qualidade inferior e superior.

- **Análise descritiva e visual:**
  - Variáveis como quality, alcohol, density, pH, fixed.acidity foram analisadas;

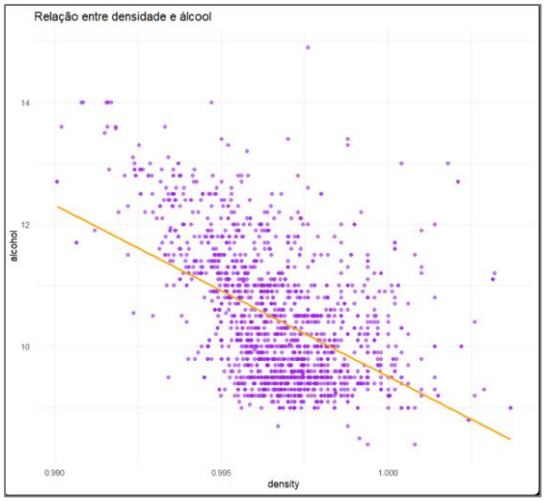

Figura 1 - Conforme o nível de álcool sobe, o nível de densidade tende a diminuir

  - Geramos histogramas, boxplots e gráficos de dispersão com linhas de regressão, sendo que a relação que mais nos chamou atenção foi a do álcool com qualidade, devido a sua reta com forte inclinação:

```
> summary(df)
 fixed.acidity   volatile.acidity  citric.acid    residual.sugar     chlorides       free.sulfur.dioxide total.sulfur.dioxide
 Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900   Min.   :0.01200   Min.   : 1.00       Min.   :  6.00
 1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00
 Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200   Median :0.07900   Median :14.00       Median : 38.00
 Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539   Mean   :0.08747   Mean   :15.87       Mean   : 46.47
 3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00
 Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500   Max.   :0.61100   Max.   :72.00       Max.   :289.00
    density             pH          sulphates         alcohol         quality      quality_bin
 Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000   Length:1599
 1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000   Class :character
 Median :0.9968   Median :3.310   Median :0.6200   Median :10.20   Median :6.000   Mode  :character
 Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
 3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
 Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

Figura 2 - Estatística descritiva básica das variáveis numéricas

- **Estatísticas:**
  - Teor alcoólico varia de 8.4% a 14.9%;
  - A qualidade varia de 3 a 8;

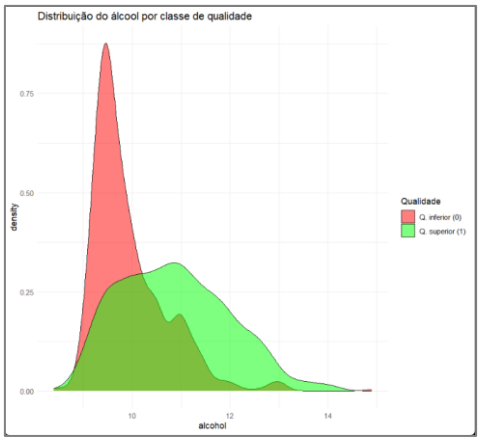

Figura 3 - Vinho de qualidade inferior tendem a ter menos álcool

- quality_bin: 744 registros de "Qualidade inferior" e 855 de "Qualidade superior", conforme mencionada a divisão das qualidades na nota 6 é uma forma de deixar variável balanceada.

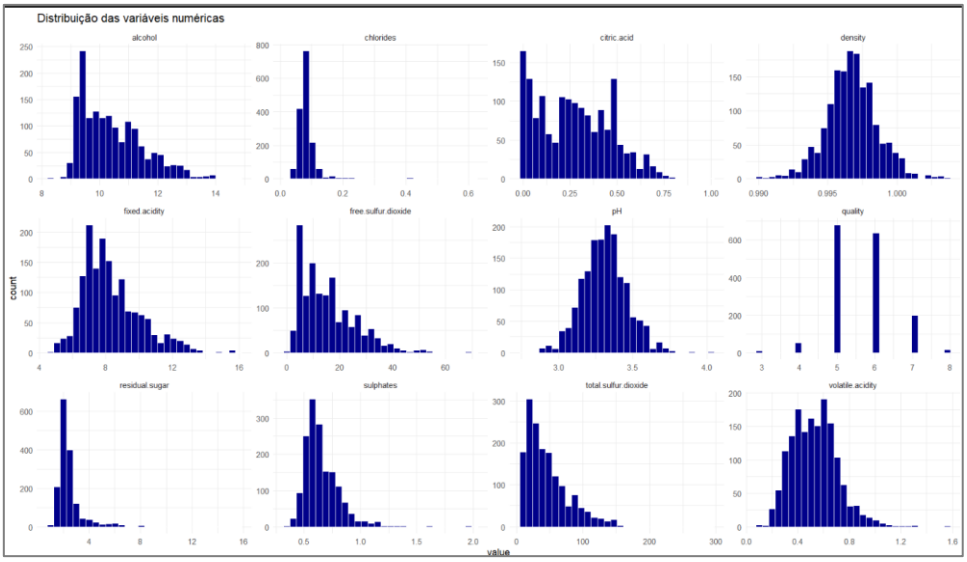

Figura 4 - Histogramas de todas as variáveis numéricas

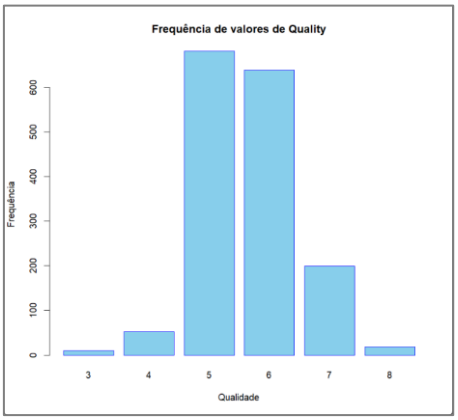

Figura 5 - A grande maioria dos registros ficam entre com a avaliação entre 5 e 6

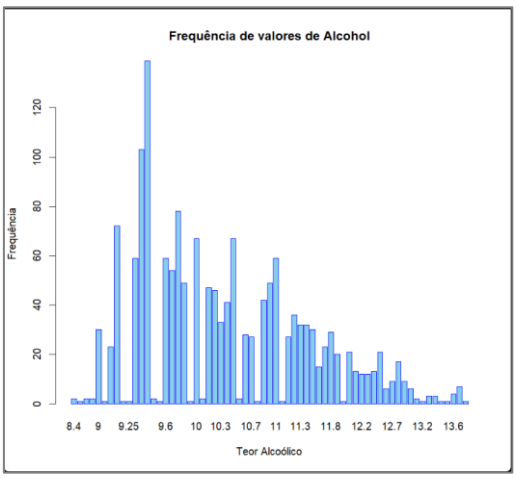

Figura 6 - Grande parte dos valores de álcool ficam entre 9 e 10

3. Testes de Normalidade

- **Teste aplicado:** Shapiro-Wilk (amostra reduzida, com menos de 2000 observações);
- **Variáveis testadas:**
    - alcohol: W = 0.92884, p < 2.2e-16
    - quality: W = 0.85759, p < 2.2e-16
    - density: W = 0.99087, p = 1.936e-08
    - fixed.acidity: W = 0.94203, p < 2.2e-16
- De modo geral, além das variáveis mencionadas acima, todas as restantes também retornaram p-valores muito baixos, indicando que nenhuma segue a distribuição normal;
- **Interpretação:** Todos os testes retornaram **p-valores muito baixos**, rejeitando a hipótese nula de normalidade. Apesar disso, a regressão linear ainda pode ser aplicada, pois o método é robusto a desvios de normalidade, principalmente com grandes amostras como esta.

4. Coeficiente de Correlação

- **Correlação entre:**
    - alcohol e quality: **r = 0.476;**
    - alcohol e density: **r = -0.496.**
- **Interpretação:**
    - A correlação entre alcohol e quality é **moderada e positiva**, indicando que vinhos com maior teor alcoólico tendem a ter melhor qualidade;
    - A correlação entre alcohol e density é **moderada e negativa**, sugerindo que vinhos com maior teor alcoólico tendem a ter menor densidade.

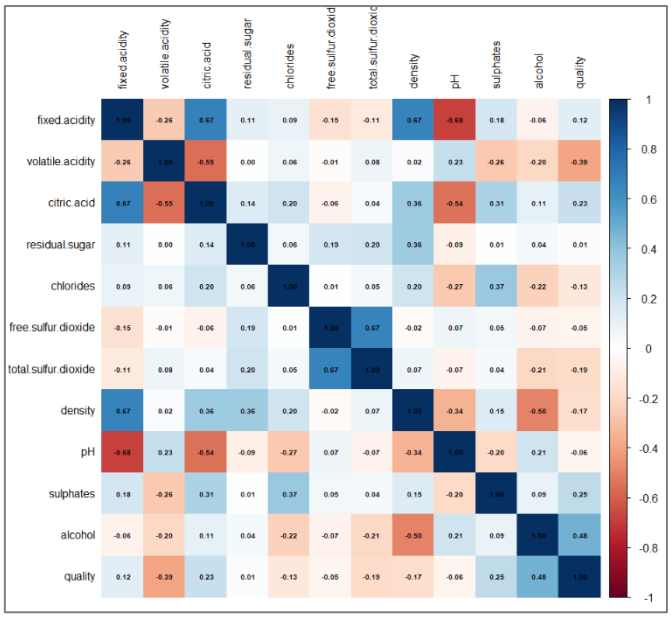

Figura 7 - Matriz de correlação entre as variáveis

5. Regressão Linear Simples (predição)

- **Modelo ajustado:** quality ~ alcohol
- **Resumo dos resultados:**
  - **Intercepto:** 1.875
  - **Coeficiente de inclinação (alcohol):** 0.361 → indica que a cada aumento de 1% no teor alcoólico, a qualidade média do vinho tende a aumentar em 0.361 pontos.
  - Ambos os coeficientes são estatisticamente significativos (**p < 2e-16**).
- **Métricas de avaliação:**
  - **R^2 = 0.2267** → cerca de 22,7% da variação da variável quality é explicada pelo teor alcoólico.
  - **Erro absoluto padrão (MAE) = 0.562** → indica que o modelo erra, em média, meio ponto na nota de qualidade;
  - **Erro padrão residual (RMSE) = 0.7104** → representa o desvio médio das previsões em relação aos valores reais da qualidade.
- **Visualização:**
  - Geramos um gráfico de dispersão com os pontos reais e a **linha de regressão ajustada**, reforçando a tendência positiva entre álcool e qualidade.

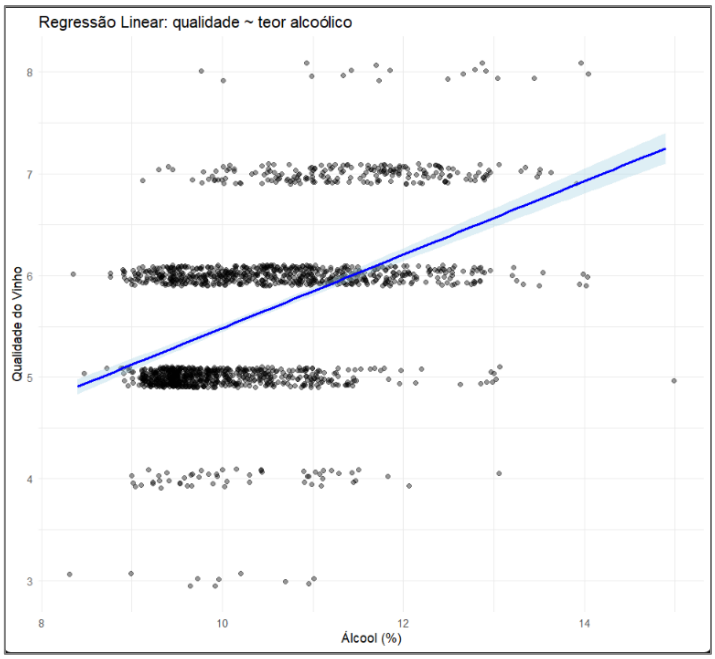

Figura 8 - Gráfico da regressão linear

6. Regressão Logística (classificação)

- **Variável-alvo:** quality_bin (binária), com as classes "Qualidade inferior" e "Qualidade superior";
- **Variável preditora:** alcohol (teor alcoólico).
- **Modelo ajustado:**
    - Treinamos um modelo de **regressão logística** com a variável alcohol como única preditora. O modelo estima a probabilidade de um vinho ser classificado como de qualidade superior;
    - **Coeficiente de alcohol:** 1.0556 ($p < 0.001$):
      → Um aumento no teor alcoólico está positivamente associado à chance de o vinho ser de qualidade superior;
    - **Intercepto:** -10.7630;
    - **Deviance residual:** 1865 (redução em relação à null deviance de 2209);
    - **AIC:** 1869.
- **Avaliação do modelo:**
    - **Matriz de confusão:**

PREDITO / REAL	QUALIDADE INFERIOR	QUALIDADE SUPERIOR
QUALIDADE INFERIOR	533	263
QUALIDADE SUPERIOR	211	592

- **Acurácia:** 70,11%.
- **Interpretação:** O modelo apresentou um desempenho consistente, com acurácia de 70,11% utilizando apenas o teor alcoólico como preditor. O coeficiente significativo e positivo de alcohol confirma sua importância na previsão da qualidade do vinho. Apesar da simplicidade do modelo, os resultados indicam uma boa separabilidade entre as classes, sugerindo que o teor alcoólico é um fator relevante para distinguir vinhos de qualidade superior.