

CENTRO UNIVERSITÁRIO UNIDOMBOSCO

POLO CAMPO BELO / MG

CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL

Projeto Integrador V A

Aluno: JULIANO FRANÇA DA MATA

RA: 23200190

CAMPO BELO / MG

Junho

JULIANO MATA

Projeto Integrador V A

Relatório apresentado como conclusão do Projeto Integrador Módulo V A do Curso
Ciência de Dados e Inteligência Artificial.

Campo Belo

2025

SUMÁRIO

1. INTRODUÇÃO.....	03
2. DESENVOLVIMENTO.....	04
3. CONCLUSÃO.....	05
REFERÊNCIAS.....	07
APÊNDICES.....	08

1. INTRODUÇÃO

A linguagem de programação Python, lançada em 1991, destaca-se por permitir o desenvolvimento de soluções com menos linhas de código em comparação a outras linguagens. Sua simplicidade, legibilidade e vasta comunidade contribuíram para sua ampla adoção em diversas áreas da tecnologia. Atualmente, o Python está integrado em praticamente todas as novas tecnologias, sendo utilizado em aplicações web, mobile, data science, machine learning, blockchain, entre outras.

Desde 2009, o Python tornou-se a linguagem padrão do curso de Ciência da Computação do Massachusetts Institute of Technology (MIT), o que reforça sua importância no meio acadêmico e científico. Além disso, foi eleito "Linguagem do Ano" pelo índice TIOBE em diferentes ocasiões: 2007, 2010, 2018 e 2020, demonstrando sua constante evolução e relevância no cenário da programação contemporânea (TIOBE, 2020).

O presente projeto integrador tem como base um tutorial publicado no site Data Flair, que aborda o pré-processamento e visualização de dados com Python. A primeira etapa do estudo contempla técnicas como normalização, padronização, transformação e binarização de dados, utilizando os pacotes Pandas e Scikit-learn. Já a segunda etapa do tutorial explora a visualização de dados por meio de histogramas e gráficos de densidade.

A base de dados utilizada refere-se à análise de vinhos portugueses e está disponível no repositório UCI Machine Learning Repository, na versão winequality-red.csv, a qual também se encontra nos arquivos da disciplina na plataforma Canvas. Como parte da proposta do projeto, será elaborado um relatório técnico no formato ABNT, contendo os programas desenvolvidos e os gráficos gerados a partir das operações realizadas sobre os dados.

2. DESENVOLVIMENTO

As atividades desenvolvidas durante o projeto seguiram uma estrutura lógica e simples, visando à criação de um modelo analítico básico sem maiores análises.

2.1 Importação das bibliotecas

Bibliotecas necessárias para análise de dados e visualizações. [Ver Apêndice , Fig. 01.](#)

2.2 Definir o caminho do arquivo CSV

O caminho do arquivo com os dados dos vinhos definido como uma única “string raw”. [Ver Apêndice , Fig. 02.](#)

2.3 Carregamento dos Dados

Carregar os dados do arquivo CSV observando o separador “;” usado nesse dataset. [Ver Apêndice , Fig. 03.](#)

2.4 Visualização das primeiras linhas

Para ter uma ideia da estrutura dos dados. [Ver Apêndice , Fig. 04 e 05.](#)

2.5 Exibir informações gerais

Todas as colunas possuem **1599 valores não nulos**, o que indica que **não há dados ausentes** no dataset. A maioria dos dados está no formato float64, com exceção da coluna quality, que está no formato int64. O uso de memória do DataFrame é de aproximadamente **150,0 KB**. [Ver Apêndice , Fig. 06 e 07.](#)

2.6 Estatísticas descritivas

Média, desvio padrão, mínimo, máximo etc. [Ver Apêndice , Fig. 08 e 09.](#)

2.7 Valores ausentes

Verificar/ confirmar se há valores ausentes em alguma coluna. [Ver Apêndice , Fig. 10 e 11.](#)

2.8 Visualização da distribuição da qualidade dos vinhos

Mostra que a qualidade dos vinhos está concentrada principalmente nas notas intermediárias (5 e 6), com poucos exemplos de vinhos muito bons ou muito ruins. [Ver Apêndice , Fig. 12 e 13.](#)

2.9 Matriz de correlação

A matriz de correlação indica que **álcool**, **sulfatos** e **ácido cítrico** estão associados positivamente com a **qualidade do vinho**, enquanto **acidez volátil** está negativamente associada. Isso pode ajudar a direcionar os esforços para identificar as características mais relevantes na produção de vinhos de melhor qualidade. [Ver Apêndice , Fig. 14 e 15.](#)

2.10 Distribuição da acidez fixa

A **acidez fixa** tende a se concentrar em uma faixa média, sendo incomum encontrar vinhos com valores muito baixos ou muito altos dessa variável. Isso indica que a acidez fixa segue um padrão consistente na produção da maioria dos vinhos analisados. [Ver Apêndice , Fig. 16 e 17.](#)

2.11 Boxplots de variáveis importantes

As variáveis volatile acidity e residual sugar possuem alta quantidade de outliers, o que pode afetar análises estatísticas e modelos preditivos.

Já alcohol e pH têm distribuição mais estável, com menos outliers. Esses dados ajudam a decidir se será necessário tratar ou remover outliers antes de análises mais avançadas. [Ver Apêndice , Fig. 18 e 19.](#)

2.12 Teor alcoólico e qualidade do vinho

Existe uma correlação positiva entre teor alcoólico e qualidade do vinho: vinhos de melhor qualidade geralmente têm maior teor alcoólico. Essa informação pode ser usada em modelos de predição de qualidade com base em variáveis físico-químicas. [Ver Apêndice , Fig. 20 e 21.](#)

2.13 Agrupar a variável “quality” em categorias

O dataset está desbalanceado, com forte concentração na categoria de qualidade média.

Isso pode impactar análises e modelos preditivos, que tenderão a prever a classe média com mais frequência. [Ver Apêndice , Fig. 22 e 23.](#)

2.14 Machine Learning

Preparação dos dados, normalização, divisão em conjunto treino e teste, treinar um modelo de classificação, avaliar o modelo. [Ver Apêndice , Fig. 24 a 30.](#)

3. CONCLUSÃO

- A análise inicial forneceu uma visão geral do conjunto de dados, incluindo a estrutura e estatísticas básicas.

- A matriz de correlação identificou variáveis com maior influência na qualidade do vinho.
- A distribuição da acidez e da qualidade revelou padrões importantes.
- Boxplots ajudaram a detectar outliers que podem impactar modelos futuros.
- A relação entre teor alcoólico e qualidade mostrou uma tendência clara de maior qualidade com mais álcool.
- Também foi aplicado um modelo de Random Forest para prever a qualidade do vinho com base nas variáveis do dataset.
- Neste caso, a qualidade foi agrupada em três categorias (baixa, média e alta) para facilitar a predição.
- As métricas de avaliação ajudam a entender a performance e precisão do modelo com essa nova abordagem.

Próximas etapas recomendadas:

- Otimização de hiperparâmetros do modelo (com GridSearchCV, por exemplo).
- Teste com outros algoritmos como SVM, KNN, ou redes neurais
- Análise de importância das variáveis para compreender quais características mais influenciam na qualidade.

REFERÊNCIAS

FACULDADE DOM BOSCO. *Curso de Ciência de Dados e Inteligência Artificial*. Disponível em: <https://faculdaadedombosco.edu.br/>. Acesso em: 6 jun. 2025.

GITHUB – JULIANOMATA. *Repositório pessoal de projetos*. Disponível em: <https://github.com/JulianoMata>. Acesso em: 6 jun. 2025.

MICROSOFT. *Visual Studio Code*. Disponível em: <https://code.visualstudio.com/>. Acesso em: 6 jun. 2025.

PANDAS. *The pandas library*. Disponível em: <https://pandas.pydata.org/>. Acesso em: 6 jun. 2025.

SCIKIT-LEARN. *Scikit-learn: Machine Learning in Python*. Disponível em: <https://scikit-learn.org/>. Acesso em: 6 jun. 2025.

SEABORN. *Seaborn: Statistical Data Visualization*. Disponível em: <https://seaborn.pydata.org/>. Acesso em: 6 jun. 2025.

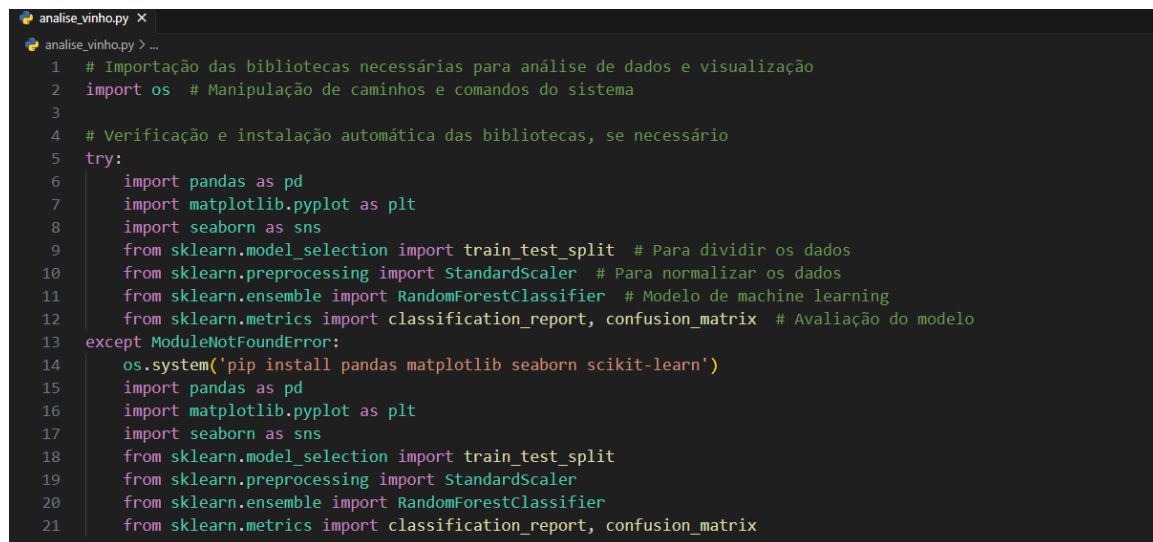
UNIVERSITY OF CALIFORNIA IRVINE. *Wine Quality Dataset*. Disponível em: <https://archive.ics.uci.edu/ml/datasets/wine+quality>. Acesso em: 6 jun. 2025.

APÊNDICES

A seguir, são apresentados os materiais complementares que ilustram as etapas do projeto. Todos os apêndices foram referenciados ao longo do texto para melhor compreensão do processo de desenvolvimento.

Apêndice – Capturas de tela do VSCODE

Fig. 01

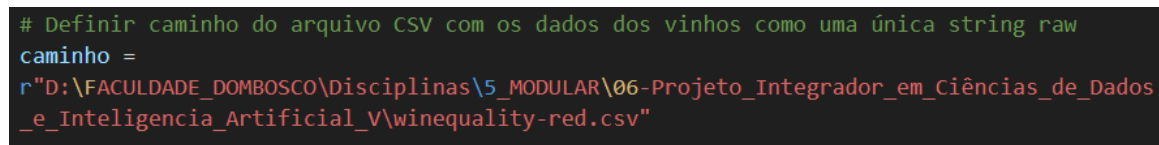


```

1  # Importação das bibliotecas necessárias para análise de dados e visualização
2  import os # Manipulação de caminhos e comandos do sistema
3
4  # Verificação e instalação automática das bibliotecas, se necessário
5  try:
6      import pandas as pd
7      import matplotlib.pyplot as plt
8      import seaborn as sns
9      from sklearn.model_selection import train_test_split # Para dividir os dados
10     from sklearn.preprocessing import StandardScaler # Para normalizar os dados
11     from sklearn.ensemble import RandomForestClassifier # Modelo de machine learning
12     from sklearn.metrics import classification_report, confusion_matrix # Avaliação do modelo
13 except ModuleNotFoundError:
14     os.system('pip install pandas matplotlib seaborn scikit-learn')
15     import pandas as pd
16     import matplotlib.pyplot as plt
17     import seaborn as sns
18     from sklearn.model_selection import train_test_split
19     from sklearn.preprocessing import StandardScaler
20     from sklearn.ensemble import RandomForestClassifier
21     from sklearn.metrics import classification_report, confusion_matrix
22

```

Fig. 02

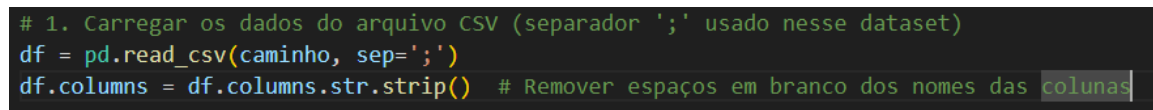


```

# Definir caminho do arquivo CSV com os dados dos vinhos como uma única string raw
caminho =
r"D:\FACULDADE_DOMBOSCO\Disciplinas\5_MODULAR\06-Projeto_Integrador_em_Ciências_de_Dados
_e_Inteligencia_Artificial_V\winequality-red.csv"

```

Fig. 03



```

# 1. Carregar os dados do arquivo CSV (separador ';' usado nesse dataset)
df = pd.read_csv(caminho, sep=';')
df.columns = df.columns.str.strip() # Remover espaços em branco dos nomes das colunas

```

Fig. 04

```
# 2. Visualizar as primeiras linhas do dataset para ter uma ideia da estrutura dos dados
print("Visualizando as primeiras linhas do dataset:")
print(df.head())
```

Fig. 05

```
Visualizando as primeiras linhas do dataset:
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

Fig. 06

```
# 3. Exibir informações gerais sobre os dados: tipo de dado, valores não nulos etc.
print("\nInformações gerais do dataset:")
print(df.info())
```

Fig. 07

```
Informações gerais do dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
#   Column                Non-Null Count  Dtype
---  -
0   fixed acidity          1599 non-null   float64
1   volatile acidity       1599 non-null   float64
2   citric acid            1599 non-null   float64
3   residual sugar         1599 non-null   float64
4   chlorides              1599 non-null   float64
5   free sulfur dioxide    1599 non-null   float64
6   total sulfur dioxide   1599 non-null   float64
7   density               1599 non-null   float64
8   pH                    1599 non-null   float64
9   sulphates             1599 non-null   float64
10  alcohol               1599 non-null   float64
11  quality               1599 non-null   int64
dtypes: float64(11), int64(1)
memory usage: 150.0 KB
None
```

Fig. 08

```
# 4. Estatísticas descritivas básicas: média, desvio padrão, mínimo, máximo etc.
print("\nEstatísticas descritivas:")
print(df.describe())
```

Fig. 09

Estatísticas descritivas:												
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	8.319637	0.527821	0.278976	2.538806	0.087467	15.874922	46.467792	0.996747	3.311113	0.658149	10.422983	5.636023
std	1.741096	0.179060	0.194881	1.489928	0.047065	10.460157	32.895324	0.001887	0.154386	0.169587	1.065668	0.807569
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070	2.740000	0.330000	8.400000	3.000000
25%	7.100000	0.300000	0.000000	1.900000	0.070000	7.000000	22.000000	0.995600	3.210000	0.550000	9.500000	5.000000
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	0.996750	3.310000	0.620000	10.200000	6.000000
75%	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000	62.000000	0.997835	3.400000	0.730000	11.100000	6.000000
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289.000000	1.003690	4.010000	2.000000	14.900000	8.000000

Fig. 10

```
# 5. Verificar se há valores ausentes em alguma coluna
total_nulos = df.isnull().sum()
print("\nValores ausentes por coluna:")
print(total_nulos)
if total_nulos.sum() == 0:
    print("\nNão há valores ausentes no dataset.")
```

Fig. 11

```
Valores ausentes por coluna:
fixed acidity      0
volatile acidity   0
citric acid        0
residual sugar     0
chlorides          0
free sulfur dioxide 0
total sulfur dioxide 0
density           0
pH                0
sulphates         0
alcohol           0
quality           0
dtype: int64

Não há valores ausentes no dataset.
```

Fig. 12

```
# 6. Visualizar a distribuição da variável alvo 'quality' (qualidade dos vinhos)
plt.figure(figsize=(8, 5))
sns.countplot(x='quality', data=df, hue='quality', palette='viridis', legend=False)
plt.title("Distribuição da Qualidade do Vinho")
plt.xlabel("Qualidade")
plt.ylabel("Contagem")
plt.show()
```

Fig. 13

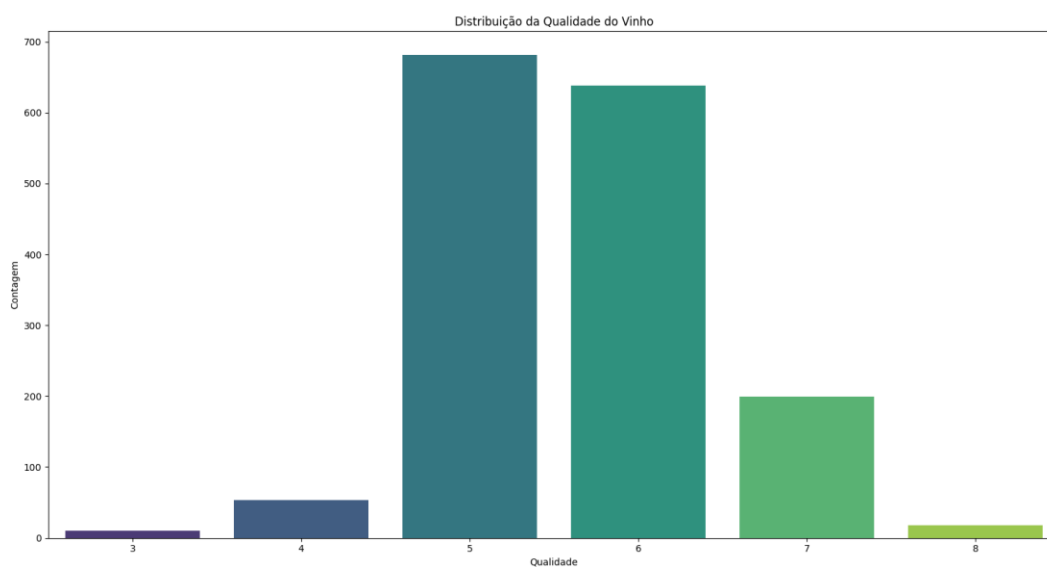


Fig. 14

```
# 7. Matriz de correlação: identificar relações entre variáveis numéricas
plt.figure(figsize=(12, 10))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm', fmt='.2f', linewidths=0.5)
plt.title("Matriz de Correlação das Variáveis")
plt.show()
```

Fig. 15

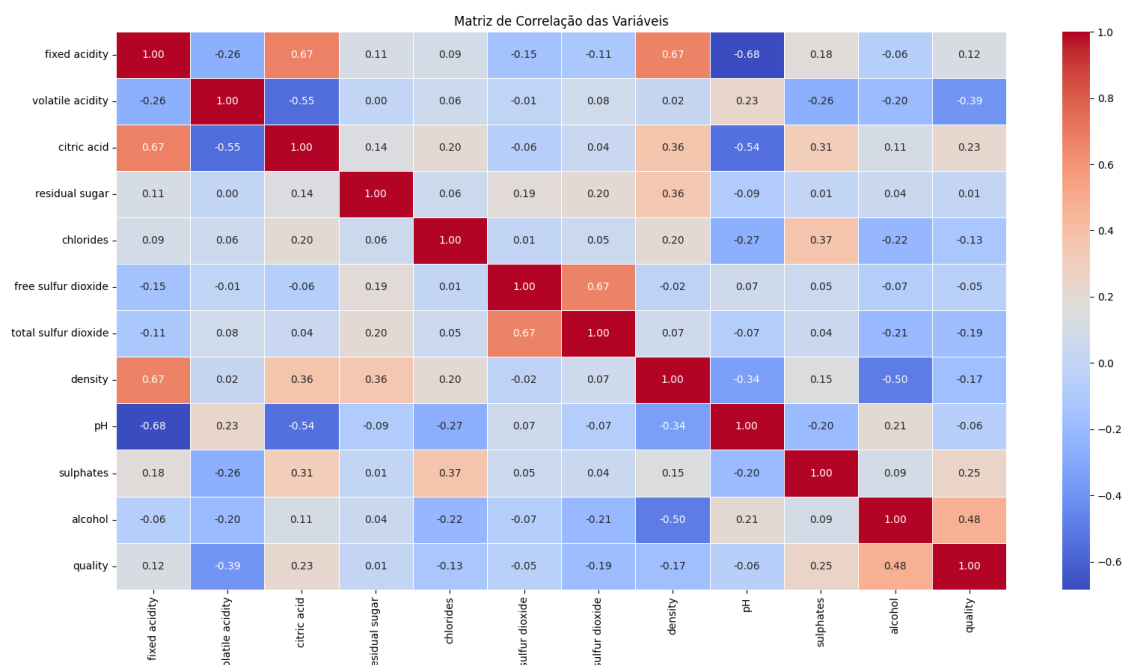


Fig. 16

```
# 8. Distribuição da acidez fixa para entender sua densidade
plt.figure(figsize=(8, 5))
df['fixed acidity'].hist(bins=30, color='skyblue', edgecolor='black')
plt.title("Distribuição da Acidez Fixa")
plt.xlabel("Acidez Fixa")
plt.ylabel("Frequência")
plt.show()
```

Fig. 17

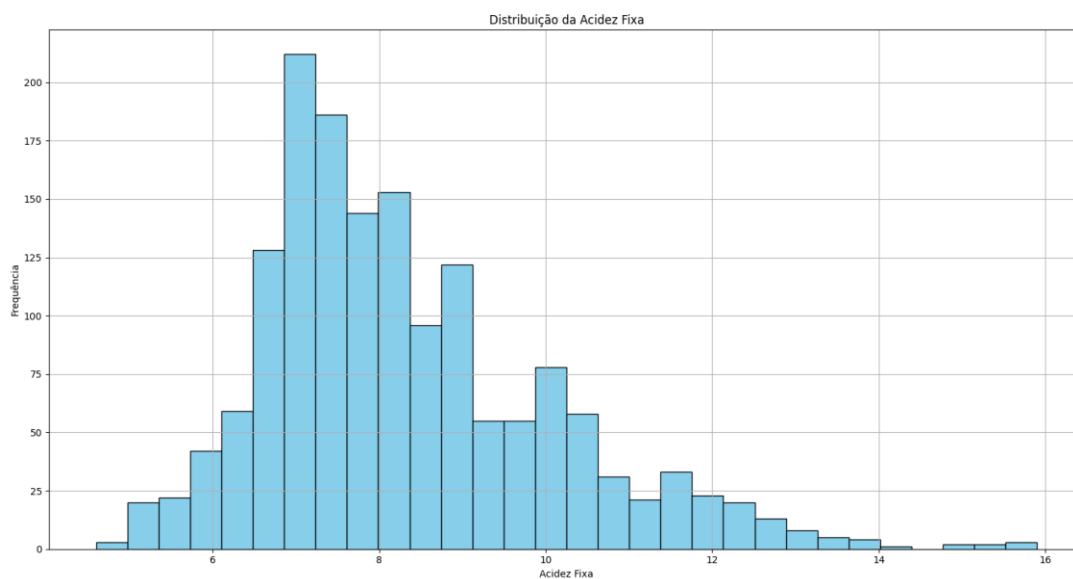


Fig. 18

```
# 9. Boxplots para detectar outliers em variáveis importantes
plt.figure(figsize=(12, 6))
sns.boxplot(data=df[['alcohol', 'volatile acidity', 'residual sugar', 'pH']], palette='Set2')
plt.title("Boxplot de Variáveis para Identificação de Outliers")
plt.show()
```

Fig. 19

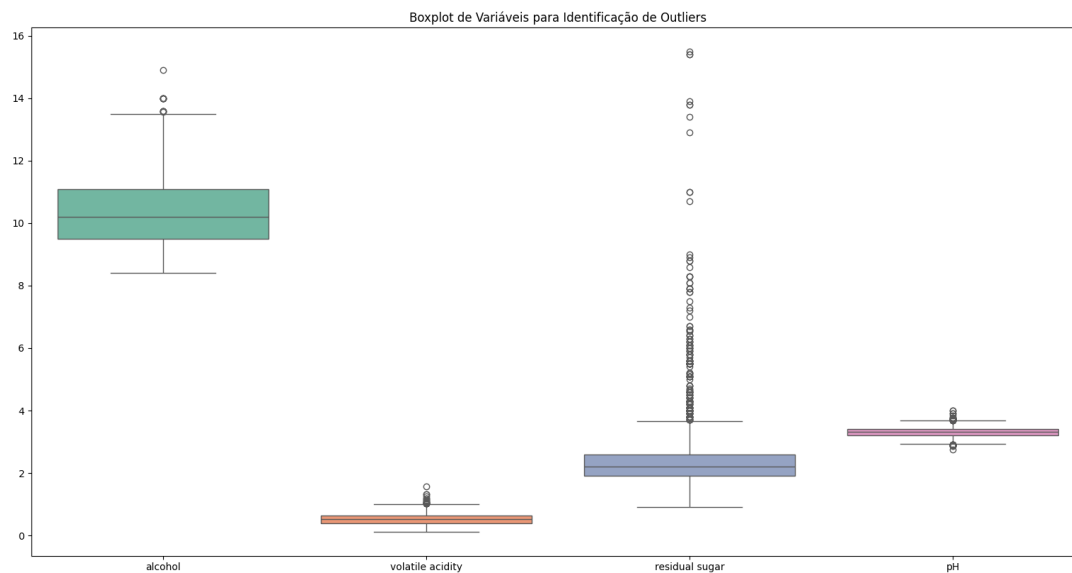


Fig. 20

```
# 10. Relação entre teor alcoólico e qualidade do vinho
plt.figure(figsize=(10, 6))
sns.boxplot(x='quality', y='alcohol', data=df, hue='quality', palette='magma', legend=False)
plt.title("Relação entre Teor Alcoólico e Qualidade")
plt.xlabel("Qualidade")
plt.ylabel("Teor Alcoólico")
plt.show()
```

Fig. 21

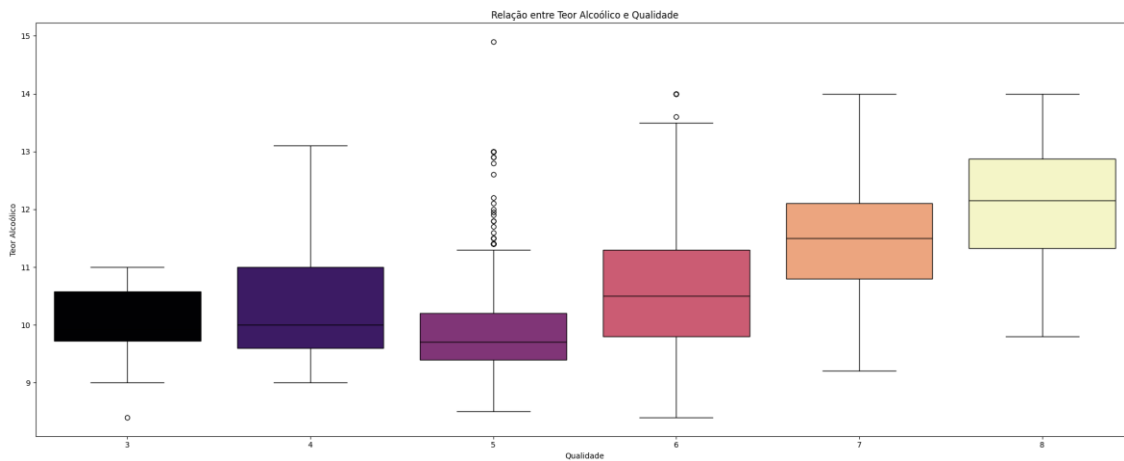


Fig. 22

```
# 11. Agrupar a variável 'quality' em categorias: baixa (0), média (1), alta (2)
def categorizar_qualidade(valor):
    if valor <= 4:
        return 0 # baixa
    elif valor <= 6:
        return 1 # média
    else:
        return 2 # alta

df['qualidade_cat'] = df['quality'].apply(categorizar_qualidade)

# Visualizar nova distribuição após o agrupamento
plt.figure(figsize=(8, 5))
sns.countplot(x='qualidade_cat', data=df, hue='qualidade_cat', palette='plasma', legend=False)
plt.title("Distribuição Agrupada da Qualidade do Vinho")
plt.xlabel("Categorias de Qualidade (0=Baixa, 1=Média, 2=Alta)")
plt.ylabel("Contagem")
plt.show()
```

Fig. 23

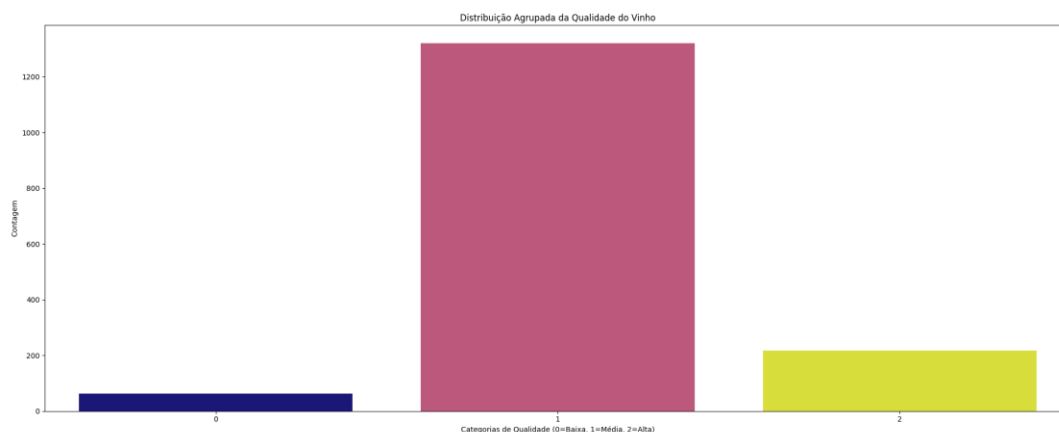


Fig. 24

```
# 12. Preparação dos dados para o modelo de machine learning
x = df.drop(['quality', 'qualidade_cat'], axis=1) # Atributos (features)
y = df['qualidade_cat'] # Variável alvo categorizada

# Normalização dos dados
scaler = StandardScaler()
x_scaled = scaler.fit_transform(x)

# Dividir em conjunto de treino e teste
x_train, x_test, y_train, y_test = train_test_split(x_scaled, y, test_size=0.2, random_state=42)
```

Fig. 25

```
# 13. Treinar um modelo de classificação (Random Forest)
modelo = RandomForestClassifier(n_estimators=100, random_state=42)
modelo.fit(x_train, y_train)
```

Fig. 26

```
# 14. Avaliar o modelo
y_pred = modelo.predict(x_test)
print("\nMatriz de Confusão:")
print(confusion_matrix(y_test, y_pred))
print("\nRelatório de Classificação:")
print(classification_report(y_test, y_pred, zero_division=0))
```

Fig. 27

```
Matriz de Confusão:
[[ 0 11  0]
 [ 0 250 12]
 [ 0 21 26]]
```


Fig. 28

```
# Isso significa:
# - Dos 11 vinhos realmente da Classe 0, todos foram classificados
  erroneamente como Classe 1.
# - Dos 262 vinhos da Classe 1, 250 foram corretamente classificados e
  12 foram confundidos com Classe 2.
# - Dos 47 vinhos da Classe 2, 26 foram corretamente classificados, mas
  21 foram confundidos com Classe 1.

"""Embora a Classe 0 não tenha sido corretamente classificada pelo
modelo, esse comportamento é aceitável dentro do escopo educacional do
projeto. Futuramente, técnicas de balanceamento de dados ou ajuste de
hiperparâmetros poderiam ser exploradas para melhorar a performance."""
```

Fig. 29

Relatório de Classificação:				
	precision	recall	f1-score	support
0	0.00	0.00	0.00	11
1	0.89	0.95	0.92	262
2	0.68	0.55	0.61	47
accuracy			0.86	320
macro avg	0.52	0.50	0.51	320
weighted avg	0.83	0.86	0.84	320

Fig. 30

```
# Relatório de Classificação:
# Mostra o desempenho do modelo em cada classe (0 = Ruim, 1 = Regular,
# 2 = Boa)
# Métricas:
# - precision: acertos entre os que foram previstos como aquela classe
# - recall: acertos entre os que realmente pertencem àquela classe
# - f1-score: equilíbrio entre precision e recall
# - support: total real de exemplos daquela classe

# Observações:
# - O modelo teve ótimo desempenho na classe "Regular" (classe 1)
# - Teve dificuldades nas classes "Ruim" (0) e "Boa" (2), possivelmente
#   por poucos exemplos (desequilíbrio de classes)
# - A acurácia geral foi de 86%, o que é um bom resultado para um
#   projeto educacional
# - A média ponderada das métricas também indica desempenho consistente
#   no geral

""" Por ser um exercício inicial, não faremos ajustes finos no modelo
neste momento"""
```