UNIVERSIDADE FEDERAL
DE MINAS GERAIS

EXERCÍCIO
ANÁLISE EXPLORATÓRIA DE DADOS

ALUNO: FELIPPE VELOSO MARINHO
MATRÍCULA: 2021072260
DISCIPLINA: APRENDIZADO DE MÁQUINA

Objetivo: Utilizando o dataset Wine, realizar uma análise exploratória completa com o objetivo de identificar possíveis características irrelevantes e redundantes. O aluno deverá usar ferramentas como matriz de correlações, scatterplots e boxplots.

1. Carregamento dos Dados:

O conjunto de dados do Wine é referente a um conjunto de dados de classificação multiclasse clássico.

Classes	3
Samples per class	[59,71,48]
Samples total	178
Dimensionality	13
Features	real, positive

Os nomes das variáveis de entrada estão descritas abaixo.

```
['alcohol', 'malic_acid', 'ash', 'alcalinity_of_ash', 'magnesium',
'total_phenols', 'flavanoids', 'nonflavanoid_phenols',
'proanthocyanins', 'color_intensity', 'hue',
'od280/od315_of_diluted_wines', 'proline']
```

A varável de saida utlizada é o “target”.

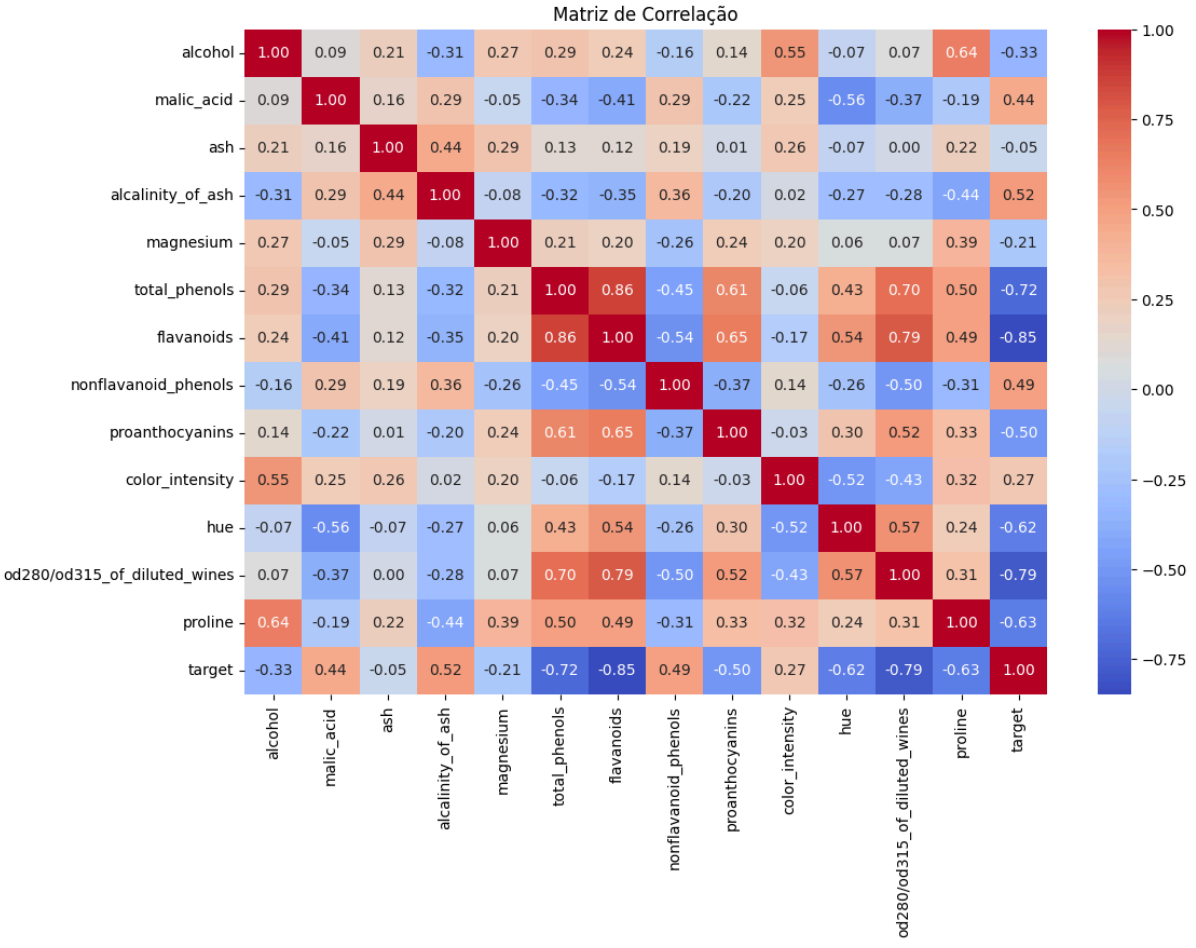
2. Descrição Estatística dos Dados:
   Através da função “df.describe()” temos o seguinte retorno:

```
# Resumo estatístico
df.describe()
```

	alcohol	malic_acid	ash	alcalinity_of_ash	magnesium	total_phenols	flavanoids
count	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000
mean	13.000618	2.336348	2.366517	19.494944	99.741573	2.295112	2.029270
std	0.811827	1.117146	0.274344	3.339564	14.282484	0.625851	0.998859
min	11.030000	0.740000	1.360000	10.600000	70.000000	0.980000	0.340000
25%	12.362500	1.602500	2.210000	17.200000	88.000000	1.742500	1.205000
50%	13.050000	1.865000	2.360000	19.500000	98.000000	2.355000	2.135000
75%	13.677500	3.082500	2.557500	21.500000	107.000000	2.800000	2.875000
max	14.830000	5.800000	3.230000	30.000000	162.000000	3.880000	5.080000

Os dados exibidos através do comando nos informam a média, desvio padrão, valores mínimos e máximos entre outras informações como 75, 50 e 25 percentis. Sendo 50, a mediana.

3. Matriz de Correlações:



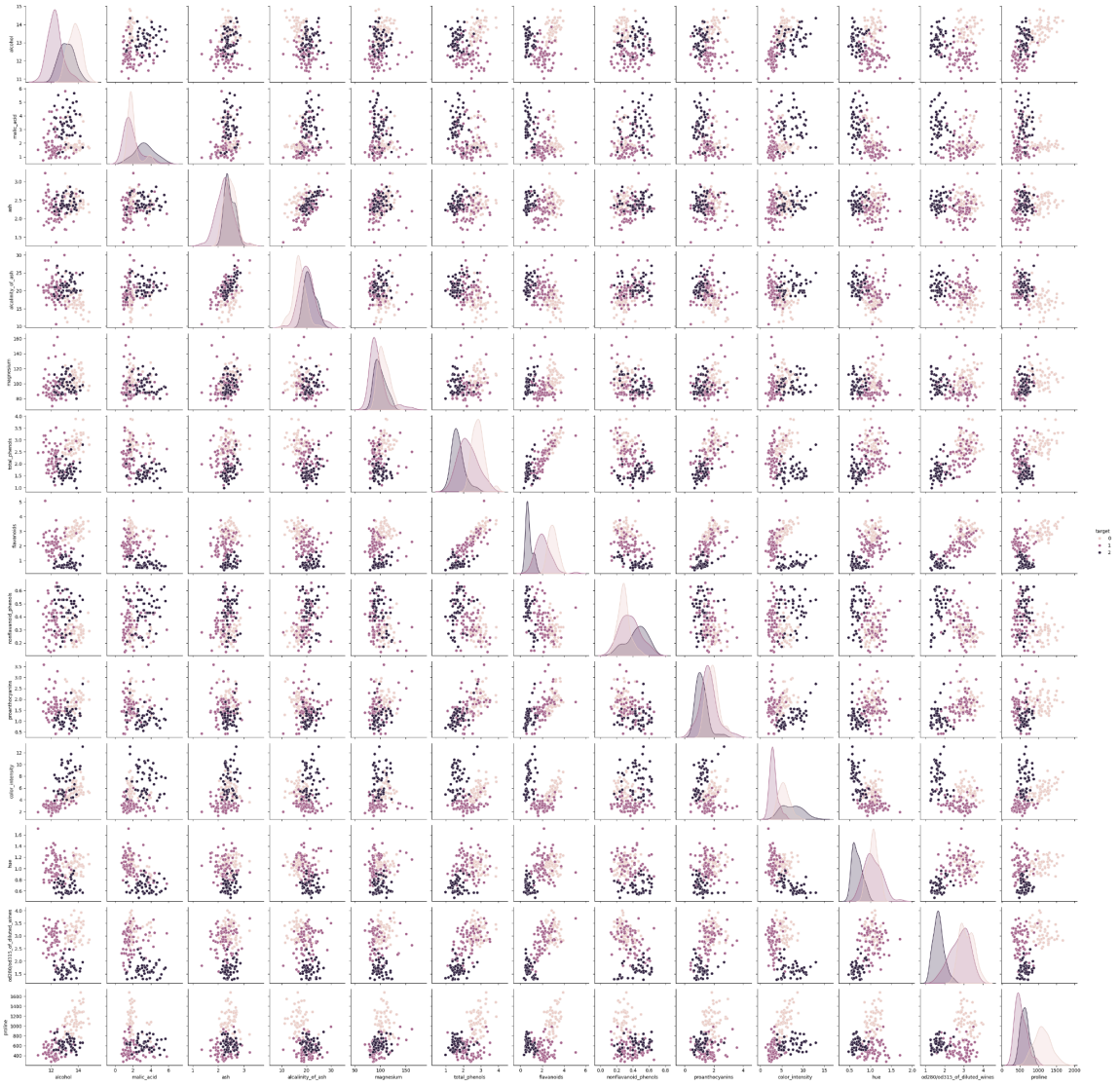
Analisando a matriz de correlação acima, é visível que as variáveis com correlação forte (próxima de 1 ou -1) ou com alta correlação entre si são representadas pelas correlações de:

- total-penols e flavanoids, com 0.86 de correlação.
- od280/od315_of_diluted_wines e flavinoids, com 79 de correlação
- od280/od315_of_diluted_wines e total-penols com, 70 de correlação

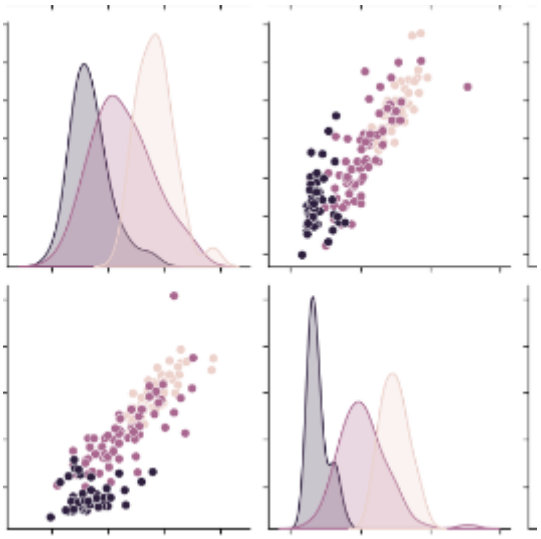
Uma alta correlação entre os dados pode representar que elas são redundantes e consideradas descartáveis.

Considerando as correlações com a saída (target), temos altas relações demonstradas, porém, nesse caso, uma alta relação pode significar que existe uma forte influência no resultado.

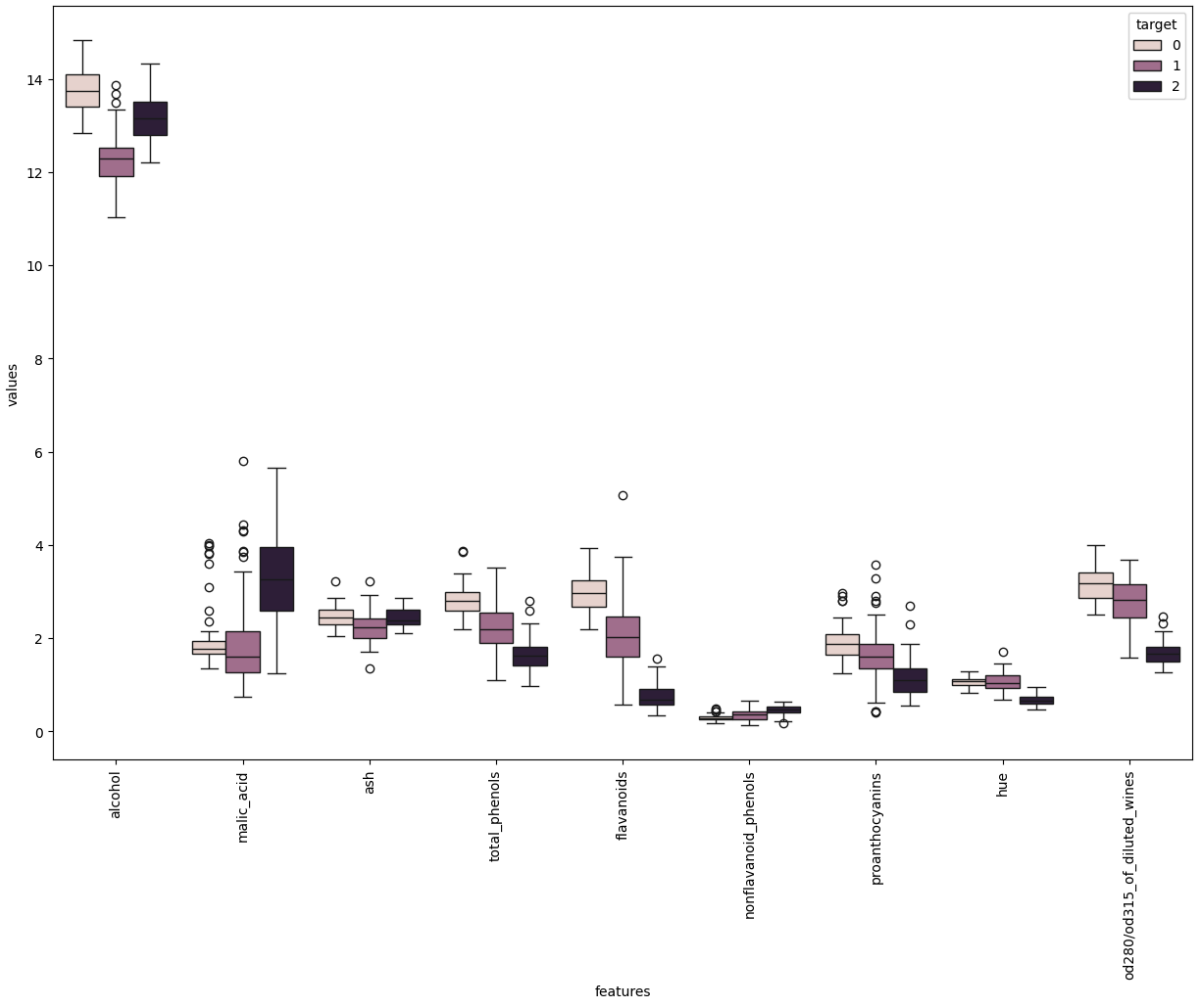
4. Matriz de Scatterplots:



Verificando a matriz scatterplots é possível perceber rapidamente que as variáveis que apresentam maior distinção são as de flavanoids e non-falvanoid-phenols.



5. Boxplots:
   Retirando algumas variáveis para melhor visualização e padronização dos dados, podemos ver que dentre as variáveis observadas, flavanoids novamente apresenta boas características, sendo que nenhuma a distribuição da variável pouco sobrepostas. Retirar esses dados poderia ser feito através de uma normalização de forma mais rebuscada.



6. Conclusão:

1. Quais variáveis apresentam alta correlação entre si? Explique por que você acredita que são redundantes.

- flavanoids e total_phenols = 0,86
- target e flavanoids = -0,85
- OD280/OD315 of diluted wines e flavanoids = 0,79
- proline e alcohol = 0,65
- flavanoids e hocyanins = 0,65
- target e proline = -0,65
- hue e malic_acid = -0,56
- proline e alcohol = 0,64

2. Há variáveis que, com base nos scatterplots e boxplots, parecem não ajudar a distinguir as classes? Quais você considera irrelevantes?

As variáveis nonflavanoid phenols, ash, magnesium, e total phenols têm padrões semelhantes entre as classes, o que sugere que elas não ajudam muito a

distinguir as classes. Isso é visto tanto nos scatterplots quanto nos boxplots, onde a separação entre as classes não é clara.

3. Quais variáveis você consideraria remover para otimizar o modelo de classificação, baseado nas observações feitas?

- nonflavanoid
- phenols
- ash
- magnesium