escola britânica de artes criativas & tecnologia

Módulo | Análise de Dados: Aprendizado de Máquina, Agrupamento

Caderno de **Exercícios**
Professor André Perez

Tópicos

1. Agrupamento;
2. Dados;
3. Treino;
4. Avaliação;
5. Predição.

Exercícios

1. Pinguins

Neste exercício, vamos utilizar uma base de dados com informações sobre flores do gênero iris. A idéia é agrupar as flores de acordo com suas características físicas (variáveis preditivas). Lembre-se das aulas, nós já temos uma idéia dos agrupamentos.

In [1]:
```
import sklearn
import numpy as np
import pandas as pd
import seaborn as sns
```

In [30]:
```
iris = sns.load_dataset('iris')
iris = iris.drop(['species'], axis=1)
```

In [ ]:
```
iris.head()
```

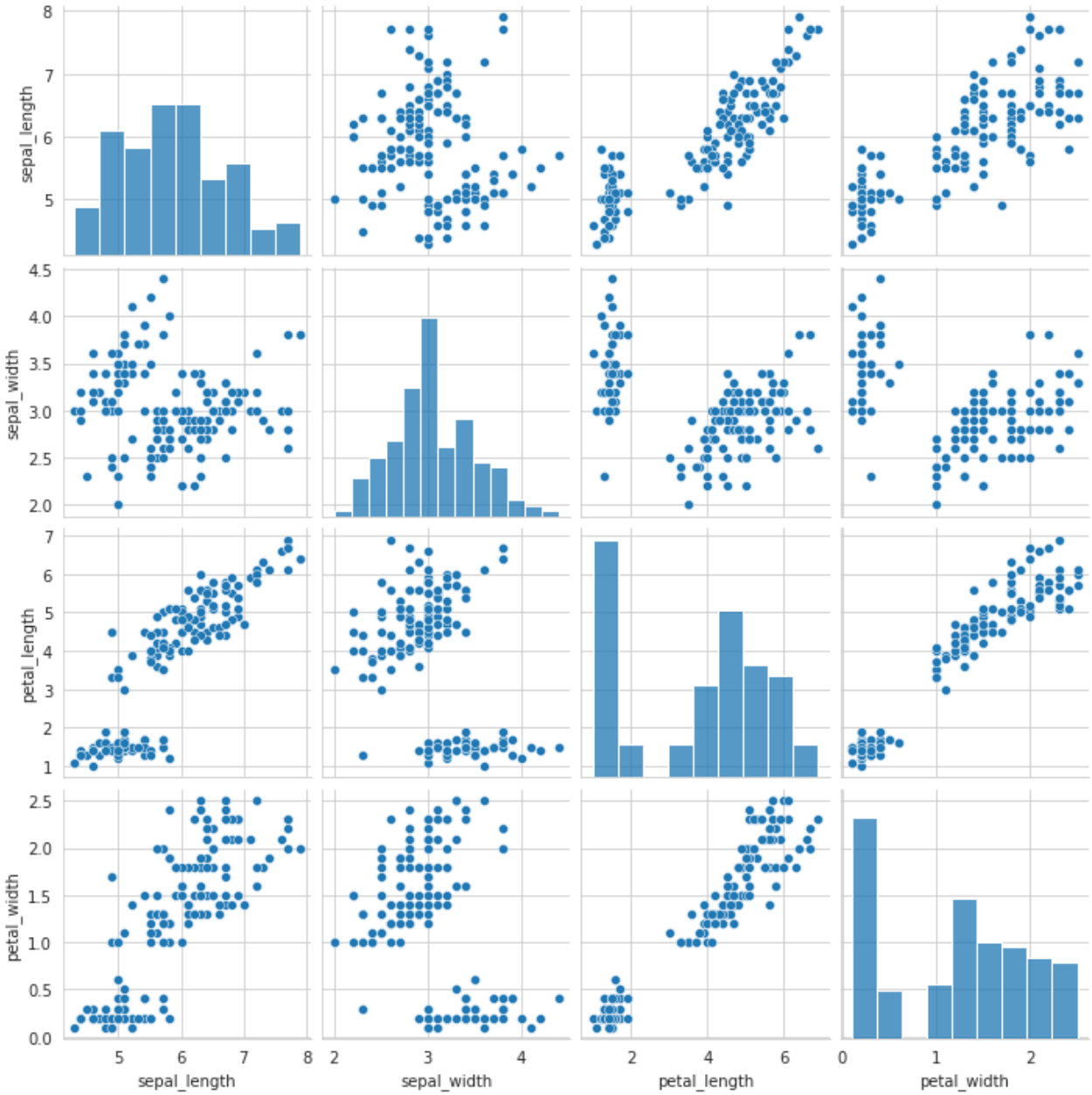
1.1. Analise exploratoria

Utilize os gráficos abaixo para entender melhor a relação entre os atributos da base de dados. Comente o que observou no gráfico.

- Atributos preditivos:

In [32]:
```
with sns.axes_style('whitegrid'):

  grafico = sns.pairplot(data=iris, palette="pastel")
```



Comentário: ?

2. Dados

2.1. Valores nulos

Avalie se a base de dados possui valores faltantes, se sim, utilize os conceitos da aula para trata-los.

In [ ]:
```
# resposta da questão 2.1
```

2.2. Variáveis numéricas

Identifique se existe a necessidade de escalar as variáveis numéricas. Se sim, crie uma nova coluna **padronizando** seus valores. A nova coluna deve ter o mesmo nome da coluna original acrescidade de "*_std*".

> **Nota**: Você não deve tratar a variável resposta.

In [ ]:
```
# resposta da questão 2.2
```

2.3. Limpeza

Caso você tenha escalado suas variáveis, descarte as colunas originais e mantenha apenas as variáveis preditivas com o sufixo *_std*", *_nom*" e "*_ord*".

In [ ]:
```
# resposta da questão 2.3
```

3. Modelagem

3.1. Treino

Treine 10 modelos de **k-médias** variando o número de clusters de 1 a 10. Para cada modelo treinado, salve o valor global do **wcss** em uma lista.

In [ ]:
```
# resposta da questão 3.1
```

3.2. Avaliação

Gere um gráfico de linha dos valores do **wcss** pelo **número de clusters**. Utilize o método do cotovelo para decidir o número final de clusters.

In [ ]:
```
# resposta da questão 3.2
```

3.3. Visualização

a) Utilizando o número de clusters final, adicione uma coluna chamada **cluster** no dataframe **iris** com o número do cluster que cada flor foi alocada.

In [ ]:
```
# resposta da questão 3.3.a
```

b) Gere a mesma visualização da sessão 1.1, agora passando como atributo **hue** a coluna **cluster**. Comente os resultados com base no valor esperado do número de clusters.

In [ ]:
```
# resposta da questão 3.3.b
```

Comentário: ?

4. Predição

4.1. Nova flor

Em qual cluster a flor abaixo seria alocada?

sepal_length	sepal_width	petal_length	petal_width
5.1	3.5	1.4	0.2

> **Atenção:** Lembre-se de pre-processar os atributos assim como nos exercício 2.2.

In [ ]:
```
# resposta da questão 4.1
```
