# Análise de dados do dataset Iris

**Camila Eleutério Gusmão**

Aprendizado de Máquina - PPGC-IC-UFF - 2021.01

## 1. Introdução: Carregando o dataset como um dataframe

O dataset Iris é uma base de dados disponibilizada no site da Universidade da Califórnia[1] contendo 150 amostras com informações sobre uma flor chamada Íris, que na verdade é o nome do gênero de 3 espécies: *Iris virginica*, *Iris setosa* e *Iris versicolor*. Nesta base de dados, temos 3 classes, que são justamente os tipos de íris, havendo 50 amostras de cada, além de suas características, expressas por meio de 4 atributos: tamanho da sépala (cm), largura da sépala (cm), tamanho da pétala (cm) e largura da pétala (cm) (ver Figura 1). Todos os atributos são quantitativos, isto é, representam quantidades e podem ser usados em operações aritméticas. A base também está disponível na biblioteca desenvolvida em Python chamada Scikit Learn[2], voltada para aprendizado de máquina.

```
1 from sklearn.datasets import load_iris
2
3 data = load_iris()
4 print('Atributos: ',data['feature_names'])
5 print('Classes: ', data['target_names'])
```

```
Atributos:  ['sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', 'petal width (cm)']
Classes:  ['setosa' 'versicolor' 'virginica']
```

**Figura 1. Classes e atributos da base Iris.**

Para a manipulação dos dados, foi utilizada a biblioteca Pandas[3], que nos permite transformar um conjunto de dados em um *dataframe* - que é uma estrutura similar a uma tabela - e realizar algumas análises a partir dele, como mostra a Figura 2.

```
1 dataset = pd.DataFrame(data= np.c_[data['data'], data['target']], columns= data['feature_names'] + ['class'])
2 print(dataset)
```

```
     sepal length (cm)  sepal width (cm)  ...  petal width (cm)  class
0                  5.1               3.5  ...               0.2    0.0
1                  4.9               3.0  ...               0.2    0.0
2                  4.7               3.2  ...               0.2    0.0
3                  4.6               3.1  ...               0.2    0.0
4                  5.0               3.6  ...               0.2    0.0
..                 ...               ...  ...               ...    ...
145                6.7               3.0  ...               2.3    2.0
146                6.3               2.5  ...               1.9    2.0
147                6.5               3.0  ...               2.0    2.0
148                6.2               3.4  ...               2.3    2.0
149                5.9               3.0  ...               1.8    2.0

[150 rows x 5 columns]
```

**Figura 2. Dataframe da base Iris.**

[1] http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data
[2] https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_iris.html
[3] https://pandas.pydata.org/

Na criação do *dataframe* foi utilizada a biblioteca NumPy[4] para adicionar a coluna "class", que contém a classe de cada amostra.

## 2. Exploração dos dados

O primeiro passo da exploração de dados é através da estatística descritiva obter informações que resumam quantitativamente as características da base de dados a ser trabalhada a fim de auxiliar na tomada de decisões em etapas posteriores, como o pré-processamento.

A Tabela 1 apresenta algumas medições sobre a base de dados iris, contendo os valores de mínimo e máximo, a média, a mediana, o desvio padrão, o 1º quartil (Q1), o 3º quartil (Q3), a obliquidade e a curtose. Estas medições foram realizadas por meio da biblioteca Pandas.

**Tabela 1. Estatísticas da base Iris.**

| | Tamanho da sépala (cm) | Largura da sépala (cm) | Tamanho da pétala (cm) | Largura da pétala (cm) |
|---|---|---|---|---|
| Máximo | 7,90 | 4,40 | 6,90 | 2,50 |
| Mínimo | 4,30 | 2,00 | 1,00 | 0,10 |
| Média | 5,84 | 3,06 | 3,76 | 1,20 |
| Mediana | 5,80 | 3,00 | 4,35 | 1,30 |
| Desvio padrão | 0,83 | 0,44 | 1,77 | 0,76 |
| Q1 | 5,10 | 2,80 | 1,60 | 0,30 |
| Q3 | 6,40 | 3,30 | 5,10 | 1,80 |
| Obliquidade | 0,31 | 0,32 | -0,27 | -0,10 |
| Curtose | -0,55 | 0,23 | -1,40 | -1,34 |

Os valores de média e mediana estão bem parecidos, o que indica que os dados estão bem distribuídos, o que pode indicar baixa ocorrência de *outliers*, que são resultados que têm baixa probabilidade de ocorrer. A obliquidade positiva dos atributos da sépala indicam cauda direita, enquanto que a obliquidade negativa indica cauda esquerda, no entanto são valores próximos de zero, o que indicaria simetria dos dados. A curtose negativa dos atributos indica uma distribuição achatada, com exceção da curtose do atributo *largura da sépala (cm)*, que é positiva, indicando uma distribuição afunilada dos dados deste atributo.

## 3. Boxplots

Os boxplots são uma forma de visualização dos quartis dos dados de cada atributo, sendo importantes na identificação de *outliers*. A Figura 3 mostra um gráfico contendo o boxplot de cada atributo da base de dados, onde fica nítido que no atributo *sepal width (cm)* temos a presença de *outliers*, representados pelos círculos acima do limite superior e abaixo do limite inferior. A Figura 4 mostra os boxplots de cada atributo a partir das classes presentes nas amostras. Ambas as figuras tiveram seus gráficos gerados a partir da biblioteca Seaborn[5], voltada para visualizações de dados em Python.

[4] https://numpy.org/
[5] https://seaborn.pydata.org/

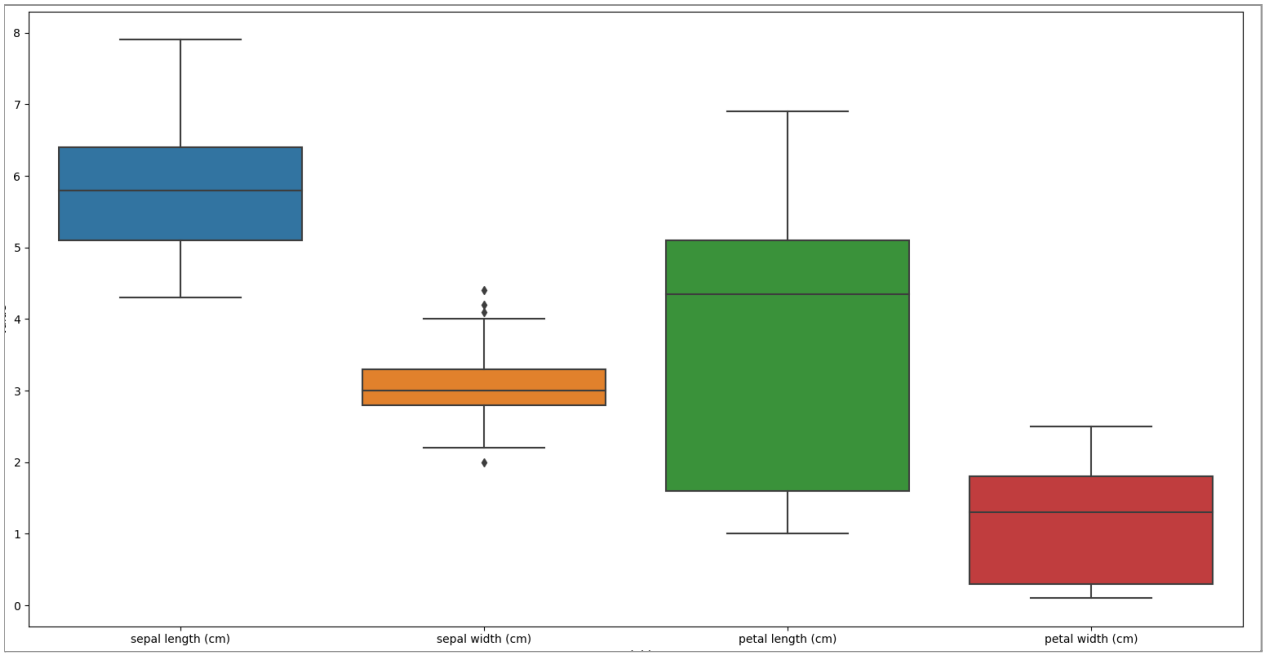


**Figura 3. Boxplots da base Iris.**

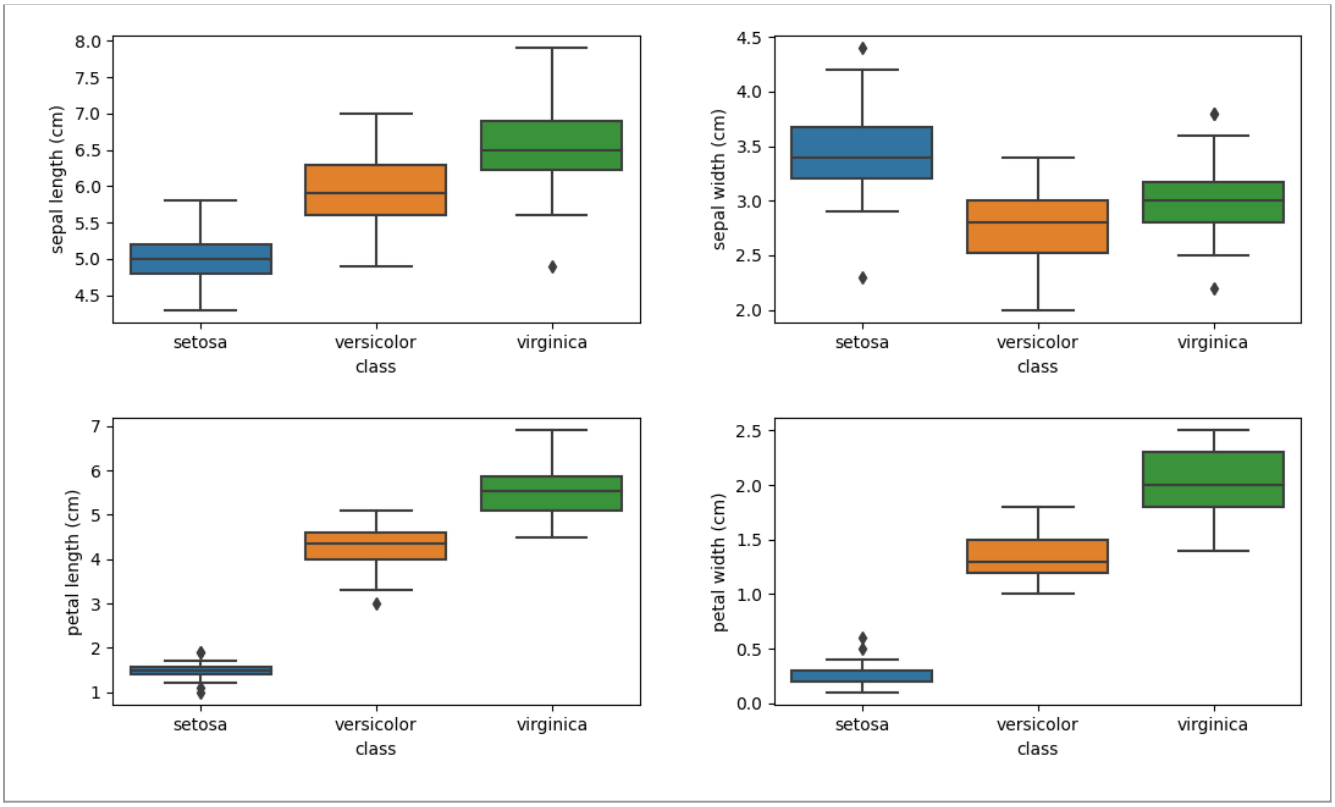


**Figura 4. Boxplots de cada atributo distribuídos por classe.**

## 4. Histogramas

Os histogramas servem para a visualização da distribuição dos dados, onde os valores do assumidos pelo atributo são divididos em cestas (ou intervalos), nas quais é possível através da altura das barras perceber a frequência dos dados. Na Figura 5 são apresentados os

histogramas para cada atributo, sendo gerados através da biblioteca Seaborn utilizando `n_bins = 10`. Através da análise destes histogramas percebe-se que o atributo *sepal length (cm)* é mais achatado do que os demais, pois seus dados têm frequências parecidas, diferentemente do atributo *sepal width (cm)*, que tem os seus valores concentrados nos intervalos do meio. Já os atributos referentes à pétala (*petal length (cm)* e *petal width (cm)*) têm um comportamento parecido porque forma 2 grupos de ocorrências, havendo no atributo *petal length (cm)* inclusive um intervalo sem nenhum dado entre dois blocos com dados existentes, denotando menor variabilidade de valores possíveis para estes atributos.

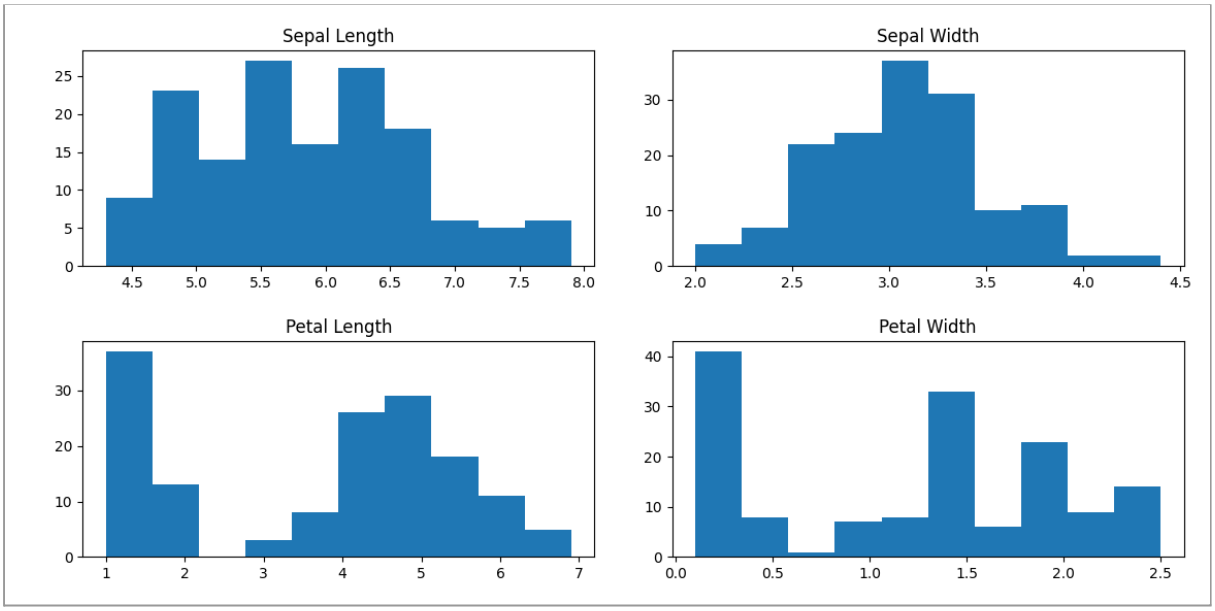


**Figura 5. Histogramas de cada atributo.**

## 5. Scatterplots

Os scatterplots são diagramas que nos ajudam a identificar a relação entre diferentes atributos, ilustrando a correlação linear entre dois atributos. A partir dessas relações de dois em dois atributos, é possível plotar uma matriz de scatterplot, como mostra a Figura 6, também provida por meio da biblioteca Seaborn.

Através da matriz de scatterplot, é possível verificar que as amostras da classe *setosa* (representada pela cor azul) formam agrupamentos mais bem definidos do que as amostras das outras classes

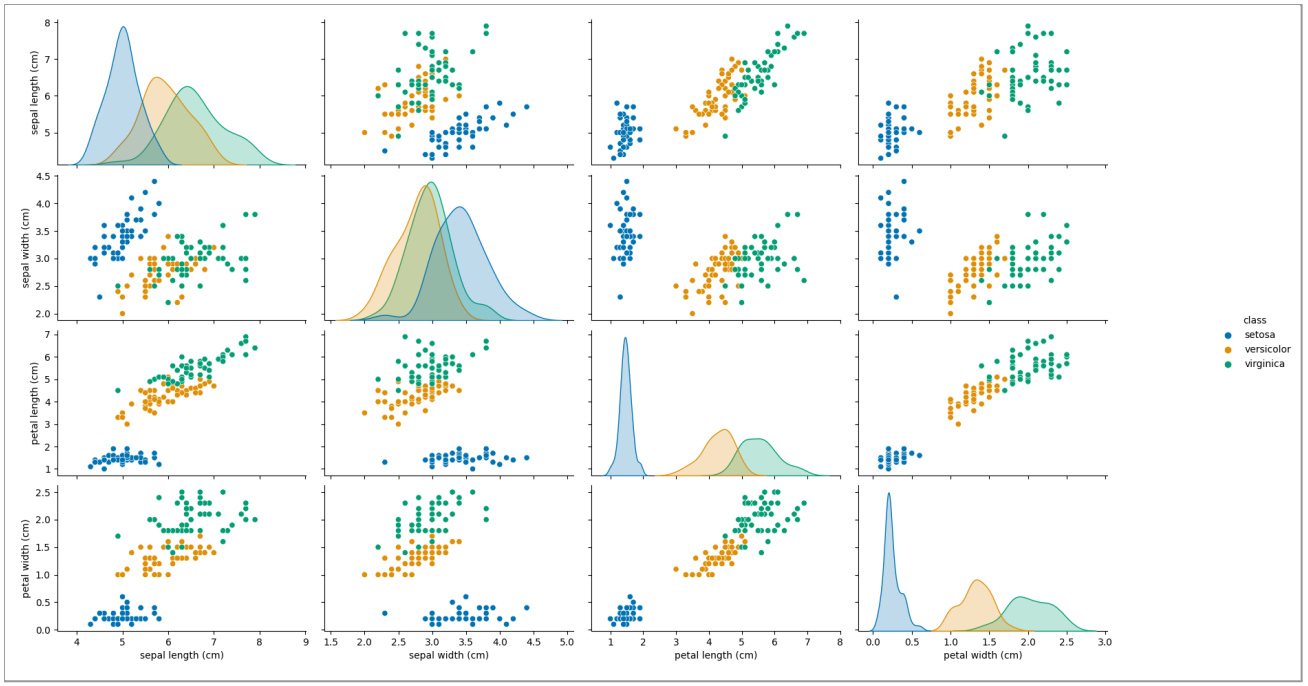


**Figura 6. Matriz de scatterplot da base Iris.**