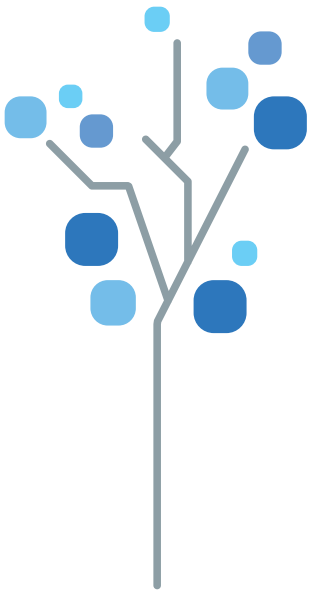
Fundamentos de Inteligência Artificial [5COP099]



DC
Departamento de Computação

Dr. Sylvio Barbon Junior

Departamento de Computação - UEL

1º Semestre

Aula 8
Análise de Dados

Sumário

Introdução

- A análise da características presentes em um conjunto de dados permite a descoberta de padrões e tendências que podem fornecer informações sobre os processos que os gerou.
- Muitas informações podem ser obtidas por estatística simples.
- Outras por meio de técnicas de visualização.
- No entanto, muitos padrões e conhecimento são obtidos com uso de técnicas mais sofisticadas como o Aprendizado de Máquina.

Caracterização dos Dados

- Um conjunto de dados são representados por **objetos**;
- Objetos podem representar objetos físicos ou noções abstratas;
- Os objetos são representados por **descritores**, também denominados como atributos, campos ou variáveis;
- Cada propriedade está associada a uma propriedade do objeto;
- Formalmente os dados podem ser representados como uma matriz $X_{n \times d}$, onde **n** é o número de objetos e **d** o número de atributos.
- O valor de **d** define a dimensionalidade do problema.

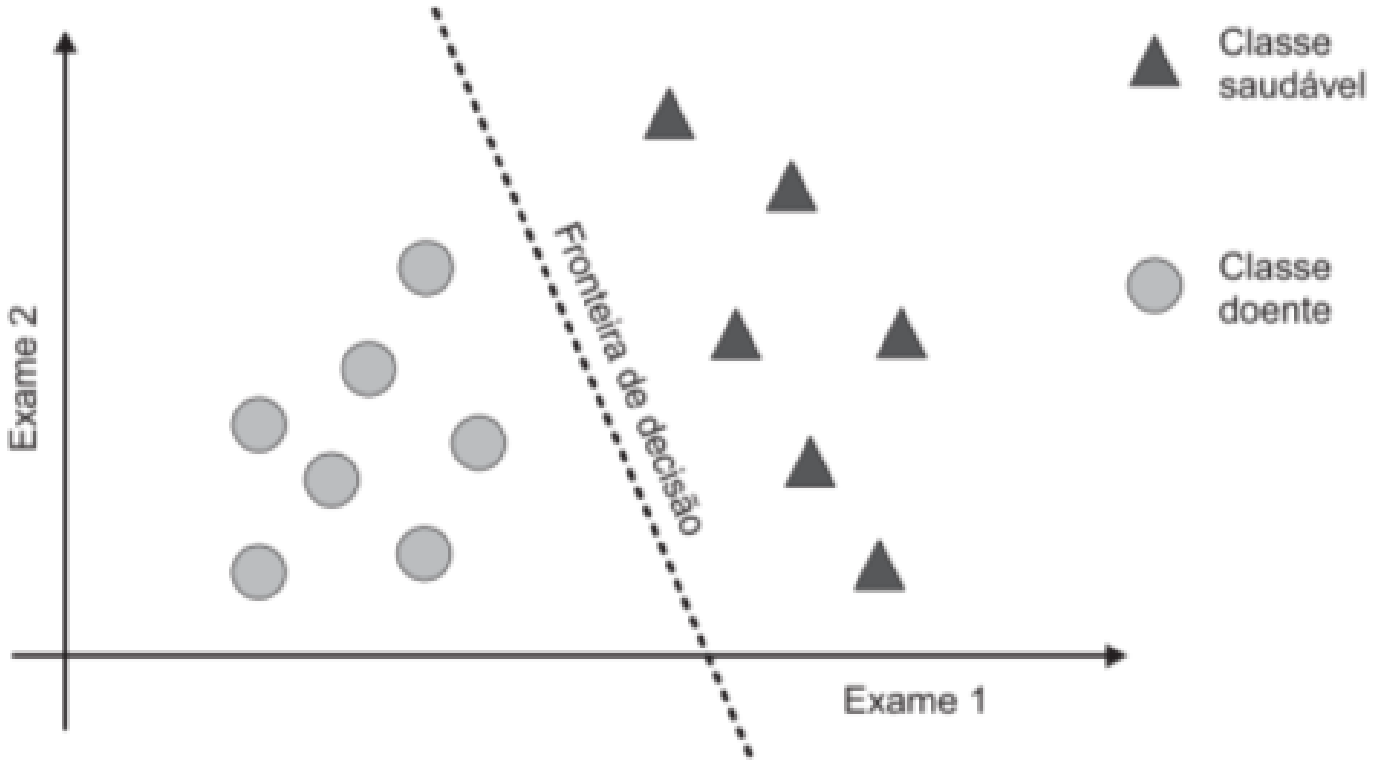
Caracterização dos Dados

Considere o conjunto de dados hospital:

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	8	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	72	Inexistentes	38,0	3	RJ	Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável

Caracterização dos Dados

Objetivo do Aprendizado de Máquina:



Caracterização dos Dados

- Atributo alvo: Chamado de atributo meta ou saída, representa o fenômeno de interesse sobre o qual se deseja fazer as previsões.
- Alguns problemas não apresentam um atributo alvo, neste caso a linha de atuação é chamada classificação **não supervisionada**. Quando existem vários atributos alvo, chamamos o problema de **multi-rótulo**.
- Quando temos a presentaça de um atributo alvo, podemos ter problemas de natureza:
  - **Classificação**, quando o atributo alvo apresenta um rótulo ou classe com valores discretos. Exemplo: 1, 2, ..., k ou Saudável, Doente. Estes problemas podem apresentar classes com maior (**majoritária**) ou menor (**minoritária**) na descrição do problema.
  - **Regressão**, quando o atributo alvo são valores numéricos contínuos. Exemplo: Prever a temperatura ou tempo ideal para colheita.

Caracterização dos Dados

- Os valores podem assumir diferentes formas. Podemos avaliá-los quanto ao:
  - **Tipo**: diz respeito ao grau de quantização dos dados;
  - **Escala**: está relacionada a significância relativa dos dados.
- Com relação ao tipo, podemos assumir os seguintes:
  - **Qualitativo** (simbólico): quando o atributo está associado a uma categoria. Exemplo, pequeno, médio, grande.
  - **Quantitativos** (numéricos): quando o atributo é numérico e pertence a um conjunto de valores, podendo ser **contínuos** ou **discretos**. Estes valores podem ser inteiros, reais ou binários.

Caracterização dos Dados

- Com relação à **escala**, podemos assumir os seguintes:
  - Nominal: são apenas nomes diferentes, sem carregar qualquer informação associada, não existindo relação de ordem. Exemplo: CPF
  - Ordinal: apresentam valores que podem ser manipulados pelos operadores de grandeza (exemplo $>$ e $<$). Exemplo: frio, morno e quente.
  - Intervalar: quando os números variam dentro de uma escala e a ordem representa a magnitude entre os valores. Exemplo: temperatura;
  - Racional: são os atributos que agregam mais valores, com um significado absoluto. Exemplo, "visitas ao hospital", onde 0 significa ausência do atributo.

Exploração dos Dados

- As técnicas de exploração dos dados podem auxiliar na seleção de técnicas posteriores como pré-processamento e aprendizado.
- O mais tradicional é utilizar a **estatística descritiva**, que permitem capturar informações como:
  - **Frequência**;
  - **Localização** e tendência central;
  - **Dispersão** ou espalhamento;
  - **Distribuição** e formato;
- A **medida de frequência** mede a proporção de vezes que um atributo assume um dado valor em determinado conjunto de dados. Pode ser aplicada a qualquer natureza de tipo de dado.

Exploração dos Dados

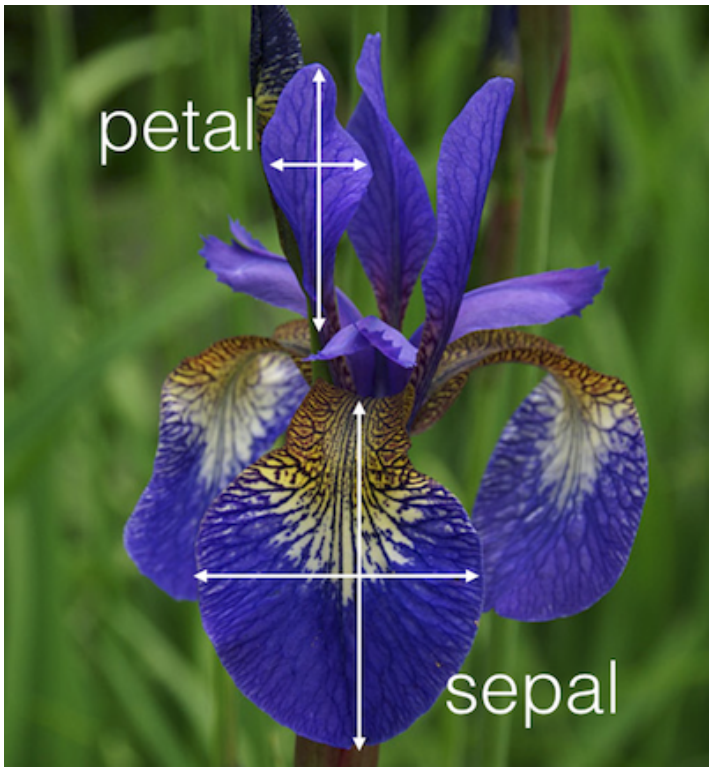
- As técnicas de exploração dos dados podem auxiliar na seleção de técnicas posteriores como pré-processamento e aprendizado.
- O mais tradicional é utilizar a **estatística descritiva**, que permitem capturar informações como:
  - **Frequência**;
  - **Localização** e tendência central;
  - **Dispersão** ou espalhamento;
  - **Distribuição** e formato;
- A **medida de frequência** mede a proporção de vezes que um atributo assume um dado valor em determinado conjunto de dados. Pode ser aplicada a qualquer natureza de tipo de dado.

Exploração dos Dados

- **Dados Univariados**: São objetos que apresentam somente um atributo.
  - **Medidas de Localidade**: Pode-se utilizar a moda (para dados simbólicos). Para atributos numéricos, pode-se utilizar média, mediana e percentil. A média é um atributo importante quando os dados não apresentam *outliers* ou distribuições não simétricas. Quando há a presença de *outliers* o ideal seria utilizar a mediana.
  - O uso do **Boxplot** é uma ferramenta importante para a identificação de medidas de localidade.

Caracterização dos Dados

Problema de classificação de "iris" (flor):



Caracterização dos Dados

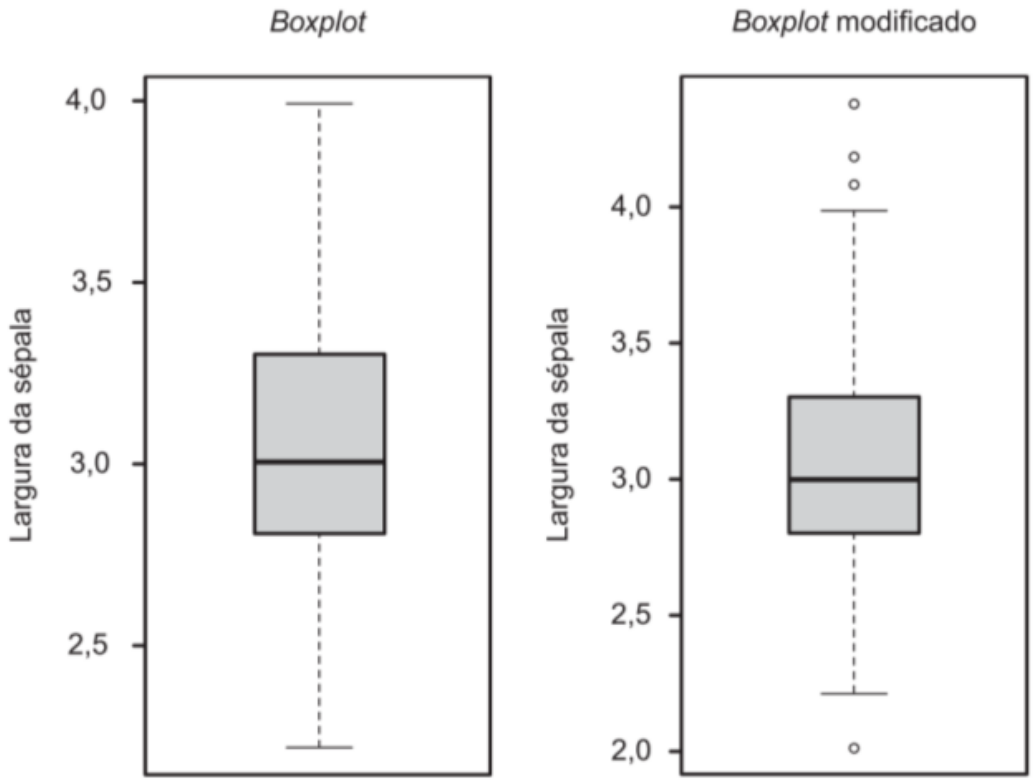
Base de dados iris

sepal length in cm	sepal width in cm	pepal length in cm	pepal width in cm	label
4.600	3.200	1.400	0.200	Iris-setosa
5.300	3.700	1.500	0.200	Iris-setosa
5	3.300	1.400	0.200	Iris-setosa
7	3.200	4.700	1.400	Iris-versicolc
6.400	3.200	4.500	1.500	Iris-versicolc

- 150 objetos: 50 de cada classe
- 4 atributos: comprimento pétala, largura da pétala, comprimento da sépala, largura da sépala
- 3 classes: setosa, versicolor, virginica

Caracterização dos Dados

Boxplot e Boxplot modificado para caracterização dos dados.

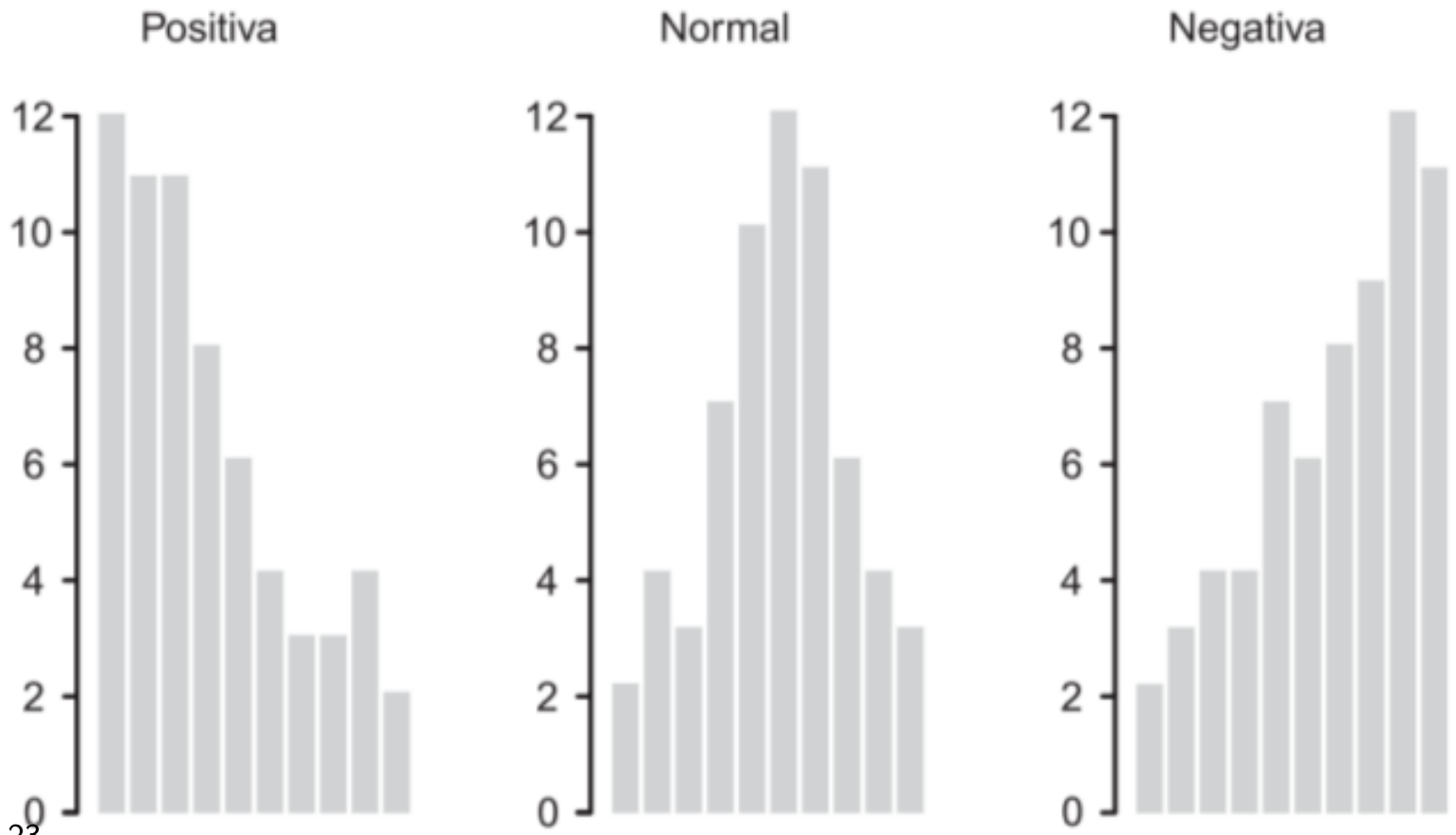


Exploração dos Dados

- **Dados Univariados**: São objetos que apresentam somente um atributo.
  - **Medidas de Espalhamento**:Permitem observar se os valores estão amplamente espalhados ou concentrados. Normalmente utiliza-se: Intervalo, Variância e Desvio Padrão (AAD, MAD e IQR - *Absolute Average Deviation*, *Median Absolute Deviation* e *Interquartil Range*).
  - **Medidas de Distribuição**:Baseia-se no momento, sendo divido em:
    - Valor 0 (k=1): quando é o primeiro momento em torno da origem (média);
    - Variância (k=2): segundo momento central;
    - **Obliquidade** (k=3) (*skewness*): terceiro momento central, mede a simetria. Obliquidade $= 0$ é simétrica, Obliquidade > 0 concentração do lado esquerdo e Obliquidade < 0 concentração do lado direito.
    - **Curtose** (k=4) (*kustosis*): quarto momento central, é a medida de dispersão (achatamento da distribuição). Curtose $= 0$ (normal), Curtose > 0 (concentrado) e Curtose < 0 (mais achatado que a normal)

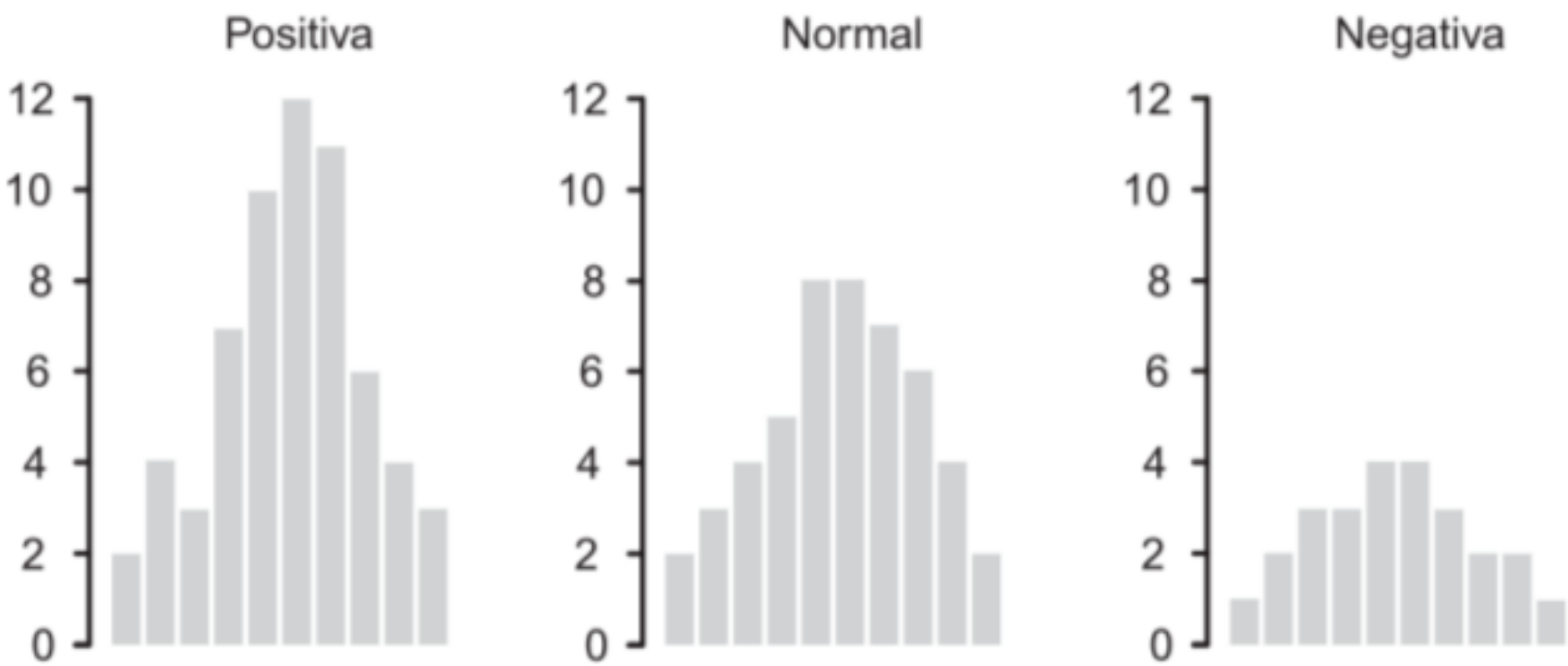
Exploração dos Dados

Gráficos comparando Skewness:



Exploração dos Dados

Gráficos comparando Kustosis:

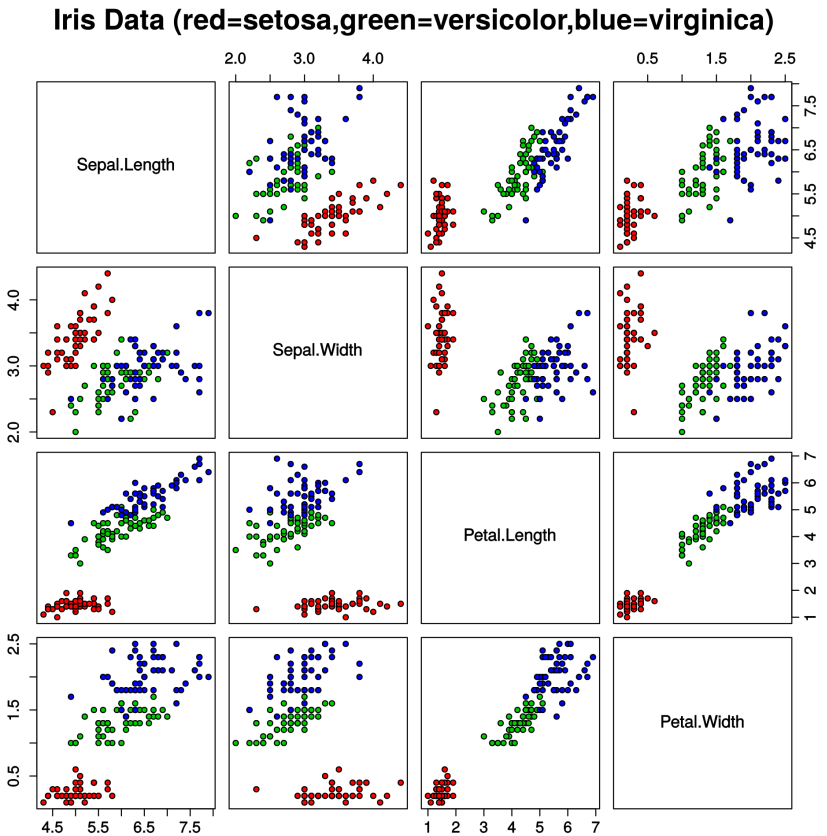


Exploração dos Dados

- **Dados Multivariados**: São aqueles que mais de um atributo descreve um objeto.
- Neste tipo de caso procura-se realizar a análise univariada para cada item e compreender qual atributo teria um comportamento desejado.
- Uma importante análise é a relação entre os atributos, para isso usamos:
  - **Covariância**: mede o grau com que cada atributo varia junto. Seu valor depende da magnitude dos atributos. A covariância negativa indicada que quando um atributo aumenta, outro diminui.
  - **correlação**: indica a relação linear entre dois atributos independente da dimensão.

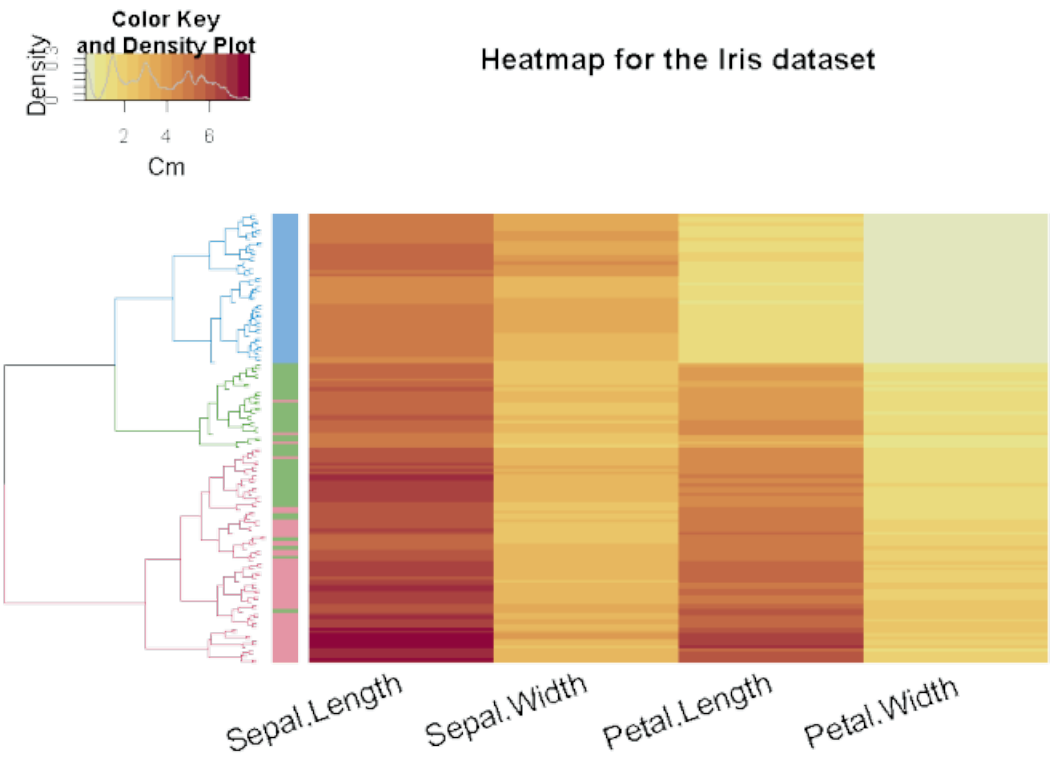
Exploração dos Dados

O mais utilizado para análise multivariada é o **Scatterplot**



Exploração dos Dados

Heatmap permite analises com mais dimensões.



Referências

1. Coppin, B. Inteligência Artificial. LTC. 2010.
2. Russell, S.; Norvig, P. Artificial Intelligence: a modern approach. Prentice Hall. 2010. Localização: BC – Número de Chamada: 519.683 R967a 3.ed.
3. Luger, G. F. Inteligência Artificial: estruturas e estratégias para a resolução de problemas complexos.Bookman. 2004. Localização: BC – Número de Chamada: 519.683 L951a 4.ed.
4. Carvalho, André, et al. "Inteligência Artificial–uma abordagem de aprendizado de máquina."Rio de Janeiro: LTC (2011).