Business Analytics

Marcelo Rosano Dallagassa

2024

Visualização de dados

Análise Exploratória de Dados (EDA)

Os cinco princípios de Edward Tufte

1º Princípio: Comparação

Mostrar comparações, contrastes e diferenças: compara-se a quê?

Análise Exploratória de Dados (EDA)

Os cinco princípios de Edward Tufte

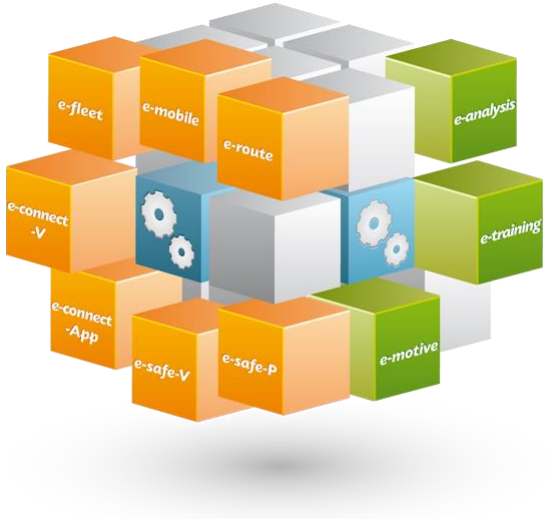
2º Princípio: Causalidade.

As relações de causa e o efeito – mostrar causalidade, mecanismo, explicação, estrutura sistemática.

Análise Exploratória de Dados (EDA)

Os cinco princípios de Edward Tufte

3º Princípio: Multidimensão - Mostrar dados multivariados (OLAP)



“To think multivariate, show multivariate” Edward Tufte.

Análise Exploratória de Dados (EDA)

Os cinco princípios de Edward Tufte

4º Princípio: Evidências (enriquecimento dos dados)

Integrar completamente palavras, números, diagramas, imagens

“Words, numbers, diagrams, graphics, charts, tables, belong together”
Edward Tufte

Análise Exploratória de Dados (EDA)

Os cinco princípios de Edward Tufte

5º Princípio: Documentação

Descreva a evidência – forneça títulos detalhados, indicar fontes com autores e apoiadores , mostrar as escalas com medições completas, aponte os aspectos relevantes.

Visualização de Dados

É a **conversão** dos dados para um formato visual ou tabular de tal forma que as características dos dados e os relacionamentos entre itens de dados ou atributos possa ser analisada ou reportada.

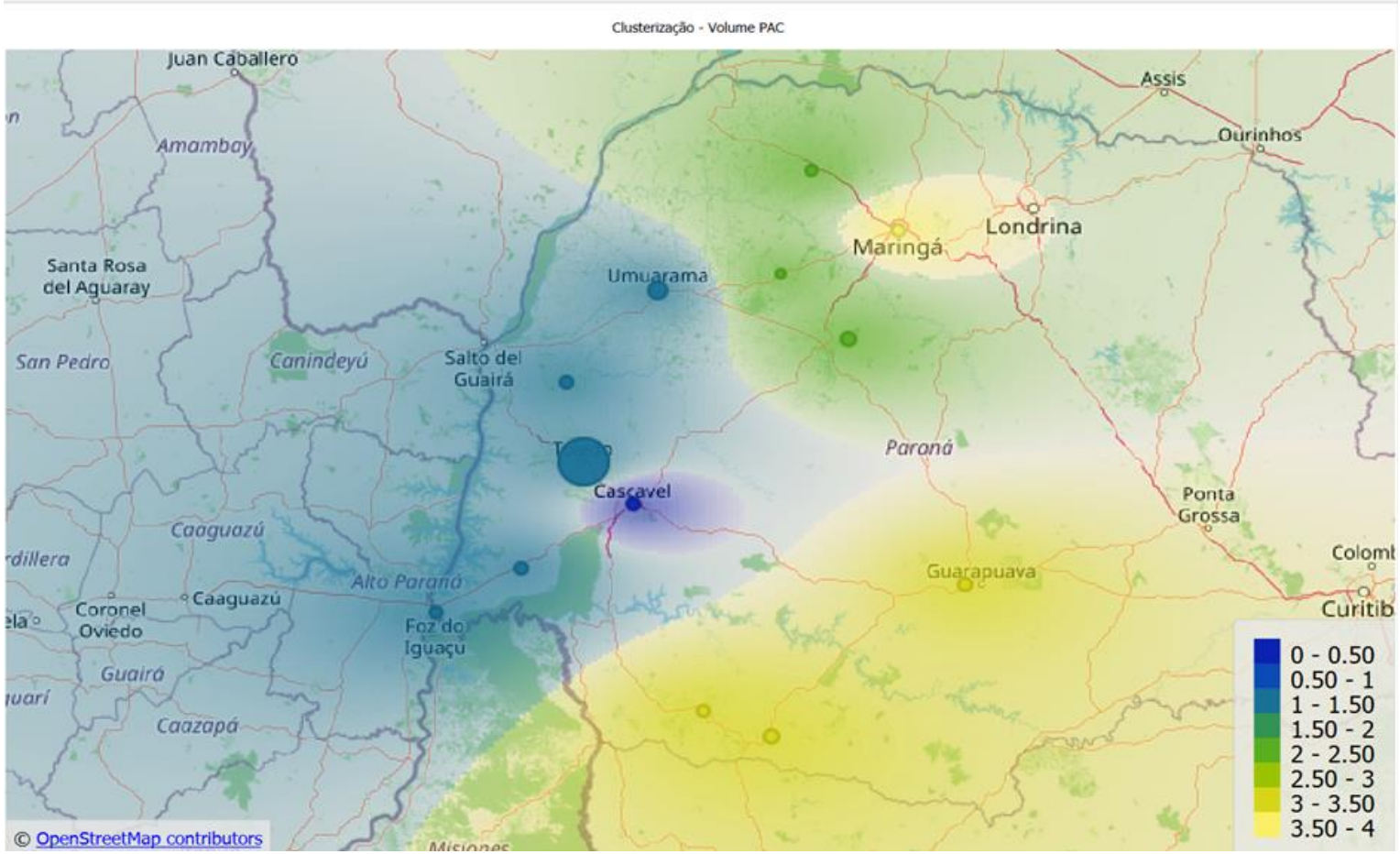
Visualização de dados é uma das técnicas de maior apelo e poder para exploração de dados.

- ✓ Os seres humanos tem uma grande habilidade de analisar grandes quantidades de informação que seja apresentada visualmente
- ✓ Pode detectar padrões gerais e tendências
- ✓ Pode detectar outliers e padrões não usuais

Visualização de Dados

Exemplo de visualização de dados:

Mapa do Estado do Paraná – Cluster de Interesse Estratégicos



Visualização de Dados

Visualização / Representação de dados:

- ✓ É o mapeamento da informação em um formato visual
- ✓ Objetos de dados, seus atributos e as relações entre objeto de dados são traduzidos em elementos gráficos tais como, pontos, linhas, formatos e cores.

Exemplos:

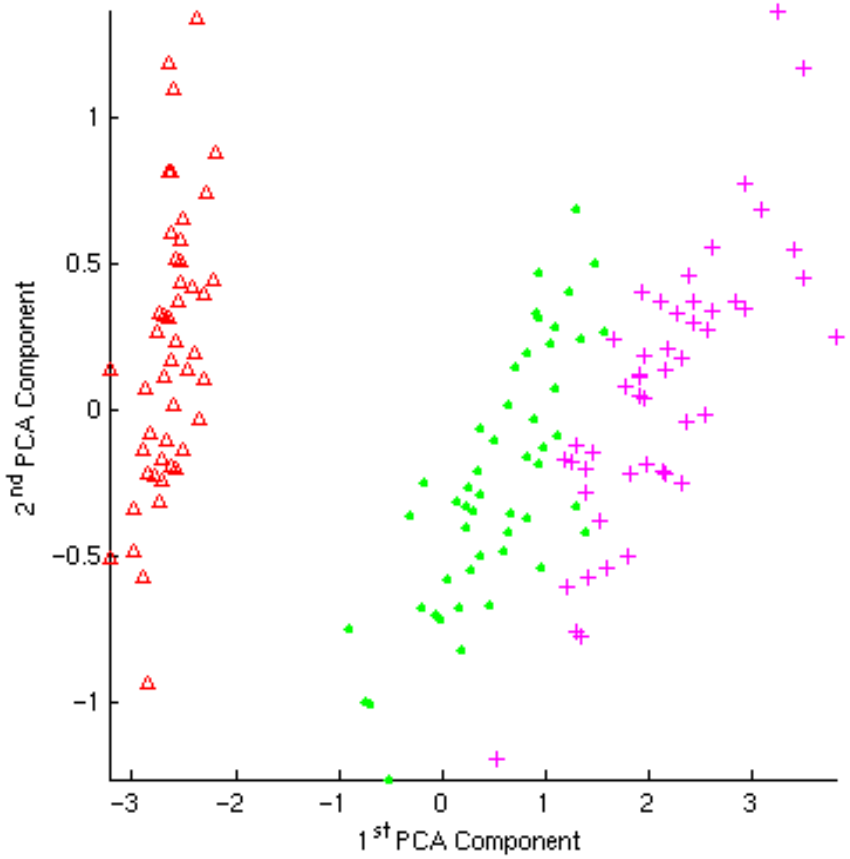
✓Objetos são frequentemente representados por pontos.

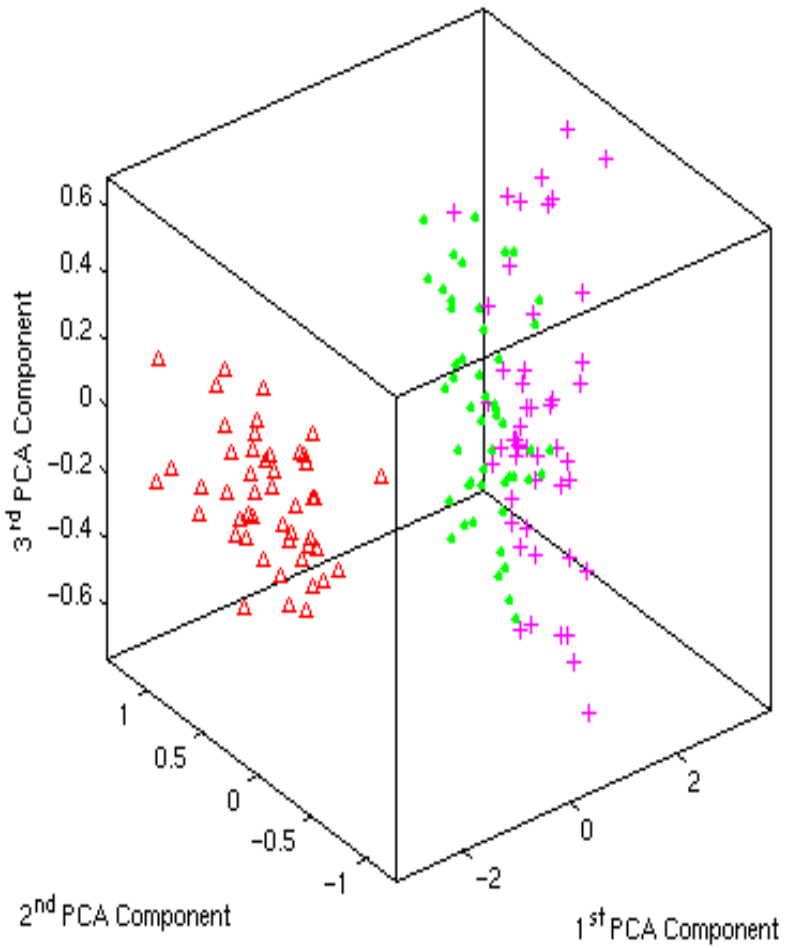
✓Seus valores de atributo podem ser representadas como pontos ou as características dos pontos. Exemplo: cor, tamanho e formato.

✓Se a distância entre eles for utilizada, um agrupamento de pontos forma um conjunto (grupos) e se for um ponto isolado então é um outlier.

Visualização de Dados

Scatterplot 2D, 3D





Visualização de Dados

Seleção:

- ✓É a retirada (eliminação) de certos atributos / objetos.
  - ✓Subconjunto de atributos
    - Redução da dimensionalidade (de 3 para 2)
  - ✓Subconjunto de objetos
    - Seleção de uma região (áreas esparsas)

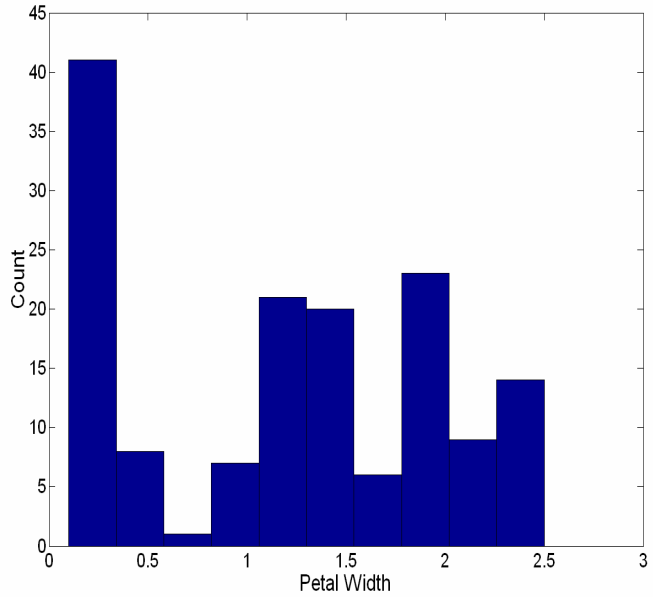
Visualização de Dados

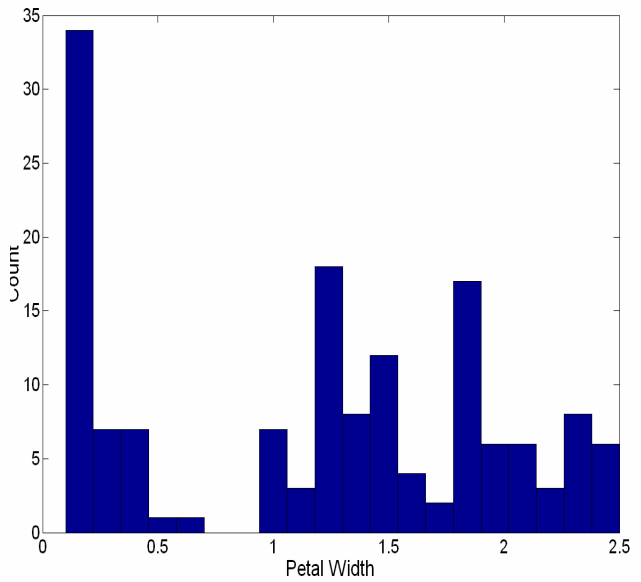
Histogramas:

Mostra distribuição de valores de uma variável e a divisão dos valores em faixas apresentando um gráfico de barras do número de objetos em cada faixa.

- A altura de cada barra indica o número de objetos
- Formato do histograma depende do número de faixas

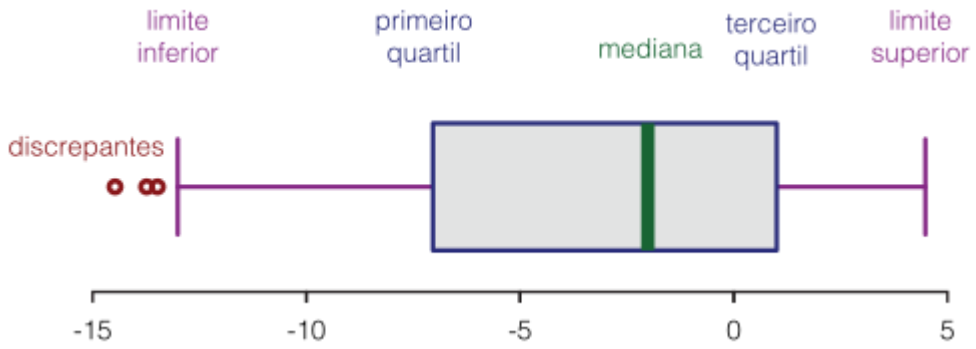
Exemplo: Comprimento da Pétala (10 e 20 faixas)

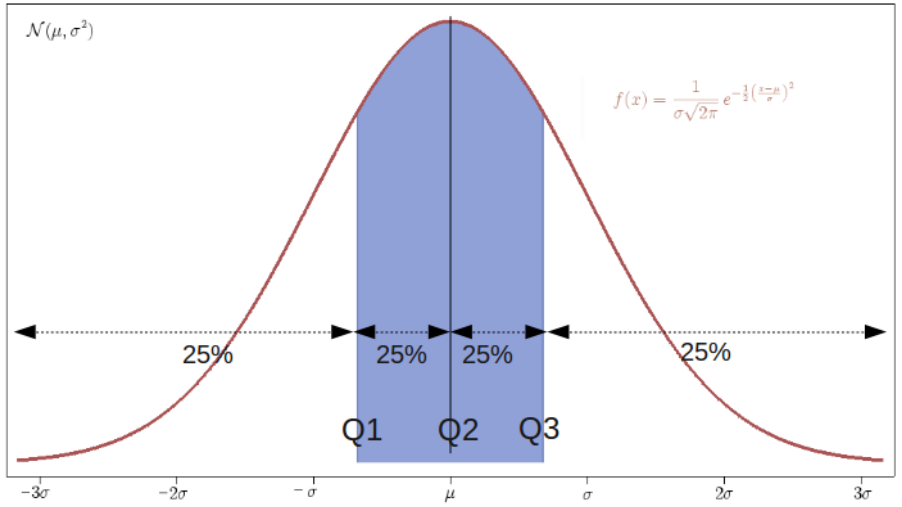




Visualização de Dados

Gráfico Box-Plot, idealizado por J.Tukey

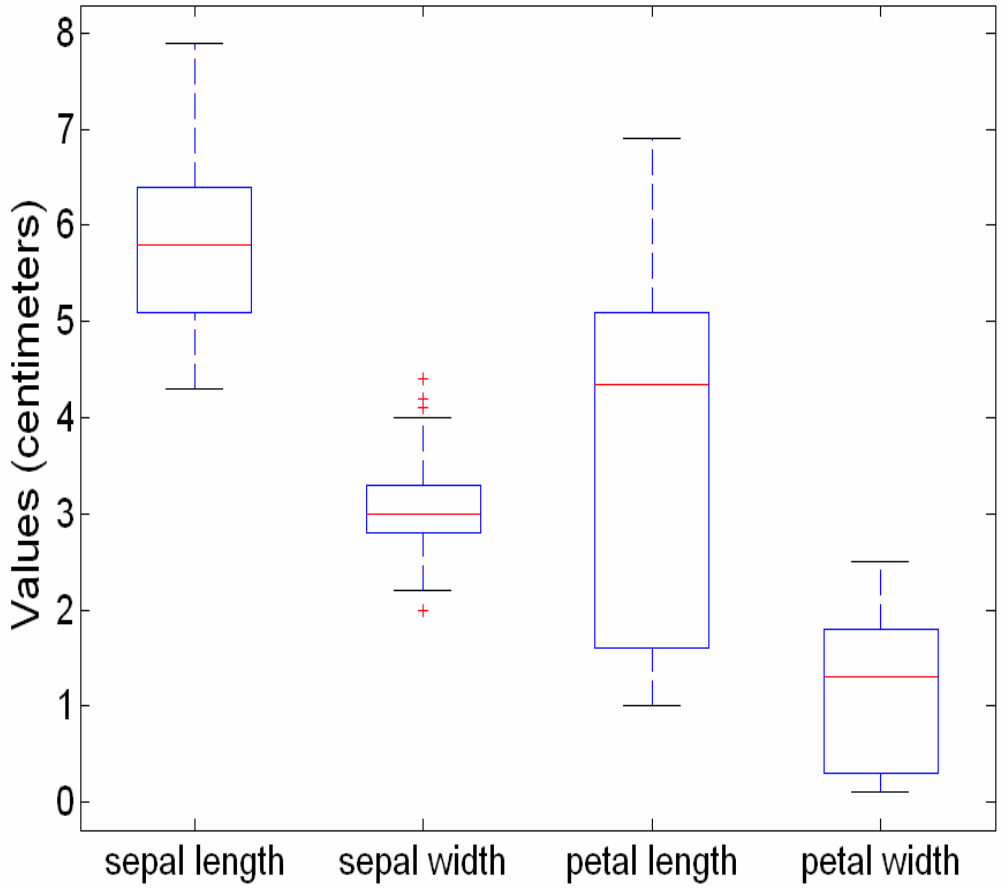




Visualização de Dados

Gráfico Box-Plot

Análises comparativas



Visualização de Dados

Gráfico de Dispersão

- Valores dos atributos determinam a posição.

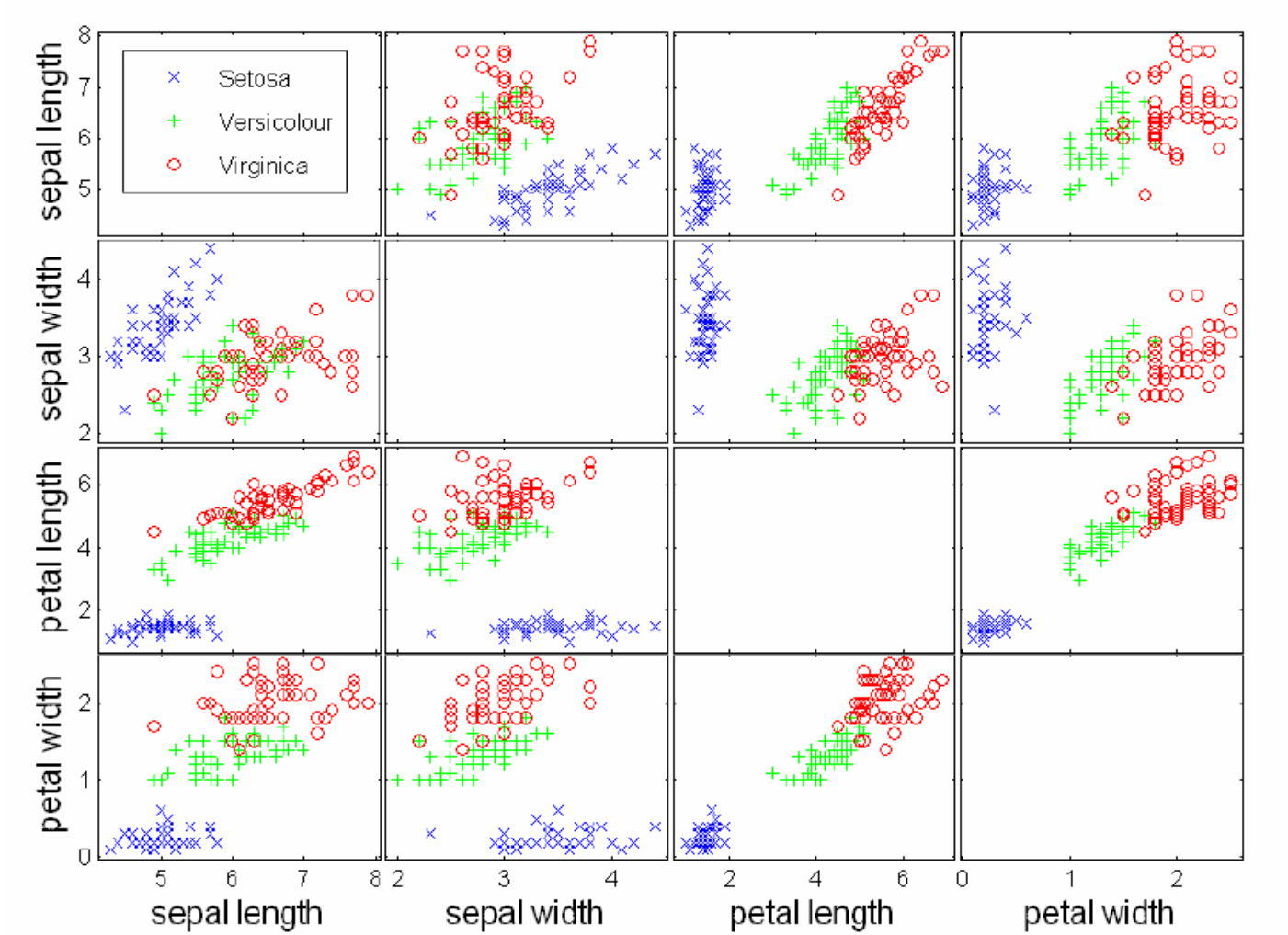
- Gráficos de dispersão bidimensionais são mais comuns, mas também há gráficos tridimensionais

- Os atributos adicionais podem ser mostrados usando tamanho, forma e cor dos marcadores que representam os objetos

- É útil ter arranjos de gráficos de dispersão para sumarizar de maneira compacta os relacionamentos de vários pares de atributos

Visualização de Dados

Gráfico de Dispersão



Visualização de Dados

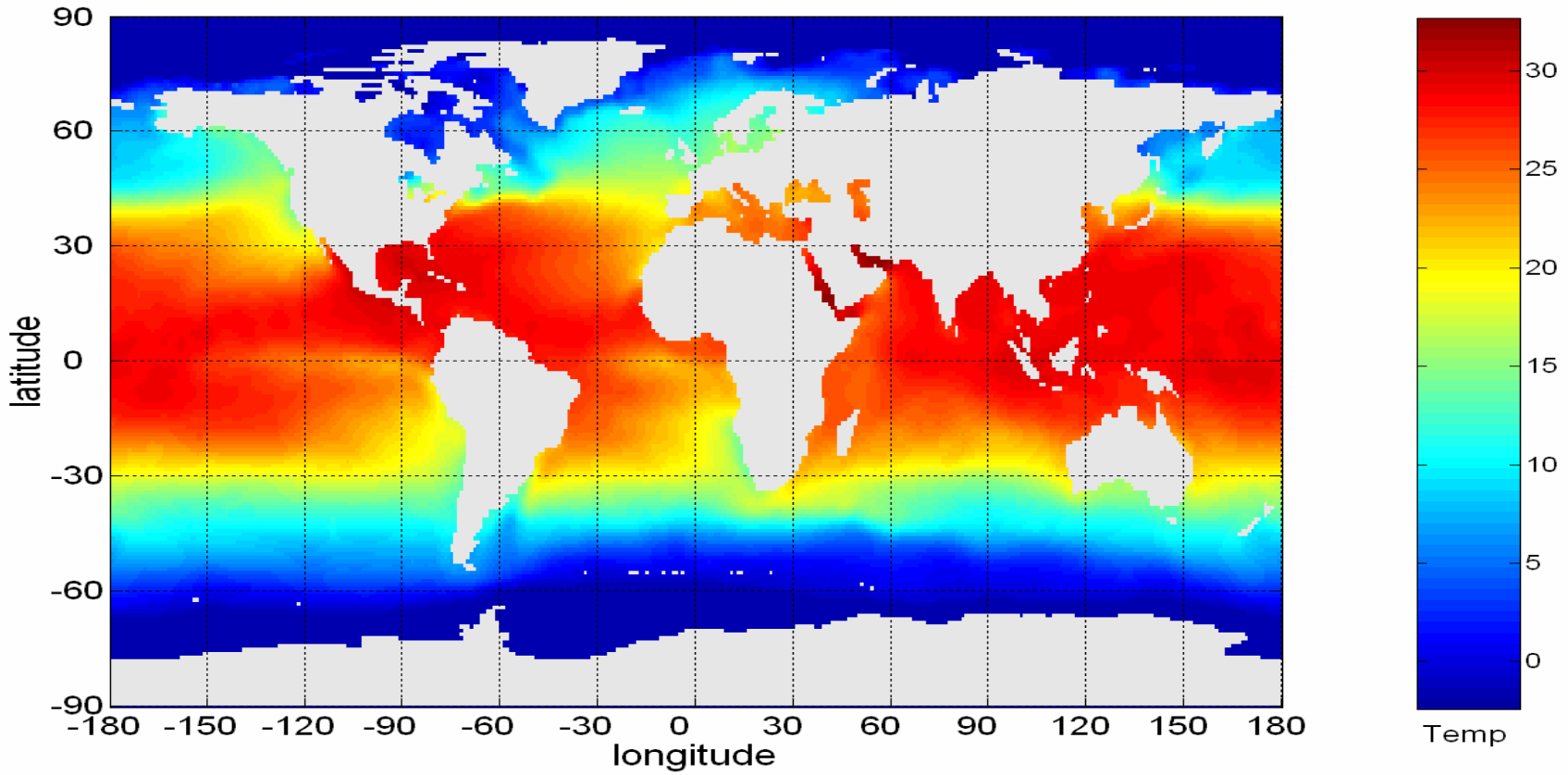
Gráfico de Contorno

✓ Útil quando um atributo contínuo é medido em uma grade espacial
✓Particionam o plano em regiões de valores similares
✓Linhas de contorno que formam os limites destas regiões conectam pontos com valores iguais
✓O exemplo mais comum são os mapas de contorno de elevação
✓Podem indicar temperatura, precipitação, pressão do ar, etc

Visualização de Dados

Gráfico de Contorno

Exemplo: temperatura dos oceanos



Visualização de Dados

Gráfico Matriciais

✓Plotam a matriz de dados.

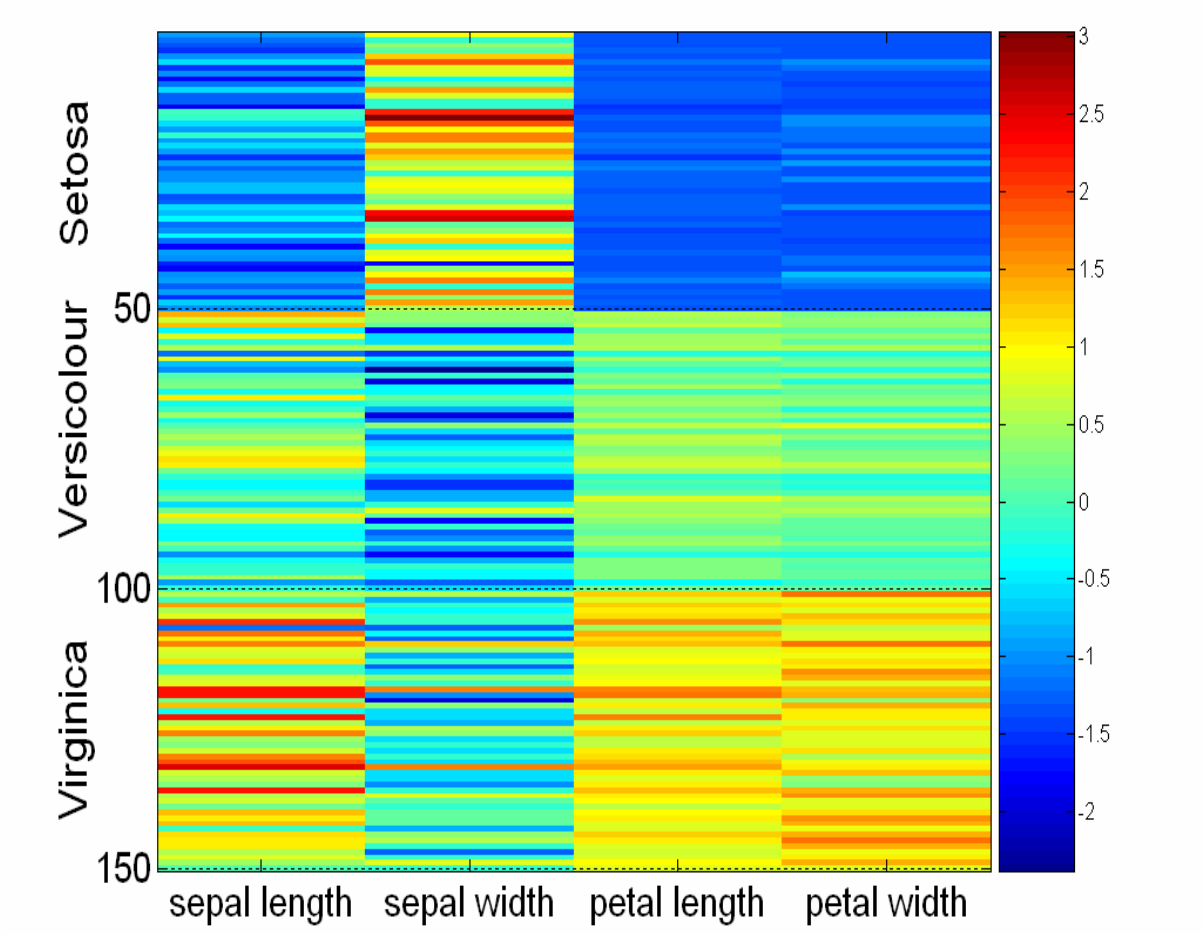
✓Podem ser útil quando os objetos são ordenados de acordo com a classe.

✓Os atributos normalmente são normalizados para evitar que um atributo domine o gráfico.

✓Gráficos de similaridade ou matrizes de distância também podem ser úteis para visualizar os relacionamentos entre objetos

Visualização de Dados

Gráfico Matriciais (Exemplo)



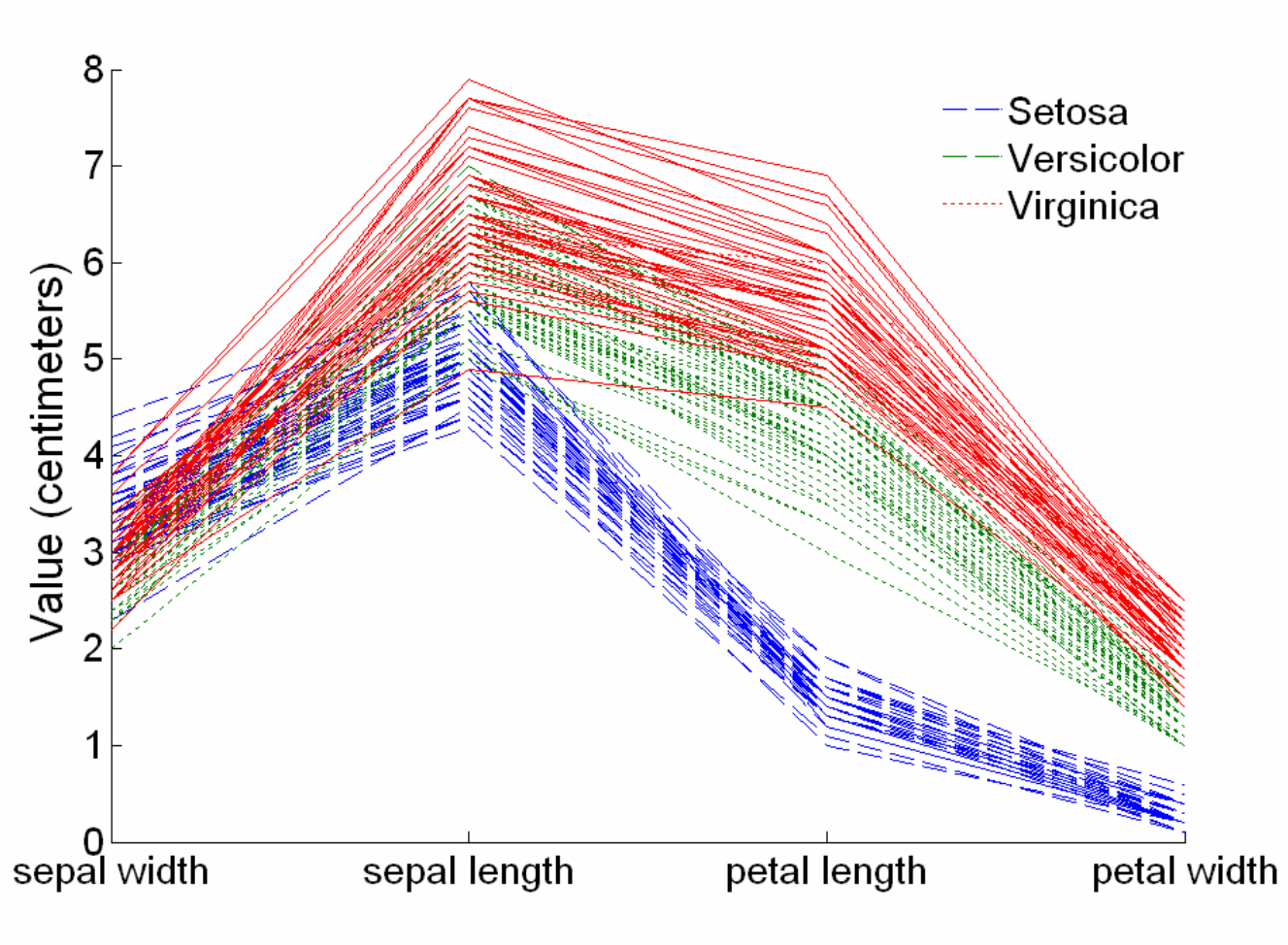
Visualização de Dados

Coordenadas Paralelas

- ✓Usadas para plotar os valores dos atributos de dados de alta dimensionalidade;
- ✓Em lugar de eixos perpendiculares, usa-se um conjunto de eixos paralelos;
- ✓Valores dos atributos de cada objeto são plotados como um ponto em cada um dos eixos coordenados correspondentes e os pontos são ligados por linhas;
- ✓Então, cada objeto é representado como uma linha;
- ✓Frequentemente linhas representam uma classe distinta de objetos agrupados, ao menos para alguns atributos
- ✓Ordenar atributos é importante para ver os grupos

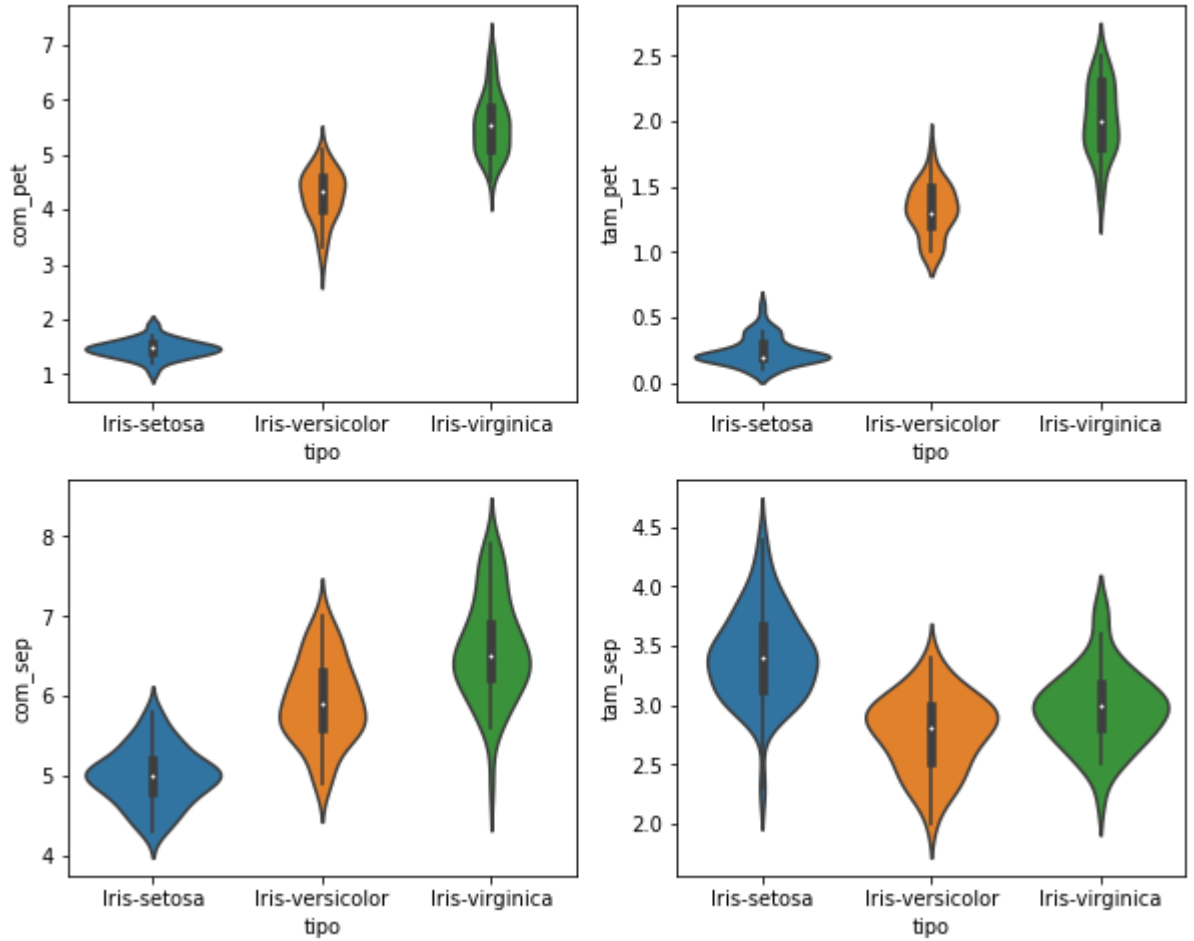
Visualização de Dados

Coordenadas Paralelas (Exemplo)



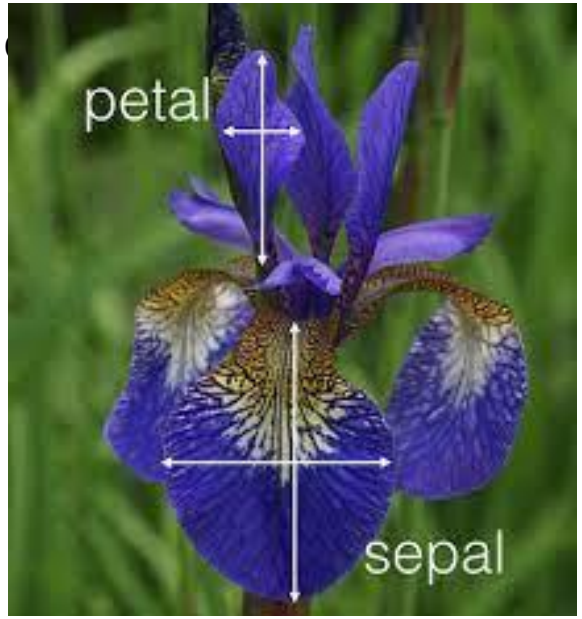
Visualização de Dados

Violin Plot



Visualização de Dados

Muitas das técnicas de exploração de dados são ilustradas com o conjunto de dados da planta Iris.



- Pode ser obtido do UCI Machine Learning Repository
  http://www.ics.uci.edu/~mlearn/MLRepository.html

- Criada pelo estatístico Douglas Fisher
- Três tipos de flores (classes):
  Setosa
  Virginica
  Versicolour
- Quatro atributos (não-classes)
  Comprimento e Largura Sépala
  Comprimento e Largura Pétala

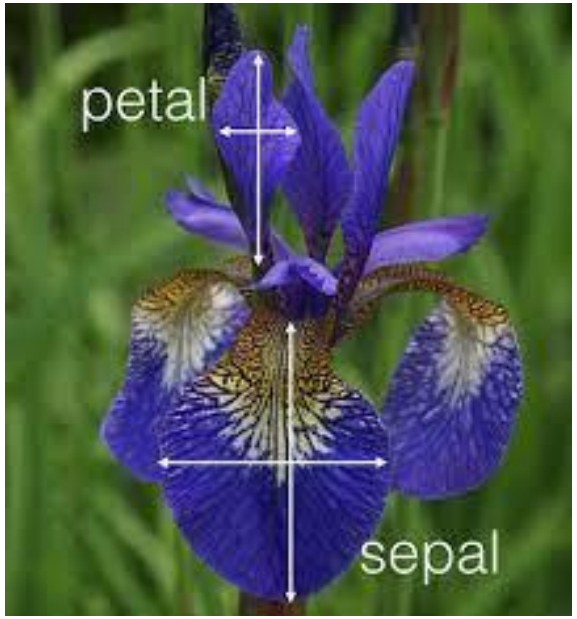
Atividade 02 – Análise dos dados IRIS_DATA

Faça uma análise por meio da visualização de dados do arquivo IRIS (iris_data.csv).

Utilize os softwares WEKA e ORANGE

Arquivo: Iris_Data.csv

Instância – N=150
Comprimento da Sépala
Largura da Sépala
Comprimento da Pétala
Largura da Pétala
Classe {íris-setosa,íris-virginica,íris-versicoulor}



Fonte: Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute

Atividade 03

Desenvolva as etapas de definição do seu projeto.

Sugestão:

a) Defina o negócio e o problema a solucionar
b) Identifique o tipo de tarefa de aprendizagem de máquina
c) Análise o problema e identifique os atributos relevantes para o processo.
d) Descreva o dicionário de dados do objeto definindo o tipo de atributo.
e) Gerar um conjunto de registro de dados para testes ou buscar uma base de dados.

Pré-Processamento

Tipos de Atributos

- **Nominal ou Categóricos:** Utilizados para nomear ou atribuir rótulos a objetos.
  - ✓ Exemplo: números de ID, cor dos olhos, estado civil, código do cep

- **Ordinal:** Assemelham-se aos nominais, porém os valores (estados) que elas podem assumir possuem um ordenamento**.**
  - ✓ Exemplo: escala entre 1-10 sobre uma pontuação de enquete, graus, altura em {alto, médio, baixo}

- **Intervalar:** Valores geralmente utilizados entre intervalos
  - ✓ Exemplo: data de calendário, temperaturas

- **Razão:** Variáveis quantitativas cujo valores possuem uma relação de ordem entre eles, geralmente representadas por um tipo de dado numérico. Seu conjunto pode ser finito ou infinito.
  - ✓ Exemplo: comprimento, tempo, contagem, renda, idade...

Propriedade dos Atributos

- **O tipo de um atributo depende das propriedades que ele possui:**

✓	Distinção:	=	<>
✓	Ordem:	<	>
✓	Adição:	+	-
✓	Multiplicação:	*	/

- **Atributo Nominal:** Distinção
- **Atributo Ordinal:** Distinção e Ordem
- **Atributo Intervalar:** Distinção, Ordem e Adição
- **Atributo Razão:** Distinção, Ordem, Adição e Multiplicação

Dados do Objeto

Dados que consistem de uma coleção de registros, cada um dos quais consiste de um conjunto fixo de atributos

=> Dados do Objeto

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Dados Matriciais

- Se os objetos de dados tem o mesmo conjunto fixo de atributos numéricos, então os objetos de dados podem ser vistos como pontos em um espaço multidimensional, em que cada dimensão representa um atributo distinto.
- Tal conjunto de dados pode ser representado por uma matriz m por n, em que há m linhas, uma para cada objeto, e n colunas, uma para cada atributo.

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Dados de Documentos

- Cada documento torna-se um vetor de “termos”,
  - Cada termo é um componente (atributo) do vetor,
  - O valor de cada componente é o número de vezes que o termo correspondente ocorre no documento.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Dados de Transações

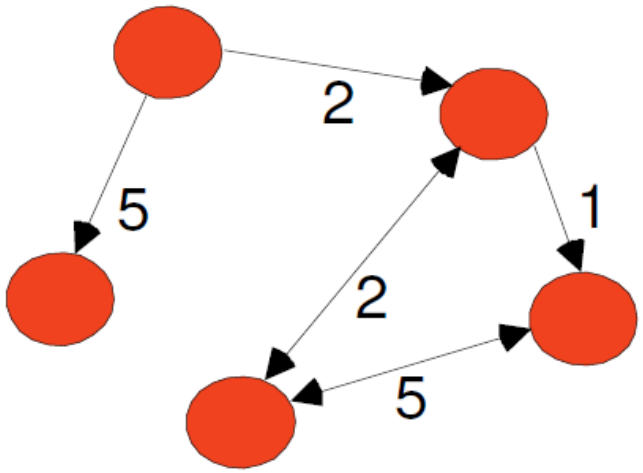
- São dados de registro de um tipo especial, em que;

  - Cada registro (transação) envolve um conjunto de itens. Exemplo: considere um supermercado. O conjunto de produtos comprados por um cliente durante as compras constitui-se uma transação, enquanto os produtos individuais são os itens.

ID	*Itens*
1	Pão, Refri, Leite
2	Cerveja, Pão
3	Cerveja, Refri, Fralda, Leite
4	Cerveja, Pão, Fralda, Leite
5	Refri, Fralda, Leite

Dados de Grafos

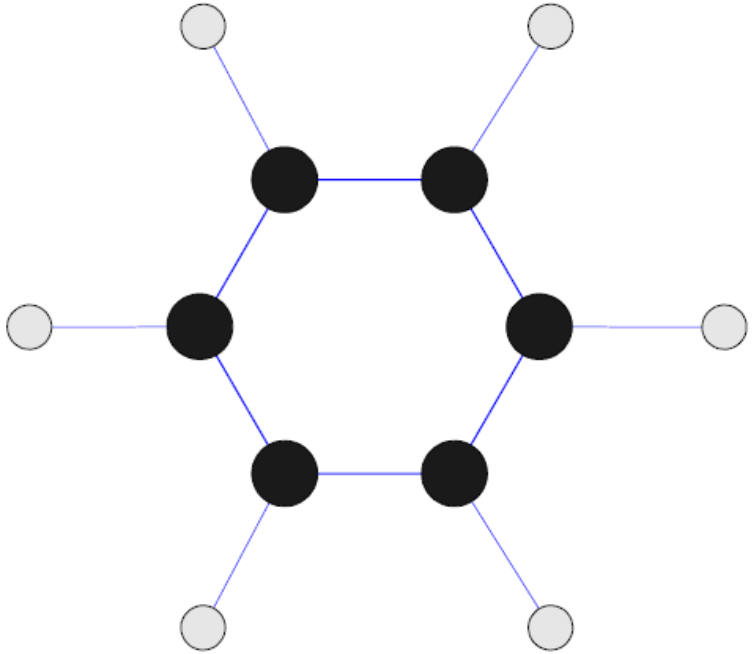
- São dados que representam uma sequência de ordenação. Exemplo: grafos genéricos e links HTML;




```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

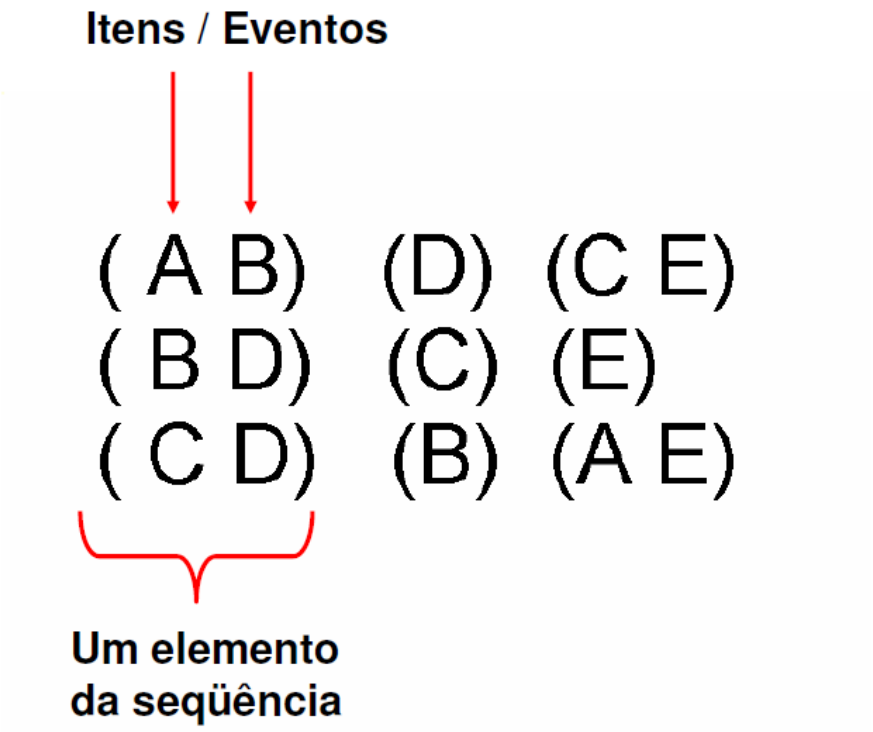
Dados de elementos Químicos

- Representa a estrutura de um elemento químico
  Exemplo: molécula de Benzeno: C_6H_6;



Dados Ordenados

- Dados que representam a sequencia de um transação, registro de eventos ou log de eventos;
  Exemplo: ordem de compras de produtos em um supermercado; registro de atendimento de pacientes, log de rastreamento de veículos, etc.



Dados de Sequência

- Consistem de um conjunto de dados que representam uma sequência de entidades individuais (como sequência de palavras ou letras).

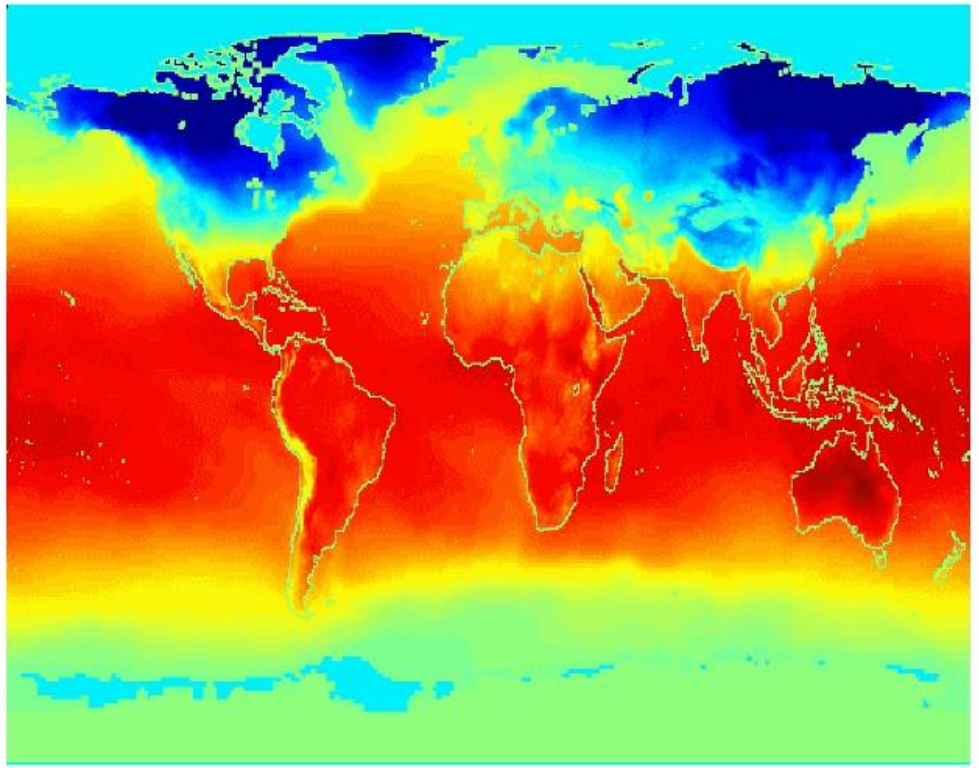
```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Dados de sequência de genoma (quatro nucleotídeos – ATG e C) do qual todo o DNA é construído.

Dados Espaço-Temporais

- Dados que identificam posições ou áreas.
  Exemplo: Dados Espaço-temporais;

Temperatura Média Mensal das terras e oceanos



Qualidade dos Dados

- **Quais são os tipos de problemas de qualidade de dados?**
- **Como se pode detectar problemas nos dados?**
- **Como resolver problemas com a qualidade dos dados?**

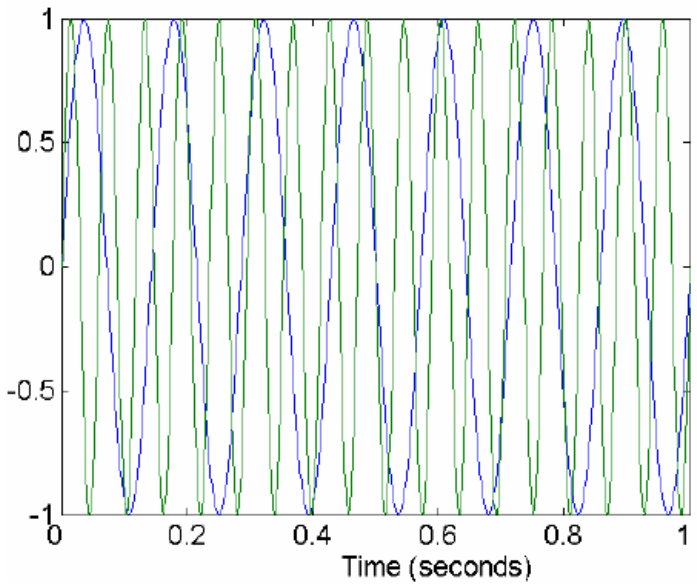
Exemplo de problemas de qualidade nos dados:

- Ruídos
- Outliers
- Dados faltantes
- Dados duplicados

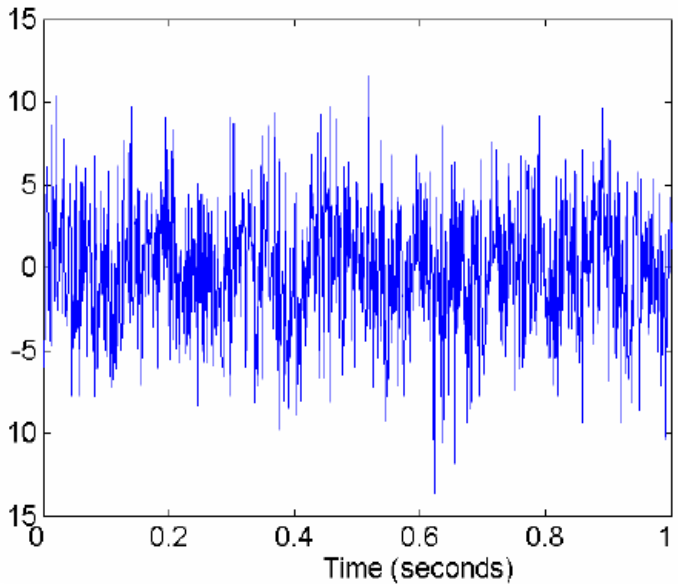
Ruído

- **Ruído** refere-se à modificação de valores originais. Também pode-se dizer que é um erro de medição.

Exemplo: distorção da voz de uma pessoa



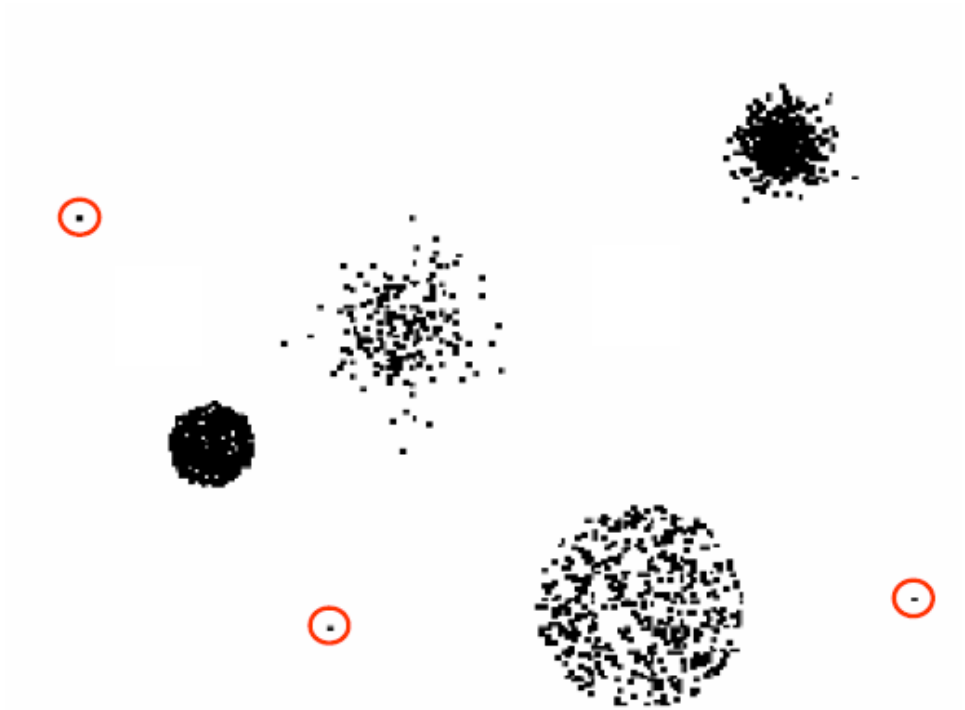
Duas ondas senoidais



Duas ondas senoidais + Ruído

Outliers

- **Outliers** são objetos de dados com características que são consideravelmente diferente da maioria dos outros objetos do conjunto de dados.



Valores Faltantes

- **Razões para valores faltantes ou ausentes**.
  - Informação não coletada.
  (Exemplo:pessoas que não fornecem sua idade e peso)
  - Atributos que não são aplicáveis a todos os casos.
  (Exemplo: salário não é aplicável as crianças).

- **Manipulando valores faltantes**.
  - **Eliminar** objeto de dados
  - **Estimar** valores faltantes
  - **Ignorar** valores faltantes durante análise
  - **Substituir** por todos os valores possíveis (ponderados por suas probabilidades).

Dados Duplicados

- **Conjunto de dados pode incluir objetos de dados que são duplicatas, ou muito semelhantes**.
  - União de dados de fontes heterogêneas.

- **Exemplos:**
  - Mesma pessoa com múltiplos endereços de email

KDD – Knowledge Discovery in Database

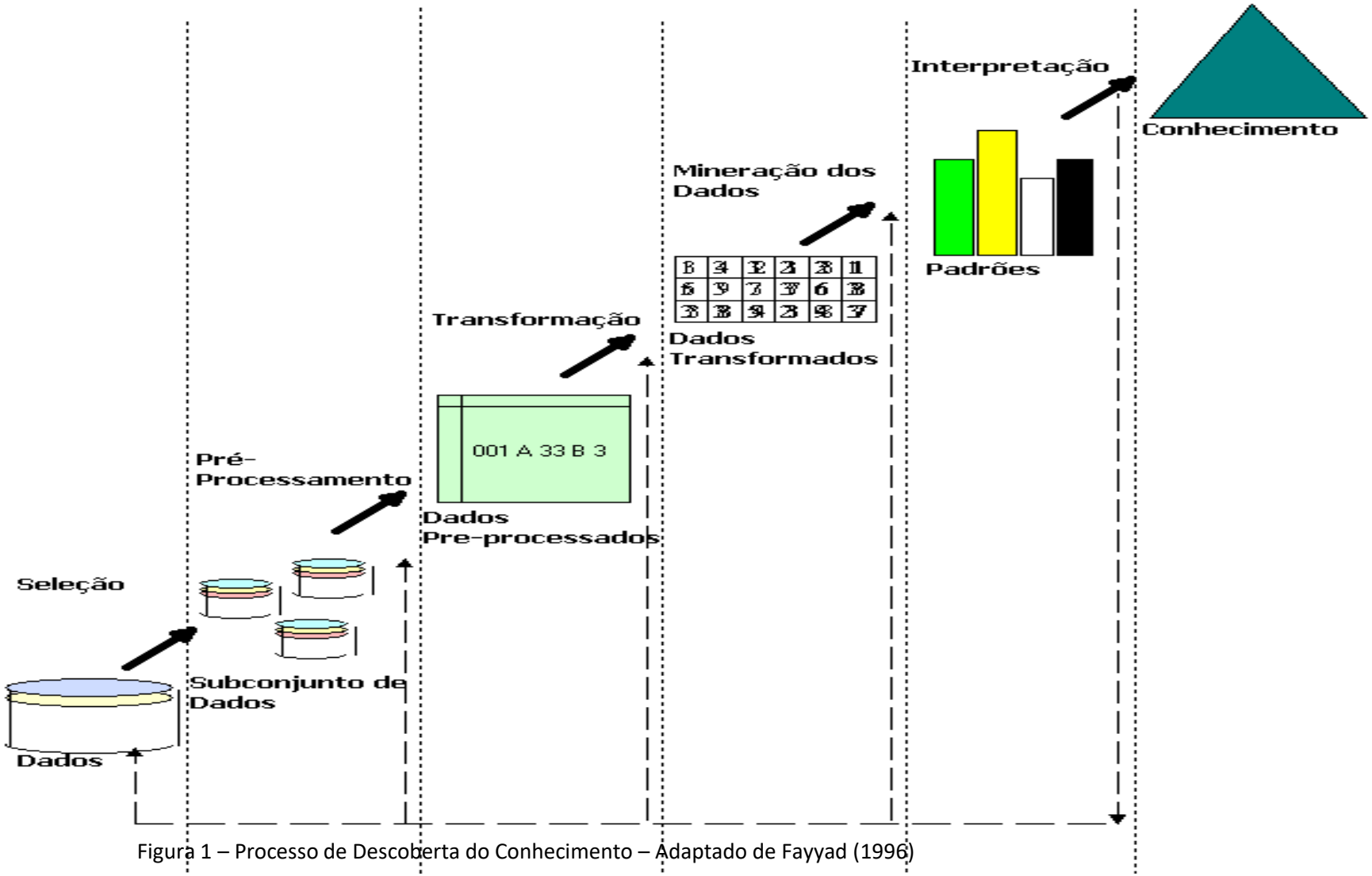


Figura 1 – Processo de Descoberta do Conhecimento – Adaptado de Fayyad (1996)

Pré-Processamento

A partir do objetivo podemos preparar um conjunto de dados.

- **Seleção.** Focar em um subconjunto
- **Limpeza.** Eliminar ruído
- **Enriquecimento.** Acrescentar dados externos
- **Transformação/Codificação.** Normalizar dados

Grande parte da preparação é feita quando temos o DW!

OLAM (mining)

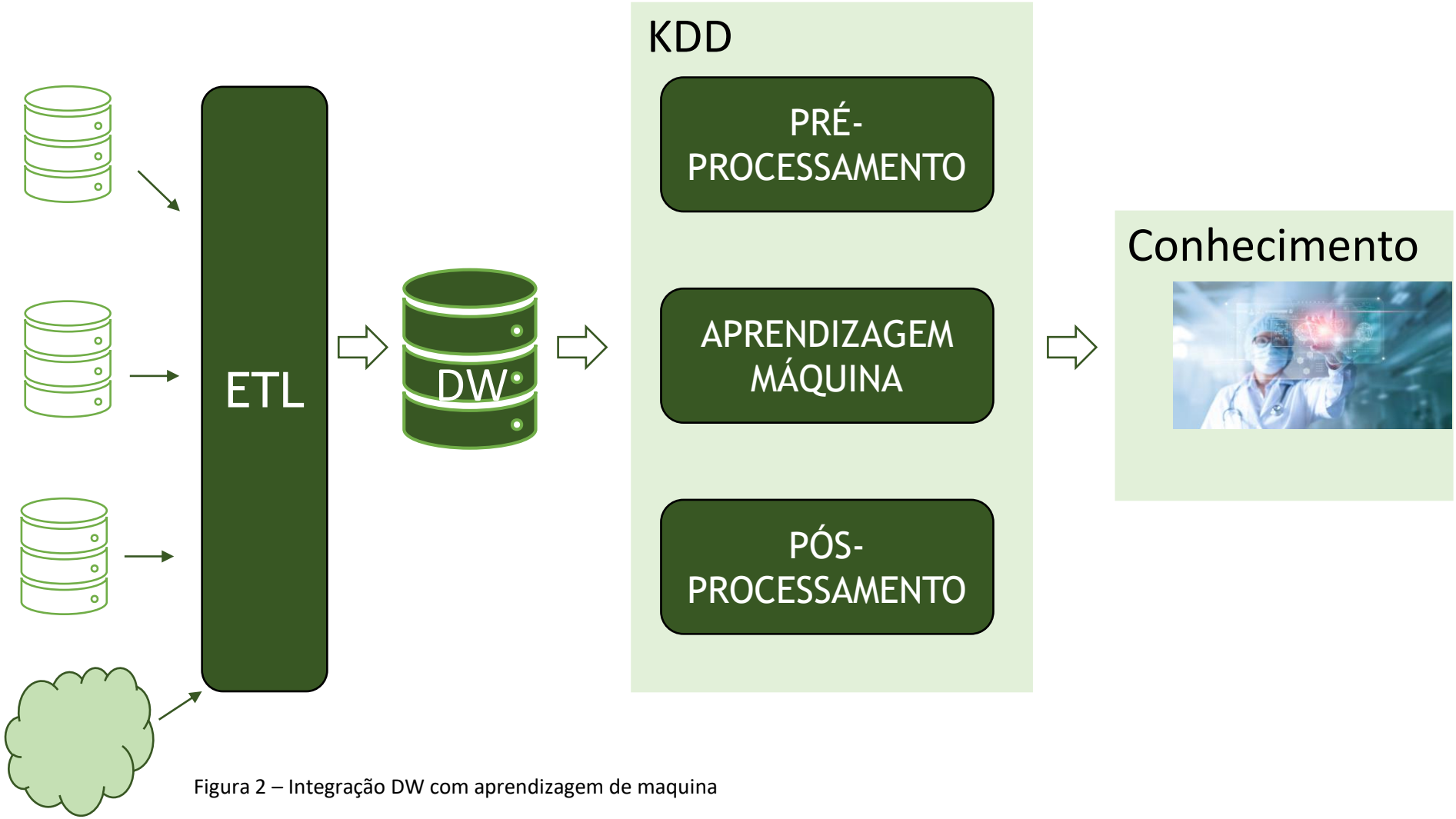


Figura 2 – Integração DW com aprendizagem de maquina

Pré-Processamento

- Agregação
- Amostragem
- Redução da Dimensionalidade
- Seleção de subconjunto de características
- Criação de características
- Discretização e Binarização
- Transformação de Atributos.

Agregação

- **Combinar diversos atributos (ou objetos) em um único atributo (ou objeto).**

- **Finalidade**

  - Redução de dados (Reduzir o número de atributos ou objetos);

  - Alteração de escala (Cidades agregadas em regiões, estados, países, ...)

  - Tornar os dados mais estáveis (Dados agregados geralmente tem menor variabilidade).

Amostragem

- **Amostragem é uma técnica empregada na seleção de dados:**

- **Finalidades:**
  - Usada frequentemente para a investigação preliminar dos dados quanto análise final dos dados;
  - Usa-se amostragem, porque obter o conjunto completo dos dados de interesse é muito caro ou consome muito tempo/recurso;
  - Amostragem é utilizada em DM porque o processamento do conjunto inteiro dos dados de interesse é muito caro ou consome muito tempo.

Amostragem

O princípio básico para amostragem:

- Usando uma amostra o processo funcionará tão bem quanto usando o conjunto completo de dados se a amostra é representativa.

- Uma amostra é representativa se ela tem aproximadamente as mesmas propriedades (de interesse) quanto o conjunto original de dados.

Tipos de Amostragem

- **Amostragem simples aleatória;**

- Probabilidade igual de selecionar qualquer item particular.

- **Amostragem sem substituição (reposição);**

- À medida que cada item é selecionado, ele é removido de todos os objetos que juntos constituem a população.

- **Amostragem com substituição (reposição);**

- Objetos não são removidos da população quando são selecionados para compor a amostra.Esse objeto pode ser selecionado mais de uma vez dentro da amostra.

- **Amostragem estratificada;**

- Divide os dados em várias partições; retira-se amostras aleatórias de cada uma das partições.

Tipos de Amostragem

8000 pontos

2000 pontos

500 pontos

Redução de Dimensionalidade

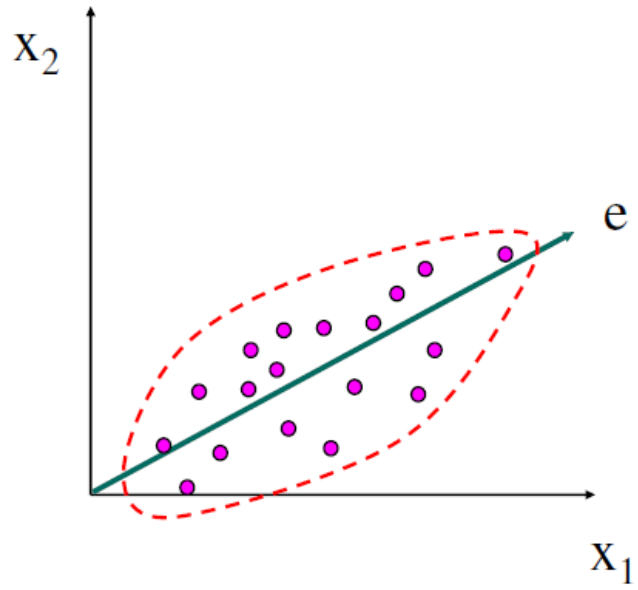
- **Finalidade**
  - Combater a maldição da dimensionalidade
  - Reduzir a quantidade de tempo e memória necessárias pelos algoritmos de mineração de dados
  - Permitir que os dados sejam mais facilmente visualizados
  - Ajudar a eliminar características irrelevantes ou a reduzir o ruído.

- **Técnicas**
  - Análise de Componentes Principais – PCA
  - Singular Value Decomposition – SVD
  - Outras: Técnicas supervisionadas e não-lineares

Redução de Dimensionalidade

Análise de Componentes Principais (PCA)

- O objetivo é encontrar a projeção que captura a maior quantidade de variação nos dados.



Seleção de Subconjuntos

Seleção de Subconjuntos de Características

Maneiras de reduzir a dimensionalidade dos dados;

- **Características redundantes**
  - Duplicam muita ou toda a informação contida em um ou mais atributos. Exemplo: preço de venda de um produto e a quantidade de taxas de venda pagas

- **Características irrelevantes**
  - Não contém informação que seja útil para a tarefa de mineração de dados. Exemplo: Id do Estudante é frequentemente irrelevante na tarefa de prever o seu desempenho

Seleção de Subconjuntos

Seleção de Subconjuntos de Características
Técnicas:

- **Abordagem de força bruta**
  - Tenta todos os subconjuntos possíveis de características como entrada para o algoritmo
- **Abordagem embutidas**
  - Seleção de características ocorre naturalmente como parte do algoritmo
- **Abordagem filtro**
  - Características são selecionadas antes que o algoritmo seja executado
- **Abordagem wrapper**
  - Uso do algoritmo como uma caixa preta para encontrar o melhor subconjunto de atributos

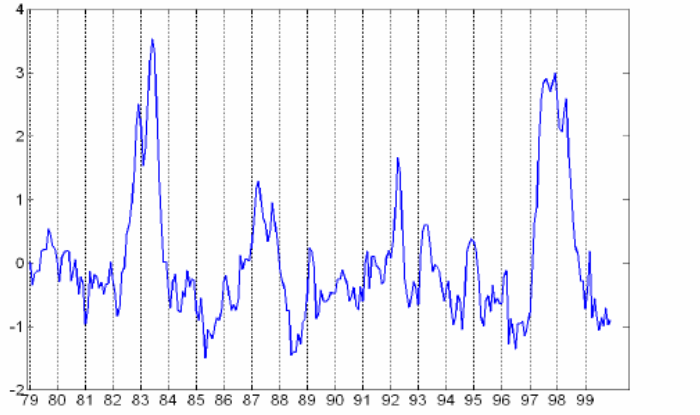
Criação de Características

- **Cria novos atributos que podem capturar informação importante em um conjunto de dados muito mais eficiente que os atributos originais**

- **Três metodologias gerais:**
  - Extração de características (específicas do domínio)
  - Mapeamento de dados para um novo espaço
  - Construção de características (combinação de características)

Transformação de Atributos

- **Uma função que mapeia o conjunto inteiro de valores de um dado atributo para um novo conjunto de valores de substituição tal que cada valor antigo pode ser identificado com um dos novos valores.**
  - Funções simples: x^k, log(x), e^x, |x|
  - Padronização e Normalização



Similaridade e Dissimilaridade

- **Similaridade**
  - Medida numérica de quão parecido dois objetos são;
  - É maior quando objetos são mais parecidos
  - Frequentemente está na faixa [0,1]
- **Dissimilaridade**
  - Medida numérica de quão diferentes dois objetos são;
  - Menor quando dois objetos são mais parecidos
  - Dissimilaridade mínima é frenquentemente 0
  - Limite superior varia.

- **Proximidade refere-se à similaridade ou dissimilaridade**

Similaridade e Dissimilaridade

p e *q* são os valores dos atributos para dois objetos de dados

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Similaridade e Dissimilaridade

- **Distância euclidiana**

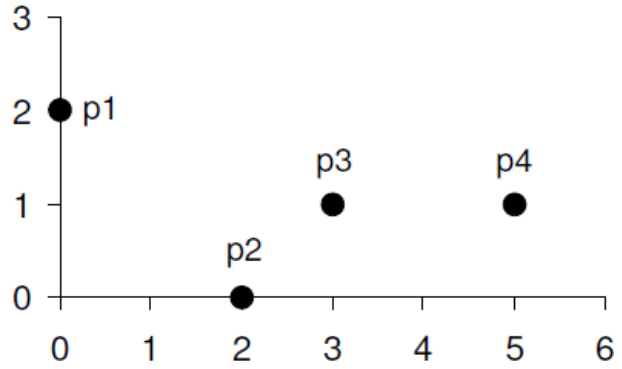
$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Em que *n* é o número de dimensões (atributos) e p_k e q_k são, respectivamente, os k-ésimo atributos (componentes) dos objetos de dados p e q.

Padronização é necessária se as escalas diferem.

Similaridade e Dissimilaridade

Distância euclidiana



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Matriz de distâncias

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Similaridade e Dissimilaridade

Distância Minkowski é uma generalização da distância Euclidiana

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

Em que *r* é um parâmetro, *n* é o número de dimensões (atributos) e p_k e q_k são, respectivamente, os k-ésimo atributos (componentes) dos objetos de dados p e q.

Similaridade e Dissimilaridade

Distância Minkowski

- **r=1.** Distância City Block (ou Manhattan, taxicab, norma L_1);

- **r =2.** Distância Euclidiana

- **Rr = ∞.** Distância "supremum" (norma L_{max}, norma L_∞).

  - É a diferença máxima entre quaisquer componentes dos vetores

  - Não confundir r com n. Exemplo: todas estas distâncias são definidas para todos os números de dimensões.

Correlação

- Correlação mede o relacionamento linear entre objetos;

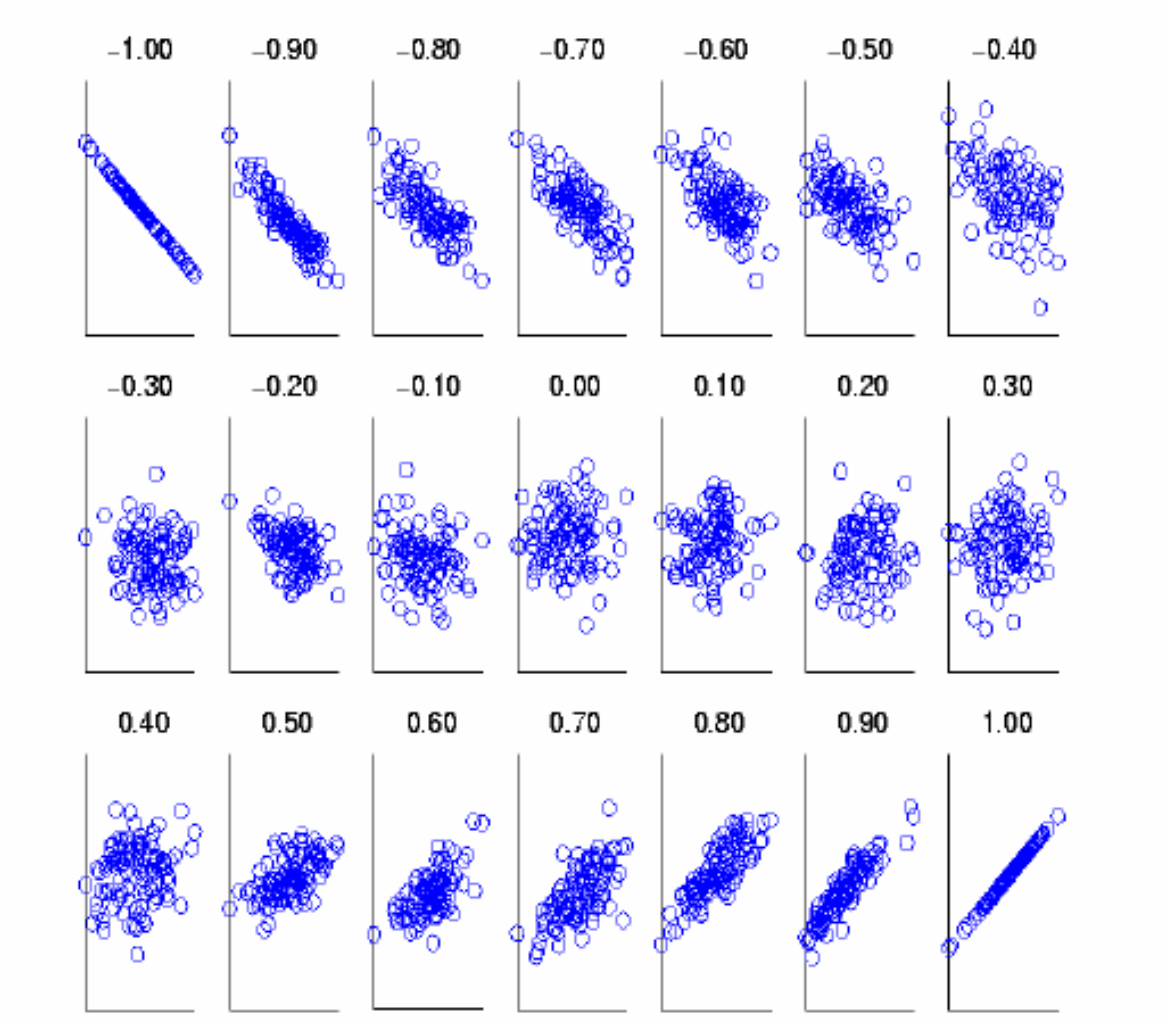
- Para calcular a correlação, padronizam-se os objetos de dados, p e q, e faz-se seu produto interno;

$$p'_k = (p_k - média(p)) / desvpad(p)$$

$$q'_k = (q_k - média(q)) / desvpad(q)$$

$$correlação(p, q) = p' \bullet q'$$

Avaliando a Correlação



Gráficos de dispersão mostrando a similaridade de –1 a 1.

Densidade

- Agrupamento baseado em densidade requer a noção de densidade;

- Exemplos;
  - Densidade euclidiana
    Densidade euclidiana = número de pontos por unidade de área

  - Densidade de probabilidade

  - Densidade baseada em grafo

Densidade

- Abordagem mais simples é dividir a região em um número de células retangulares de igual área e definir densidade como número de pontos que a célula contém;

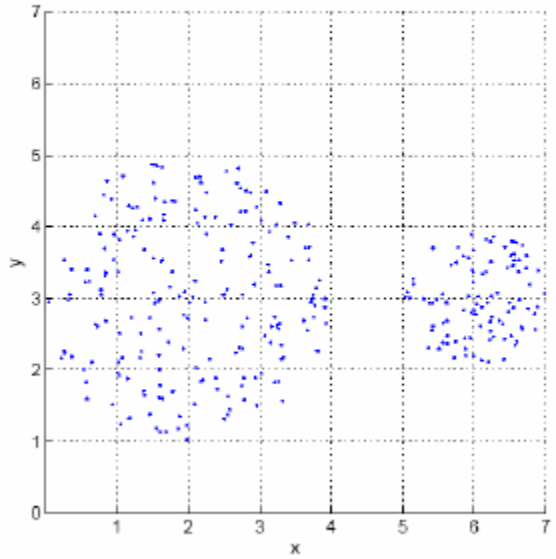


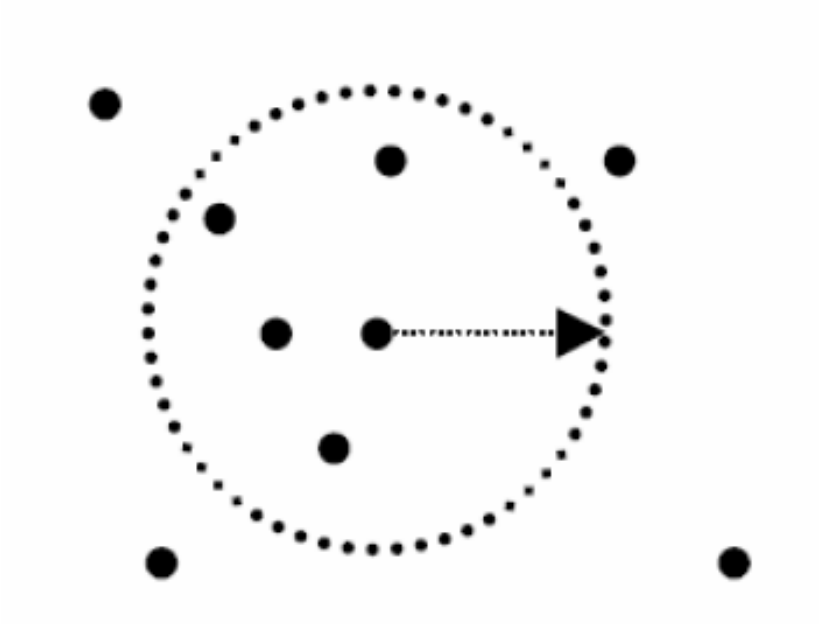
Figure 7.13. Cell-based density.

0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

Table 7.6. Point counts for each grid cell.

Densidade Euclidiana

- Densidade Euclidiana baseada em centro é o número de um raio específico a partir do ponto;



Atividade 04

Desenvolva as etapas de seleção e pré-processamento de seu projeto, definindo os dados do seu objeto;

Sugestão:

a) Defina o negócio e o problema a solucionar
b) Identifique o tipo de tarefa de aprendizagem de máquina
c) Análise o problema e identifique os atributos relevantes para o processo.
d) Descreva o dicionário de dados do objeto definindo o tipo de atributo.
e) Gerar um conjunto de registro de dados para testes ou buscar uma base de dados.

Espaço aberto para dúvidas e opiniões