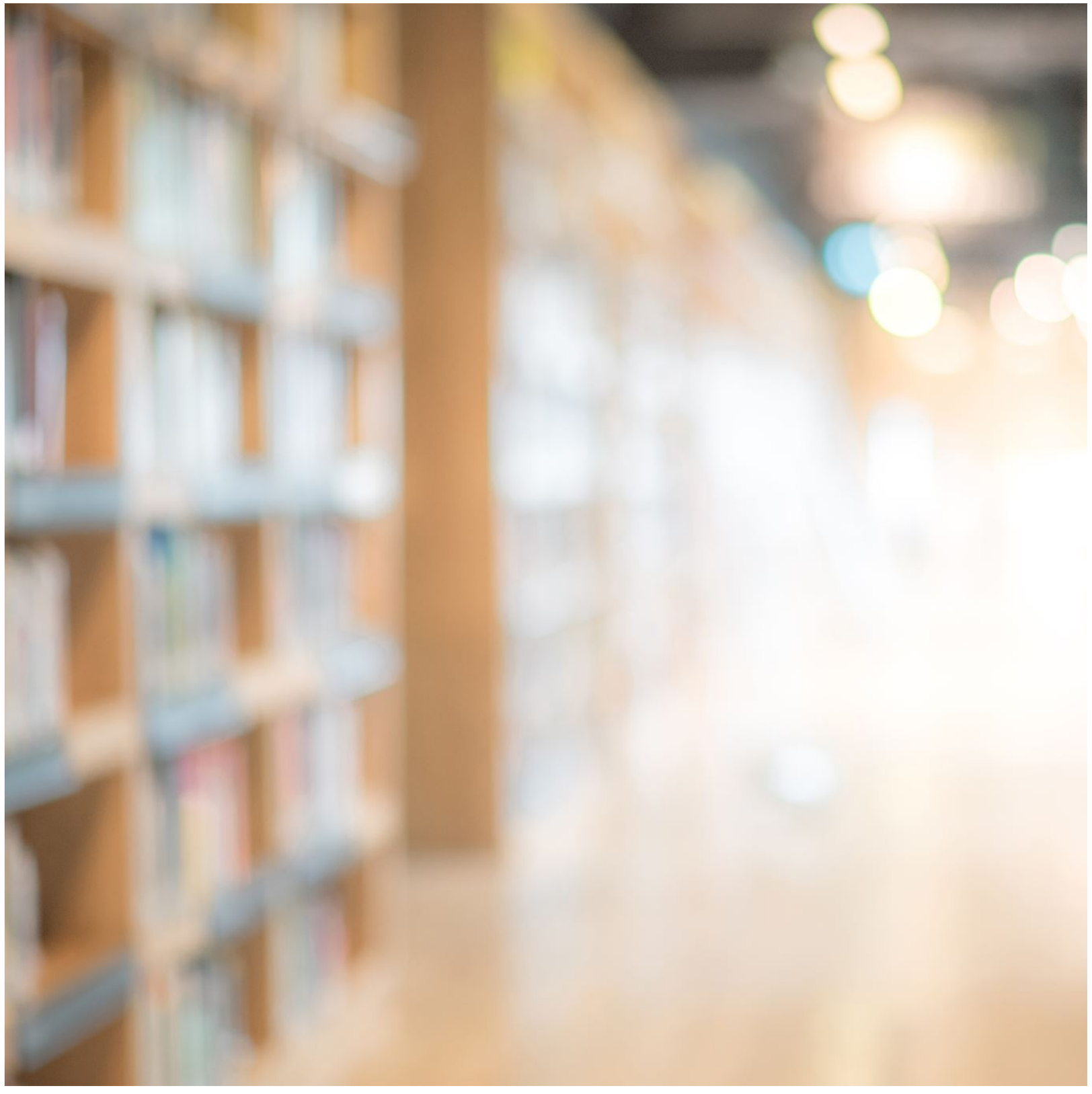

DADOS
Exploração de Dados

Prof. Dra. Karina S. Machado

PPGCOMP

C3

FURG

O que são os dados???

- Conjuntos de dados são formados por objetos que podem representar um objeto físico como uma cadeira ou uma noção abstrata, como sintomas apresentados por um paciente

- Formalmente os dados podem ser representados por uma matriz de objetos X nxd em que “n” é o número de objetos e “d” é o número de atributos de entrada de cada objeto.
- “d” -> dimensionalidade dos dados

TAN, Pang-Ning et al. Introduction to data mining. 2nd ed. New York: Pearson, 2019. 839

FACELI, Katti et al. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2017. 378 p.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. Data mining: concepts and techniques. 3rd ed. Amsterdam: Elsevier, 2012. 703 p.

O que são os Dados?????

- Coleção de objetos e seus atributos
  - Um atributo é uma propriedade ou característica de um objeto que pode variar entre objetos ou no tempo
  - Exemplo: cor do olho, temperatura, idade, etc.
  - Atributo também é conhecido como variável, campo, característica (feature).
  - Um conjunto de atributos descrevem um objeto
  - Objeto também é conhecido como registro, ponto, caso, amostra, exemplo ou **instância**

Atributos

Objetos ou Instâncias

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

TAN, Pang-Ning et al. Introduction to data mining. 2nd ed. New York: Pearson, 2019. 839

FACELI, Katti et al. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2017. 378 p.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. Data mining: concepts and techniques. 3rd ed. Amsterdam: Elsevier, 2012. 703 p.

Exemplo

Id	Nome	idade	sexo	peso	mancha	T	Es	D
1	Joao	28	M	79	Concen	38	SP	D
2	Maria	18	F	67	Inex	39	MG	D
3	Luis	49	M	92	Esp	39	RS	S
4	José	18	M	43	Inex	37	RS	D
5	Claudia	21	F	52	Uniform	37	RS	S
6	Ana	55	F	72	Esp	38	BA	D

OBJETO??
ATRIBUTOS???

ATRIBUTO ALVO??

TAN, Pang-Ning et al. Introduction to data mining. 2nd ed. New York: Pearson, 2019. 839

FACELI, Katti et al. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2017. 378 p.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. Data mining: concepts and techniques. 3rd ed. Amsterdam: Elsevier, 2012. 703 p.

Valores dos Atributos:

- Valores dos atributos são números ou símbolos associados a um atributo.
- Há diferença entre atributos e valores dos atributos:
  - O mesmo atributo pode ser mapeado para diferentes valores: por exemplo a altura pode ser medida em metros ou pés.
- Diferentes atributos podem ser mapeados para o mesmo conjunto de valores:
- Exemplo: valores de ID e Idade são ambos inteiros
- Mas as propriedades dos atributos são diferentes:
  - ID não tem limite e idade tem valores mínimos e máximos.

TAN, Pang-Ning et al. Introduction to data mining. 2nd ed. New York: Pearson, 2019. 839

FACELI, Katti et al. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2017. 378 p.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. Data mining: concepts and techniques. 3rd ed. Amsterdam: Elsevier, 2012. 703 p.

Tipos dos atributos

- Qualitativo / quantitativo

- Variáveis qualitativas:
- Variáveis quantitativas:
- Para o exemplo

Id	Nome	idade	sexo	peso	mancha	T	Es	D

Tipos dos atributos

- Qualitativo / quantitativo

- Variáveis qualitativas:
- Variáveis quantitativas:
- Para o exemplo

Id	Nome	idade	sexo	peso	mancha	T	Es	D
Quali	quali	Quant disc	Quali	quant	quali	quanti	Quali cont	quali

Tipos dos Atributos

- Há diferentes escalas para os atributos:
  - Escala Nominal
    - Apenas nomes diferentes, menor quantidade de informação possível
    - Valores não numericos e não ordenados
    - Exemplo: ID, cor do olho, CEP
  - Escala Ordinal
    - Escala não numerico e ordenado. Podem ser comparados se é maior ou menor
    - Exemplo: rankings, grau de escolaridade, etc.

Tipos dos Atributos

- Escala Interval
  - Nessa escala de valores numéricos, existe não apenas uma ordem entre os valores, mas também existe diferença entre esses valores.
  - Os atributos intervalares são números que variam dentro de um intervalo
  - O zero é relativo.
    - Ex: Temperatura em Graus Celsius
- Escala Proporcional ou racional
  - São os que carregam mais informações
  - Nessa escala de valores numéricos, além da diferença, tem sentido calcular a proporção entre valores (o zero é absoluto).
  - Temperaturas em K, Valores monetários, etc.

Escala: Propriedades dos valores dos atributos

- A escala define as operações que podem ser realizadas sobre os valores do atributo
  - Distinção ≠ =
  - Ordenação < >
  - Adição + -
  - Multiplicação * /

- Atributos nominais: distinção (sem relação de ordem)
- Atributos ordinais: distinção e ordem
- Atributos de intervalo: distinção, ordem e adição
- Atributos de proporção ou racionais: as 4 propriedades

Para o exemplo

Id	Nome	idade	sexo	peso	mancha	T	Es	D

FACELI, Katti et al. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2017. 378 p.

Para o exemplo

Id	Nome	idade	sexo	peso	mancha	T	Es	D
Nominal	Nominal	racional	Nominal	racional	nominal	intervalar	nominal	nominal

FACELI, Katti et al. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2017. 378 p.

Atributos discretos e contínuos:

- Atributo Discreto (Classes)
  - Tem somente um conjunto de valores finito e contável
  - Exemplo: CEP, conjunto de palavras de uma coleção de documentos
  - Geralmente representado por variáveis inteiras
  - Atributos binários são um caso especial de atributos discretos
- Atributo Contínuo
  - Tem um número real como valor do atributo
  - Podem assumir qualquer valor dentro de um intervalo.
  - Exemplo: temperatura, altura, peso…

Para o exemplo

Id	Nome	idade	sexo	peso	mancha	T	Es	D
Nominal	Nominal	racional	Nominal	racional	nominal	intervalar	nominal	nominal

FACELI, Katti et al. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2017. 378 p.

Para o exemplo

Id	Nome	idade	sexo	peso	mancha	T	Es	D
Nominal Discreto	Nominal Discreto	Racional Contínuo	Nominal discreto	racional contínuo	Nominal discreto	Intervalar contínuo	Nominal discreto	Nominal discreto

FACELI, Katti et al. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2017. 378 p.

Explorando bases de dados

Prof. Dra. Karina dos Santos Machado

Bases de Dados públicas

Há muitas bases de dados publicas

Em algumas aulas, vamos usar a base de dados Kaggle

Visite esse site, explore as bases disponíves

https://www.kaggle.com/datasets

Explorando uma base de dados

- Comece analisando
  - Quantidade de Instâncias
  - Quantidade de Atributos
  - Tipos dos atributos (discreto ou contínuo)
  - Tem atributo alvo? (para tarefas preditivas) Esse atributo alvo é discreto ou contínuo?

- É muito importante conhecer a qualidade dos dados antes de qualquer tarefa de pré-processamento e/ou mineração de dados.
- Para conhecer os dados e analisar sua qualidade, utilize estatística, gráficos, etc.

Explorando uma base de dados

- Estatística Descritiva
  - Resumo quantitativo das características de um conjunto de dados: media, desvio padrão, mediana, mínimo, máximo, quartis e percentis

- Exemplo para o conjunto hospital:
  - Idade média dos pacientes
  - Porcentagem dos pacientes que moram em SP

Explorando uma base de dados

- Gráficos
- Histograma: permite ver a distribuição de valores de um atributo. Você pode usar para definer como discretizar um atributo contínuo em classes, pode usar para ver se determinado atributo apresenta outliers.
- Scatter plots e box plots: permite identificar outliers, relacionar atributos
- Matrizes de correlação: permite verificar se há dados redundantes por exemplo
- Medidas estatísticas:, etc… podem ser usados para conhecermos os valores de um determinado atributo numérico.