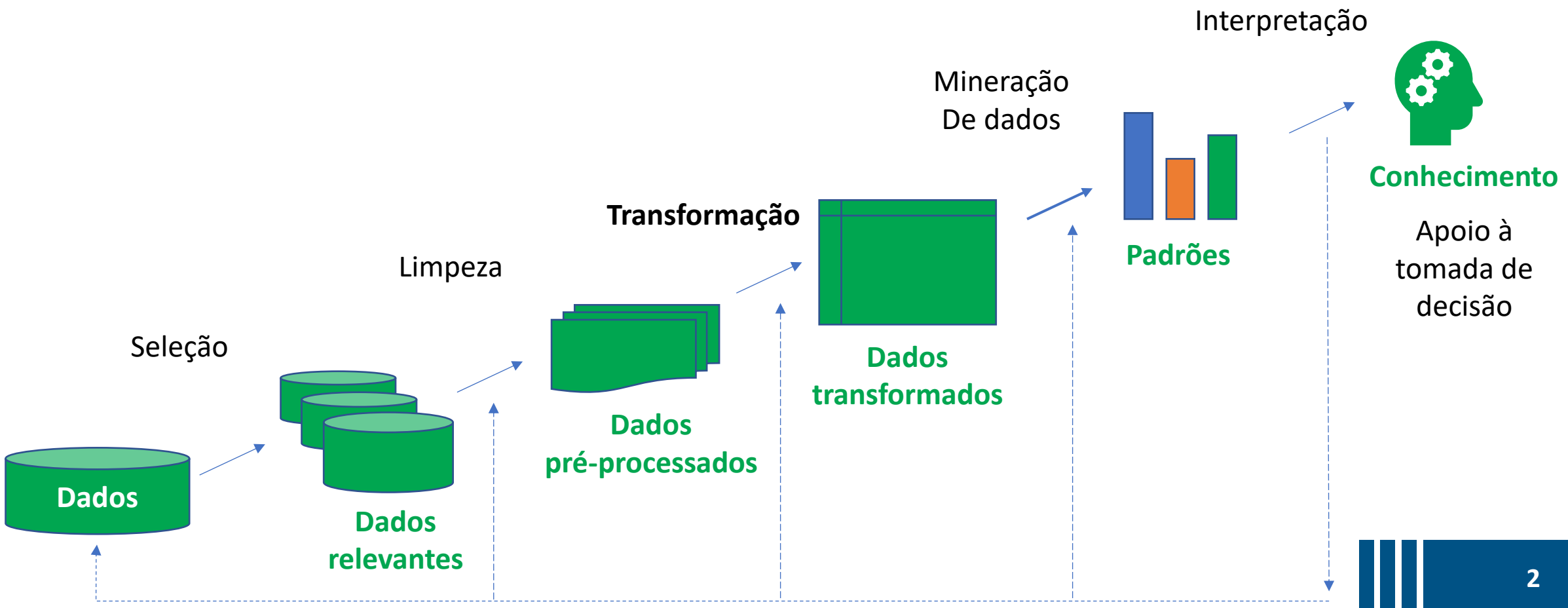
Transformação de Dados

Ciência de Dados II

Professor: Gabriel Machado Lunardi
gabriel.lunardi@ufsm.br

O processo de KDD

"É um processo de várias etapas, não trivial, **interativo** e **iterativo**, para a identificação de **padrões** válidos, novos e potencialmente úteis a partir de um grande conjunto de dados" (FAYYAD, 1996).



Introdução à Transformação de dados

- ✓Várias técnicas de aprendizado de máquina são limitadas à manipulação de valores de determinados tipos,
  - ✓valores numéricos
  - ✓ou simbólicos (categóricos).

- ✓Converter de um tipo para o outro e vice-versa conforme a necessidade.

- ✓Conversão Numérico-simbólico.

- ✓Conversão Simbólico-numérico.

- ✓Amostragem

Amostragem

Amostragem

- Algoritmos de AM podem ter:
  - Alta complexidade computacional
  - Grande demanda de memória
- Consequências:
  - Treinamento lento
  - Dificuldade de processamento

Solução: Utilização de Amostras

Geração de **subconjuntos menores de dados**

- Agilizam o treinamento
- Simplificam a geração do modelo

Técnicas de amostragem

- Amostras **não representativas** levam a:
  - Modelos com baixa eficiência
  - Falha em capturar a distribuição real dos dados

Solução: Técnicas de Amostragem Estatística

- Garantem amostras informativas e representativas
- Técnicas comuns:
  - **Amostragem Simples**
    - Seleção aleatória sem viés (com ou sem reposição)
  - **Amostragem Estratificada**
    - Divisão em subgrupos (estratos) para preservar proporções

Amostragem – exemplo

ID	Diagnóstico	Textura (1-10)	Tamanho (mm)	Idade
1	M	6	25	45
2	B	2	12	32
3	M	8	30	58
4	B	3	15	40
5	B	1	10	28
6	M	7	22	50
7	B	2	11	35
8	M	9	28	60
9	B	4	18	42
10	M	5	20	55

Considere o dataset de exemplo ao lado que versa sobre diagnósticos de câncer.

Amostragem simples

- ✓Vamos extrair uma amostra simples
- ✓Observe como tivemos um desbalanceamento

- ✓Exemplo de **amostra 1**

ID	Diagnóstico	Textura (1-10)	Tamanho (mm)	Idade
1	M	6	25	45
3	M	8	30	58
6	M	7	22	50
7	B	2	11	35

Amostragem simples

- ✓Vamos extrair uma amostra simples
- ✓Observe como tivemos um **balanceamento por sorte**

- ✓Exemplo de **amostra 2**

ID	Diagnóstico	Textura (1-10)	Tamanho (mm)	Idade
1	M	6	25	45
4	B	3	15	40
6	M	7	22	50
7	B	2	11	35

Amostragem simples

✅ Quando usar

- **Dados são naturalmente balanceados**

❌ Limitações

- Risco de **amostras desbalanceadas** em pequenos conjuntos de dados.
- Pode **subrepresentar classes minoritárias**
  - ex.: tumores raros.

Amostragem estratificada

- ✓ Observe como a proporção original das classes é mantida.

- ✓ Exemplo de **amostra**
  - ✓ Se extraíssemos outra amostra, teríamos o mesmo comportamento

ID	Diagnóstico	Textura (1-10)	Tamanho (mm)	Idade
1	M	6	25	45
6	M	7	22	50
4	B	3	15	40
9	B	4	18	42

Amostragem simples

✅ **Quando usar:**

- **Dados desbalanceados**
  - ex.: 90% benignos vs. 10% malignos.
- **Classes minoritárias são importantes**
- **Validar modelos**
  - métricas precisas (ex.: recall para câncer).

❌ **Limitações:**

- Requer **conhecimento prévio** das distribuições das classes.
- Pode ser **computacionalmente mais complexo**

Exemplo:

- Prever fraudes em transações
  - fraudes são raras, mas críticas.

Comparação das técnicas de amostragem

Critério	Amostragem Simples	Amostragem Estratificada
Balanceamento	Assume equilíbrio	Força proporções
Complexidade	Baixa	Moderada
Uso Típico	Exploratório/Rápido	Modelos críticos
Risco de Viés	Alto (se desbalanceado)	Baixo

Conversão numérico-simbólico

Conversão numérico-simbólico

- Se o atributo numérico foi discreto e binário, a conversão é trivial. Basta associar um nome a cada valor
  - Exemplo

Atributo original (valores possíveis)	Atributo transformado
1 ou 0	Sim ou Não

- Caso não exista ordem entre os atributos numéricos, basta associar um nome ou uma categoria a cada um.
  - Exemplo

Atributo original (valores possíveis)	Atributo transformado
2, 4, 1, 7	Casa, apartamento, terreno, trailer

Conversão numérico-simbólico

- Nos demais casos, métodos de **discretização** permitem transformar atributos quantitativos (numéricos) em qualitativos. Para isso, transformam valores numéricos em intervalos ou categorias.

  - Dependem do tipo de algoritmo que será utilizado na etapa de mineração.
  - Discretização paramétrica.
  - Discretização não-paramétrica.

Normalização

Normalização de dados

- ✓ É o processo de **uniformizar os valores** dos dados.

- ✓ É recomendável quando os limites de valores de atributos distintos são muito diferentes. Ela evita que um atributo predomine sobre o outro (a menos que isso seja importante).

- ✓ Normalização por amplitude

Normalização por amplitude

- ✓ Pode acontecer por **reescala** ou por **padronização**.

- ✓ Reescala (min-max)
  - ✓ Define uma nova escala de valores, limites mínimo e máximo, para todos os atributos.

$$x_{ij} = \frac{x_{ij} - \min_j}{\max_j - \min_j}$$

- ✓ Padronização
  - ✓ Define um valor central e um valor de espalhamento comuns para todos os atributos.

$$x_{ij} = \frac{x_{ij} - \bar{x}_{.j}}{s_j}$$

Quando usar uma ou outra?

Característica	Min-Max (Reescala)	Z-Score (Padronização)
Intervalo	[0, 1] ou [-1, 1]	Sem limites fixos
Média	Não necessariamente 0	Sempre 0
Desvio Padrão	Varia	Sempre 1
Efeito em Outliers	Sensível	Mais robusto
Uso típico	Redes Neurais, K-NN	Regressão, SVM, PCA

- **Min-Max** se precisar de dados em uma escala fixa e sem outliers extremos.
- **Z-Score** se os dados tiverem outliers ou se o algoritmo assumir normalidade.

Conversão simbólico-numérico

Conversão simbólico-numérico

- Atributo categórico **com dois valores possíveis**
  - Um dígito binário é suficiente
  - Exemplo 1: atributo categórico **nominal**

Tem manchas na pele	
Valores possíveis do atributo original	Valores possíveis para o atributo transformado
Sim ou Não	1 ou 0

  - Exemplo 2: atributo categórico **ordinal**

Risco de colisão	
Valores possíveis do atributo original	Valores possíveis para o atributo transformado
Baixo ou Alto	0 ou 1

Conversão simbólico-numérico

- ✓ Exemplo 3: conjunto de dados hospital

Nome	Idade	Sexo	Peso	Manchas	Temp.	#Int.	Estado	Diagnóstico
João	28	M	79	Grandes	38,0	2	SP	Doente
Maria	18	F	67	Pequenas	39,5	4	MG	Doente
Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
José	18	M	43	Grandes	38,5	8	MG	Doente
Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
Marta	19	F	87	Grandes	39,0	6	AM	Doente
Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Conversão simbólico-numérico

- ✓ Exemplo 3: conjunto de dados hospital

Nome	Idade	Sexo	Peso	Manchas	Temp.	#Int.	Estado	Diagnóstico
João	28	M	79	Grandes	38,0	2	SP	Doente
Maria	18	F	67	Pequenas	39,5	4	MG	Doente
Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
José	18	M	43	Grandes	38,5	8	MG	Doente
Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
Marta	19	F	87	Grandes	39,0	6	AM	Doente
Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Conversão simbólico-numérico

- ✓ Exemplo 3: conjunto de dados hospital

Nome	Idade	Sexo	Peso	Manchas	Temp.	#Int.	Estado	Diagnóstico
João	28	0	79	Grandes	38,0	2	SP	Doente
Maria	18	1	67	Pequenas	39,5	4	MG	Doente
Luiz	49	0	92	Grandes	38,0	2	RS	Saudável
José	18	0	43	Grandes	38,5	8	MG	Doente
Cláudia	21	1	52	Médias	37,6	1	PE	Saudável
Ana	22	1	72	Pequenas	38,0	3	RJ	Doente
Marta	19	1	87	Grandes	39,0	6	AM	Doente
Paulo	34	0	67	Médias	38,4	2	GO	Saudável

M = 0
F = 1

Conversão simbólico-numérico

- ✓ Atributo categórico **com mais de dois** valores possíveis
  - ✓ A conversão dependerá se o atributo é **nominal** ou **ordinal**.

Nome	Idade	Sexo	Peso	Manchas	Temp.	#Int.	Estado	Diagnóstico
João	28	M	79	Grandes	38,0	2	SP	Doente
Maria	18	F	67	Pequenas	39,5	4	MG	Doente
Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
José	18	M	43	Grandes	38,5	8	MG	Doente
Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
Marta	19	F	87	Grandes	39,0	6	AM	Doente
Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Conversão simbólico-numérico

- ✓ Atributo categórico **nominal** **com mais de dois** valores possíveis
  - ✓ A inexistência de ordem entre os valores deve ser mantida após a conversão.
  - ✓ Para isso, utilizar a codificação canônica (também conhecido como **one-hot-encoding**).
  - ✓ Podemos utilizar uma quantidade **C de bits** para representar os valores.
  - ✓ Exemplo isolado: atributo cor.

Cor
Azul
Amarelo
Vermelho
Verde
Azul
Vermelho
Amarelo
Amarelo

Cor
Azul
Amarelo
Vermelho
Verde

Cor
1000
0100
0010
0001

A distância entre um valor e outro é a mesma para qualquer valor, respeitando a inexistência de ordem nos valores transformados em reação aos originais.

Dependendo do quantidade de valores categóricos distintos, essa codificação gerará cadeias de bits muito grandes. Por exemplo: 193 nomes de países.

Conversão simbólico-numérico

- Atributo categórico **nominal** **com mais de dois** valores possíveis

  - Quando tivermos muitas categorias como os países do exemplo anterior, podemos representa-los por um **conjunto de pseudoatributos** do tipo binário, inteiro ou real.

Pseudoatributo	Valor	Tipo
Continente	7	Inteiro
PIB	1	Real
População	1	Inteiro
Temp Méd. anual	1	Real
Área	1	Real

  - A combinação desses 5 pseudoatributos criados representa um país.

Conversão simbólico-numérico

- ✓ Atributo categórico ordinal **com mais de dois** valores possíveis
  - ✓ A relação de ordem deve ser preservada.
  - ✓ Devemos ordenar os valores ordinais e codificar cada um de acordo com sua posição na ordem com um valor inteiro ou um valor real.
  - ✓ **Exemplo 1**: atributo isolado ranking:

Ranking
Primeiro
Terceiro
Segundo
Primeiro
Segundo
Quarto
Segundo
Quarto

Ranking
Primeiro
Segundo
Terceiro
Quarto

Ranking
0
1
2
3

A distância entre um valor e outro varia de acordo com a posição. Portanto, isso evidencia a preservação da ordem em relação aos dados originais

Conversão simbólico-numérico

- ✓ Atributo categórico **ordinal** **com mais de dois** valores possíveis
  - ✓ **Exemplo 2**: dataset hospital

Nome	**Idade**	**Sexo**	**Peso**	**Manchas**	**Temp.**	**#Int.**	Estado	Diagnóstico
João	28	M	79	Grandes	38,0	2	SP	Doente
Maria	18	F	67	Pequenas	39,5	4	MG	Doente
Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
José	18	M	43	Grandes	38,5	8	MG	Doente
Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
Marta	19	F	87	Grandes	39,0	6	AM	Doente
Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Conversão simbólico-numérico

- ✓ Atributo categórico **ordinal** **com mais de dois** valores possíveis
  - ✓ **Exemplo 2**: dataset hospital

Nome	**Idade**	**Sexo**	**Peso**	**Manchas**	**Temp.**	**#Int.**	Estado	Diagnóstico
João	28	M	79	Grandes	38,0	2	SP	Doente
Maria	18	F	67	Pequenas	39,5	4	MG	Doente
Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
José	18	M	43	Grandes	38,5	8	MG	Doente
Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
Marta	19	F	87	Grandes	39,0	6	AM	Doente
Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Conversão simbólico-numérico

- ✓ Atributo categórico **ordinal** **com mais de dois** valores possíveis
  - ✓ **Exemplo 2**: dataset hospital

Nome	**Idade**	**Sexo**	**Peso**	**Manchas**	**Temp.**	**#Int.**	Estado	Diagnóstico
João	28	M	79	3	38,0	2	SP	Doente
Maria	18	F	67	1	39,5	4	MG	Doente
Luiz	49	M	92	3	38,0	2	RS	Saudável
José	18	M	43	3	38,5	8	MG	Doente
Cláudia	21	F	52	2	37,6	1	PE	Saudável
Ana	22	F	72	1	38,0	3	RJ	Doente
Marta	19	F	87	3	39,0	6	AM	Doente
Paulo	34	M	67	2	38,4	2	GO	Saudável

1 – Pequena
2 - Média
3 - Grande

Conversão simbólico-numérico

- ✓ Atributo categórico **ordinal** **com mais de dois** valores possíveis
  - ✓ Caso seja preciso converter valores ordinais em valores binários, pode ser utilizado o código termômetro. O aumento dos valores de assemelha ao aumento de temperatura em um termômetro analógico.
  - ✓ Exemplo do atributo ranking.

Ranking
Primeiro
Terceiro
Segundo
Primeiro
Segundo
Quarto
Segundo
Quarto

Ranking
Primeiro
Segundo
Terceiro
Quarto

Ranking
000**1**
00**11**
0**111**
1111

A distância entre um valor e outro varia de acordo com a posição. Portanto, isso evidencia a preservação da ordem em relação aos dados originais