

Tópicos Especiais em Desenvolvimento de Software 1 – IMD0179

Atividade Prática – Ciência de Dados

Pré-Processamento

Conjunto de Dados

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

Atributos: **Id** (identificação do paciente, **Nome**, **Idade**, **Sexo** (Gênero), **Peso**, **Manchas** (presença e distribuição de manchas no corpo), **Temp.** (temperatura do corpo), **#Int.** (número de internações, **Est.** (estado de origem) e **Diagnóstico** (classe).

FACELI K, et. al., 2011

Eliminação Manual de Atributos

- Há atributos que não contribuem para o aprendizado.

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

**Não contribuem
para estimar se um
paciente está
doente ou não**



Eliminação Manual de Atributos

- ❑ Conjunto é definido de acordo com a experiência do especialista (médico nesse caso).

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
28	M	79	Concentradas	38,0	2	SP	Doente
18	F	67	Inexistentes	39,5	4	MG	Doente
49	M	92	Espalhadas	38,0	2	RS	Saudável
18	M	43	Inexistentes	38,5	8	MG	Doente
21	F	52	Uniformes	37,6	1	PE	Saudável
22	F	72	Inexistentes	38,0	3	RJ	Doente
19	F	87	Espalhadas	39,0	6	AM	Doente
34	M	67	Uniformes	38,4	2	GO	Saudável

**Esse atributo
é realmente
relevante?**



Eliminação Manual de Atributos

❑ Conjunto de dados (hospital):

❖ Após a eliminação manual dos atributos.

classe

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
28	M	79	Concentradas	38,0	2	Doente
18	F	67	Inexistentes	39,5	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	M	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
19	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável

Atributos Qualitativos

- Também chamados de “**categórico**” ou “**discreto**”:
 - ❖ Porém: “discreto” implica em uma ordem.

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
28	M	79	Concentradas	38,0	2	Doente
18	F	67	Inexistentes	39,5	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	M	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
19	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável

Atributos Quantitativos

□ Também chamados de “numérico” ou “contínuo”.

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
28	M	79	Concentradas	38,0	2	Doente
18	F	67	Inexistentes	39,5	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	M	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
19	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável

Atributos Quantitativos

□ Também chamados de “numérico” ou “contínuo”.

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
28	M	79	Concentradas	38,0	2	Doente
18	F	67	Inexistentes	39,5	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	M	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
19	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável

Por que fazer pré-processamento?

Essa é
realmente uma
boa
pergunta!



Por que fazer pré-processamento?

- ❑ Os dados possuem defeitos:
 - ❖ Ausência de valores (*missing values*).

Idade	Sexo	Peso	Manchas	Temp.	#Int.	Diagnóstico
??	M	79	??	38,0	??	Doente
18	F	67	Inexistentes	39,5	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	??	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
??	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável

Por que fazer pré-processamento?

- ❑ Os dados possuem defeitos:
 - ❖ Valores com erro ou fora da realidade

Idade	Sexo	Peso	Manchas	Temp.	#Int.	Diagnóstico
10	M	120	Concentradas	38,0	2	Doente
18	F	67	Inexistentes	39,5	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	M	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
19	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável

Por que fazer pré-processamento?

- ❑ Os dados possuem defeitos:
 - ❖ Valores discrepantes (*outliers*).

Idade	Sexo	Peso	Manchas	Temp.	#Int.	Diagnóstico
28	M	79	Concentradas	38,0	2	Doente
18	F	67	Inexistentes	41,7	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	M	243	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
109	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável

Etapas do pré-processamento

- ❑ Limpeza dos dados:
 - ❖ Preencher dados ausentes, “alisar” ruído, identificar e/ou remover valores aberrantes, resolver inconsistências.

- ❑ Transformação de dados:
 - ❖ Transformação de tipos;
 - ❖ Binarização;
 - ❖ Normalização.

- ❑ Redução de Dados:
 - ❖ Redução no volume de dados (instâncias e atributos).

Limpeza dos Dados

Idade	Sexo	Peso	Manchas	Temp.	#Int.	Diagnóstico
??	M	79	??	38,0	??	Doente
18	F	67	Inexistentes	39,5	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	??	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
??	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável

- + / - 40% de perda;
- desbalanceamento.

Limpeza dos Dados

❑ Método de atribuição.

❖ Estimar valores perdidos com base em valores válidos do mesmo atributo.

- Substituição pela **média**;
- Substituição pela **mediana**; ou
- Substituição pela **moda**.

❖ Válido apenas para atributos numéricos e categóricos.

Limpeza dos Dados

❑ Substituição com Mediana:

Idade	Sexo	Peso	Manchas	Temp.	#Int.	Diagnóstico
??	M	79	??	38,0	??	Doente
18	F	67	Inexistentes	39,5	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	??	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
??	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável



Vetor ordenado							
0	0	18	18	21	22	34	49
				19,5			

Limpeza dos Dados

❑ Substituição com Mediana:

Idade	Sexo	Peso	Manchas	Temp.	#Int.	Diagnóstico
19	M	79	??	38,0	??	Doente
18	F	67	Inexistentes	39,5	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	??	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
19	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável

Limpeza dos Dados

❑ Substituição com Média:

Idade	Sexo	Peso	Manchas	Temp.	#Int.	Diagnóstico
??	M	79	??	38,0	??	Doente
18	F	67	Inexistentes	39,5	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	??	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
??	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável



0	18	49	18	21	22	0	34
			20,25				

Limpeza dos Dados

❑ Substituição com Média:

Idade	Sexo	Peso	Manchas	Temp.	#Int.	Diagnóstico
20	M	79	??	38,0	??	Doente
18	F	67	Inexistentes	39,5	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	??	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
20	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável

Limpeza dos Dados

❑ Substituição com Moda:

Idade	Sexo	Peso	Manchas	Temp.	#Int.	Diagnóstico
28	M	79	??	38,0	2	Doente
18	F	67	Inexistentes	39,5	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	M	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
19	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável

Limpeza dos Dados

❑ Substituição com Moda:

Inexistentes	Espalhadas	Uniformes
3	2	2

Idade	Sexo	Peso	Manchas	Temp.	#Int.	Diagnóstico
28	M	79	??	38,0	2	Doente
18	F	67	Inexistentes	39,5	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	M	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
19	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável

Limpeza dos Dados

❑ Substituição com Moda:

Idade	Sexo	Peso	Manchas	Temp.	#Int.	Diagnóstico
28	M	79	Inexistentes	38,0	2	Doente
18	F	67	Inexistentes	39,5	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	M	43	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
19	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável

Limpeza dos Dados

- ❑ Detecção de ruído e/ou valores aberrantes.
- ❑ Técnicas para **identificar** valores **ruidosos** ou **aberrantes**:
 - ❖ Amplitude Interquartil;
 - ❖ Regressão linear.

Limpeza dos Dados

- ❑ Detecção de ruído e/ou valores aberrantes:

Idade	Sexo	Peso	Manchas	Temp.	#Int.	Diagnóstico
28	M	39	Inexistentes	38,0	2	Doente
18	F	67	Inexistentes	41,7	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	M	243	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
109	F	87	Espalhadas	39,0	6	Doente
34	M	??	Uniformes	38,4	2	Saudável

Limpeza dos Dados

□ Amplitude Interquartil:

Idade	Sexo	Peso	Manchas	Temp.	#Int.	Diagnóstico
28	M	39	Inexistentes	38,0	2	Doente
18	F	67	Inexistentes	41,7	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	M	243	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
109	F	87	Espalhadas	39,0	6	Doente
34	M	0	Uniformes	38,4	2	Saudável

Limpeza dos Dados

□ Amplitude Interquartil:

		M = $(n/2) + [(n+1)/2] \Rightarrow 4+5$		
		Q1 = $(n + 1)/4 \Rightarrow 2,25$		
		Q3 = $[3x(n + 1)]/4 \Rightarrow 6,75$		
		Inferior = $[Q1 - (Inter * 1,5)]$		
		Superior = $[Q3 + (Inter * 1,5)]$		
1	0			
2	39	50,5	Q1	
3	62			
4	67	69,5	M	
5	72			
6	77	82,5	Q3	
7	88			
8	243	32	InterQuartil	
		Inferior	2,5	
		Superior	130,5	

Limpeza dos Dados

□ Amplitude Interquartil:

		$M = (n/2) + [(n+1)/2] \Rightarrow 4+5$		
		$Q1 = (n + 1)/4 \Rightarrow 2,25$		
		$Q3 = [3x(n + 1)]/4 \Rightarrow 6,75$		
		Inferior = $[Q1 - (Inter * 1,5)]$		
		Superior = $[Q3 + (Inter * 1,5)]$		
1	→ 0			
2	39	50,5	Q1	
3	62			
4	67	69,5	M	
5	72			
6	77	82,5	Q3	
7	88			
8	→ 243	32	InterQuartil	
		Inferior	2,5	
		Superior	130,5	

Limpeza dos Dados

❑ Atributo limpo (corrigido):

Idade	Sexo	Peso	Manchas	Temp.	#Int.	Diagnóstico
28	M	39	Inexistentes	38,0	2	Doente
18	F	67	Inexistentes	41,7	4	Doente
49	M	92	Espalhadas	38,0	2	Saudável
18	M	69	Inexistentes	38,5	8	Doente
21	F	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
109	F	87	Espalhadas	39,0	6	Doente
34	M	69	Uniformes	38,4	2	Saudável

Transformação de Dados

❑ Normalização ou mudança de escala:

❖ Propósito da normalização: minimizar os problemas oriundos do **uso de unidades** e **dispersões distintas** entre as variáveis.

❖ As variáveis podem ser normalizadas segundo a amplitude ou segundo a distribuição.

$$Att_1 = \left(\frac{x_i - \min}{\max - \min} \right)$$



Ano	Vlr (norm.)
1900	0,0
1914	0,1
1950	0,4
1981	0,7
1999	0,9
2005	0,9
2014	1,0

Transformação de Dados

- ❑ Transformação de Numérico para Ordinal:
 - ❖ Discretização



Transformação de Dados

❑ Transformação de Nominal para Numérico:

❖ *Atributos ordinais* (e.g., `grau_de_satisfação` com um produto) podem ser convertidos para números preservando a ordem natural.

- Muito Satisfeito \Rightarrow 0.8
- Satisfeito \Rightarrow 0.6
- Pouco Satisfeito \Rightarrow 0.4
- Insatisfeito \Rightarrow 0.2

❖ Por que é importante preservar a ordem natural?

- Para permitir comparações que façam sentido:
`grau_de_satisfação > 0.4`

Dúvidas ...



Limpeza e Transformação com PDI

Prática PDI – Ciência de Dados

□ Fonte:

Pregnancies	Preg_Ordinal	Plasma	Blood pressure	Blood Sugar	A1C	Height	Weight	BMI	age	Age_Ordinal
6		148	72	120		1,55	90,30		50	
1		85	66	205		1,65	75,50		31	
8		183	64	290		1,55	202,30		32	
1		89	66	194		1,87	74,80		21	
0		137	40	168		1,54	77,10		33	
5		116	174			1,77	65,80		30	
3			50	188		1,55	69,40		26	
2		197	70	143		1,92	61,50		53	
8		125	96	151		1,39	66,80		54	
4		110	92	100		1,78	55,40		30	
1		189	60	346		1,81	63,53		59	
5		166	72	175		1,83	95,84		51	
7		100		155		1,80	12,48		32	
0		118	84	230		1,34	63,58		31	
7		107	74	265		2,04	55,85		31	
1		103	30	183		1,82	58,45		33	
1		115	70	196		1,71	51,27		32	
3		126	88	235		1,96			27	
8		99	84	156		1,45	93,97		50	
7		196	90	149		1,79	81,20		41	
9		119	80	130		1,65	66,58		29	
7		147	76	147			89,47		43	
1		97	66	140		1,80	52,32		22	
5		117	92	273		1,48	50,80		38	
5		109	75	147		1,46	80,77		60	

Prática PDI – Ciência de Dados

❑ Descrição dos atributos:

1. Pregnancies (número de gestações [0-9]);
2. Plasma (concentração de glicose no plasma [40-200]);
3. Blood pressure (pressão sanguínea diastólica [40-120]);
4. Blood sugar (ABS - média de açúcar no sangue em duas horas [100-300]);
5. Height (altura em centímetros [1,34-2,04]);
6. Weight (peso em quilogramas [50,0-96,0]);
7. Age (idade [20-70]).

Prática PDI – Ciência de Dados

□ Atributos a serem criados:

1. Preg_Ordinal (número de gestações expresso em classes de valores – discretização [“nenhuma”, “baixa”, “alta” e “média”]);
2. A1C (padrão clínico para medir o nível de açúcar no sangue):
 - i. $A1C = (ABS + 46,7) / 28,7$;
3. BMI (índice de massa corporal):
 1. $BMI = Weight / (Height * Height)$;
4. Age_Ordinal (idade expressa em classes [“jovem”, “adulto jovem”, “adulto” e “idoso”]).

Prática PDI – Ciência de Dados

❑ Cuidados necessários:

Pregnancies	Plasma	Blood pressure	Blood Sugar	Height	Weight	age
6	148	72	120	1,55	90,30	50
1	85	66	205	1,65	75,50	31
8	183	64	290	1,55	202,30	32
1	89	66	194	1,87	74,80	21
0	137	40	168	1,54	77,10	33
5	116	174	?	1,77	65,80	30
3	?	50	188	1,55	69,40	26
2	197	70	143	1,92	61,50	53
8	125	96	151	1,39	66,80	54
4	110	92	100	1,78	55,40	30
1	189	60	346	1,81	63,53	59
5	166	72	175	1,83	95,84	51
7	100	?	155	1,80	12,48	32
0	118	84	230	1,34	63,58	31
7	107	74	265	2,04	55,85	31
1	103	30	183	1,82	58,45	33
1	115	70	196	1,71	51,27	32
3	126	88	235	1,96	?	27
8	99	84	156	1,45	93,97	50
7	196	90	149	1,79	81,20	41
9	119	80	130	1,65	66,58	29
7	147	76	147	?	89,47	43
1	97	66	140	1,80	52,32	22

Prática PDI – Ciência de Dados




❑ Cuidados necessários:

Pregnancies	Plasma	Blood pressure	Blood Sugar	Height	Weight	age
6	148	72	120	1,55	90,30	50
1	85	66	205	1,65	75,50	31
8	183	64	290	1,55	202,30	32
1	89	66	194	1,87	74,80	21
0	137	40	168	1,54	77,10	33
5	116	174		1,77	65,80	30
3		50	188	1,55	69,40	26
2	197	70	143	1,92	61,50	53
8	125	96	151	1,39	66,80	54
4	110	92	100	1,78	55,40	30
1	189	60	346	1,81	63,53	59
5	166	72	175	1,83	95,84	51
7	100		155	1,80	12,48	32
0	118	84	230	1,34	63,58	31
7	107	74	265	2.04	55,85	31
1	103	30	183	1,82	58,45	33
1	115	70	196	1,71	51,27	32
3	126	88	235	1,96		27
8	99	84	156	1,45	93,97	50
7	196	90	149	1,79	81,20	41
9	119	80	130	1,65	66,58	29
7	147	76	147		89,47	43
1	97	66	140	1,80	52,32	22

Prática PDI – Ciência de Dados

□ Tarefa:

TAREFAS INDIVIDUAIS

Título	Período de Entrega	Envios
Mercado Imobiliário - WS Implemente um web scraper (robot) em qualquer linguagem para extrair os dados de apartamentos que estão a venda na cidade de Natal. Seu código deverá gerar um arquivo CSV como resultado da extração. Submeta código, arquivo CSV e mais um arquivo TXT com os nomes dos componentes do grupo, caso você deseje trabalhar em dupla.	de 15/01/2021 às 10h00 a 24/01/2021 às 23h59	15   
Limpeza e Transformação com PDI Baixe o arquivo Diabetes.xls contendo onze atributos. Faça a limpeza e transformação necessárias nos atributos como foi demonstrado nos slides que explicam a referida tarefa. Submeta o arquivo resultante (csv ou xls), o arquivo de transformação (*.ktr), e mais um arquivo txt contendo os nomes dos componentes do grupo, caso você esteja trabalhando em dupla.	de 05/02/2021 às 08h00 a 14/02/2021 às 23h59	0 