Fotos de Kelly Sikkema, disponível na Unsplash. Editada pelo autor.

ELDORADO

Trilha Python

Aula 10

TIC em trilhas

Chamada!

TIC em trilhas

Faísca

TIC em trilhas

Atividades de hoje

- **Introdução ao pré-processamento**
- **Técnicas de Pré-processamento**
- Exemplos práticos sem introduzir modelos.

Introdução

Ao Pré-Processamento

TIC em trilhas

Para o que serve?

- Minimizam ou eliminam problemas existentes em um conjunto de dados
- Podem tornar os dados mais adequados para sua utilização por um determinado algoritmo de AM
- Limpeza dos dados para remoção de ruídos
- Fusão dos dados de múltiplas fontes
- Observações de duplicadas
- Observação de características importantes a tarefa de mineração de dados

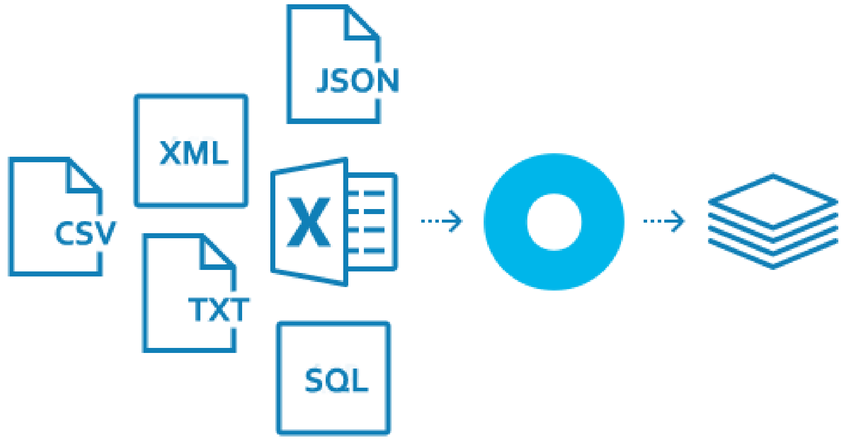
Eliminação Manual de Atributos

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	Claudia	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

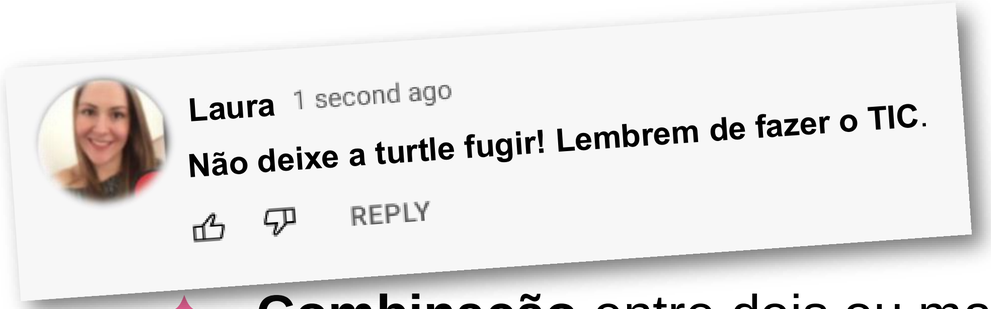
Quais **atributos** são **irrelevantes** pra um algoritmo de Aprendizado de Máquina determinar o diagnóstico dos pacientes?

Integração de Dados

- Realizado quando os dados a serem utilizados em uma aplicação de AM estão distribuídos em **diferentes conjuntos de dados,** (como bancos de dados distintos, planilhas, APIs etc.).

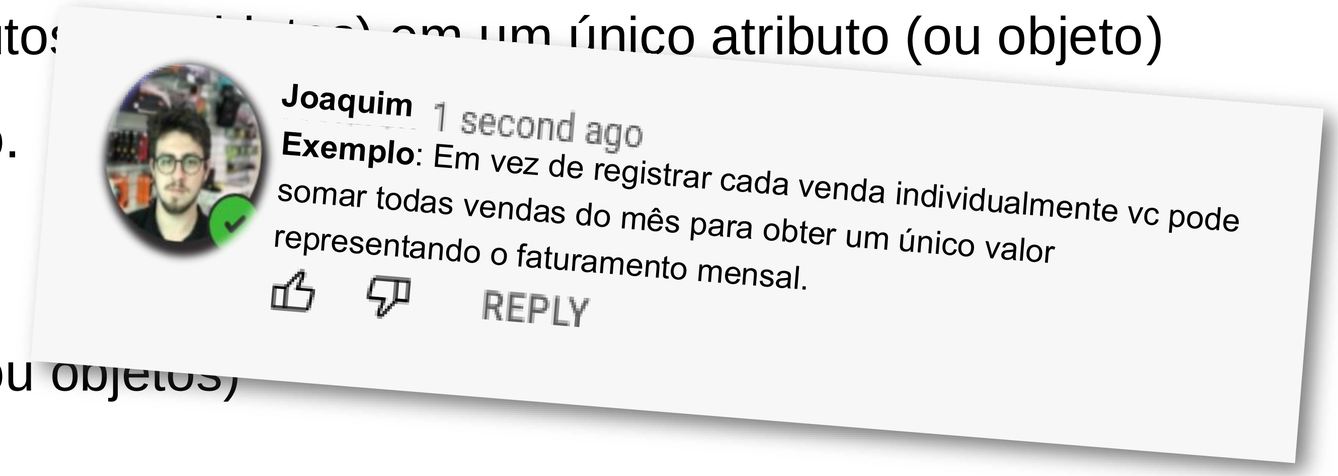


- Necessário identificar quais são os objetos que estão presentes nos diferentes conjuntos a serem combinados. Identificar as **chaves** ou **colunas** em **comum** entre os datasets.



Agregação de Dados

- **Combinação** entre dois ou mais atributos [illegible] em um único atributo (ou objeto) para simplificar e resumir a informação.



- **Objetivo**

  Redução dos dados (atributos ou objetos)

  Troca de escala

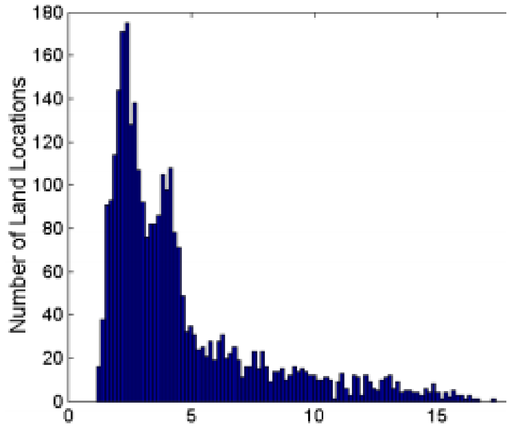
    Cidades -> agregadas em regiões, estados, país, etc...

    Dias -> agregados em semanas, meses ou anos.

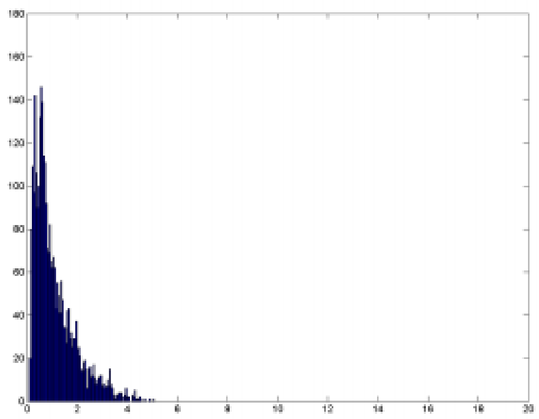
  Dados mais estáveis

    Dados agregados tendem a ter menos variabilidade.

Agregação de Dados: **Exemplo**



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

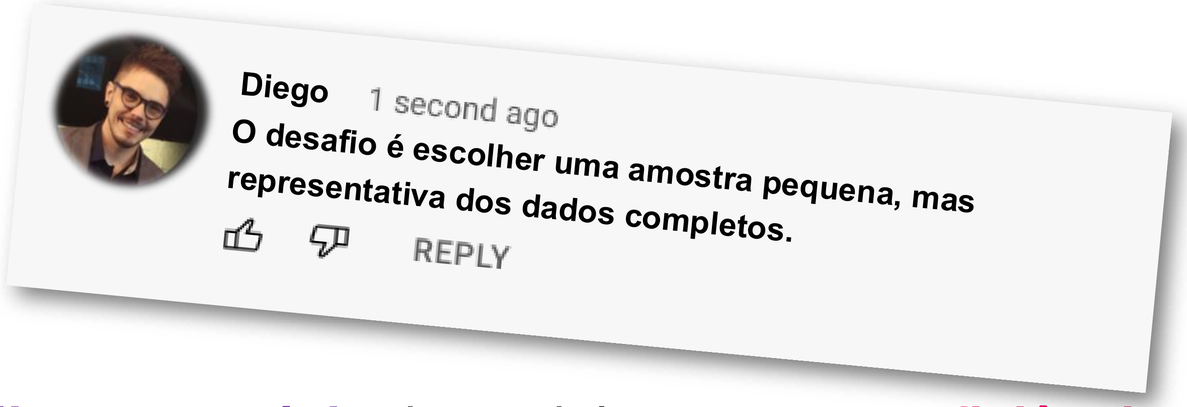
- Variação da Precipitação de chuvas na Austrália
- **Histograma** apresenta todas as medidas de precipitações em centímetros.
- A precipitação **média anual** varia menos do que a precipitação **média mensal**.

Amostragem

De Dados

TIC em trilhas

Amostragem de Dados

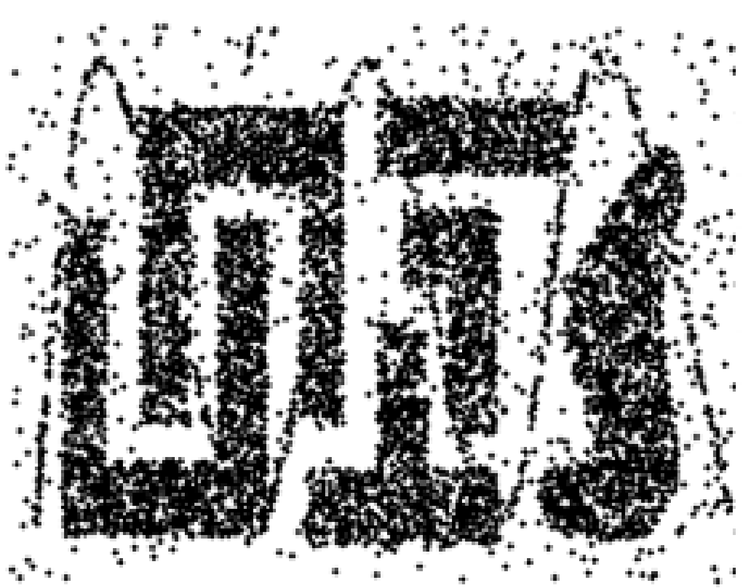


- Quanto **maior** a quantidade de dados, **melhor a acurácia** do modelo e **menor a eficiência computacional** do processo indutivo.

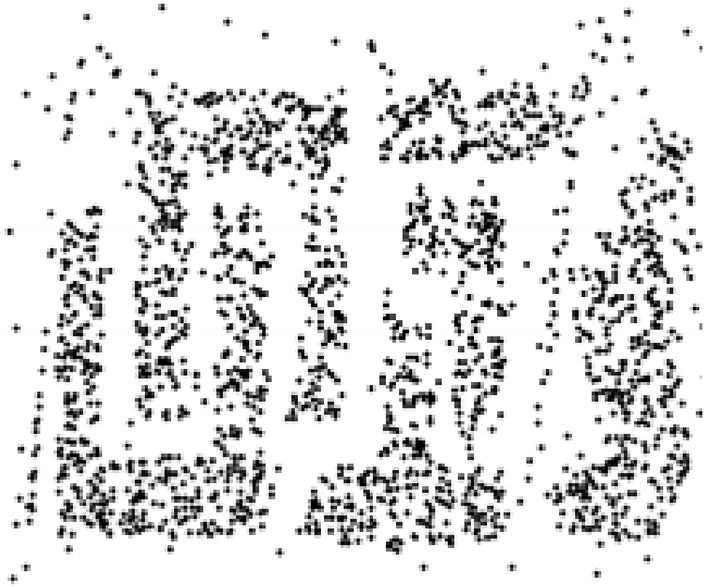
Muitos algoritmos de AM podem apresentam saturação de memória (falhar) quando um conjunto de dados tem um grande número de instâncias.

- A amostra deve ser representativa do conjunto de dados original
- Deve obedecer a mesma distribuição estatística que gerou o conjunto de dados original.

Amostragem de Dados



8000 points



2000 Points

500 Points

Amostragem de Dados

Existem **3 abordagens** para garantir a amostragem estatística dos dados:

Amostragem aleatória simples

- Mesma probabilidade de selecionar qualquer item
- Sem reposição: O item não volta ao conjunto após ser selecionado.
- Com reposição: O mesmo item pode ser escolhido várias vezes.

Amostragem com Reposição

Bagging (Bootstrap Aggregating):

- Utiliza amostragem com reposição para criar múltiplos subconjuntos de dados. Pode ser usado em Random Forests. Aumenta a diversidade entre os modelos, reduz overfitting e melhora a generalização.

Bootstrap

- Técnica estatística para estimar a performance de modelos. Treina o modelo em subconjuntos gerados com reposição. Fornece uma estimativa robusta do erro do modelo.

Amostragem sem Reposição

Boosting:

- Utiliza amostragem sem reposição para focar em exemplos mal classificados. Por Exemplo em AdaBoost. Melhora a performance ao focar em exemplos difíceis.

Validação Cruzada (Cross-Validation):

- Divide os dados em folds distintos (sem reposição). Avalia a performance do modelo com precisão em diferentes subconjuntos.

Amostragem de Dados

Amostragem estratificada

- Usada em conjunto de dados desbalanceados
- Manter o mesmo número de instâncias para cada classe.
- Exemplo: Garantir que cada categoria (gênero, classe social, etc.) esteja proporcionalmente representada.

Amostragem progressiva

- Começa com uma amostra pequena e aumenta gradativamente o tamanho da amostra, enquanto a acurácia preditiva continuar a melhorar(até obter uma acurácia estável).

Balanceamento

De Dados

TIC em trilhas

Dados Desbalanceados

- Ocorrem quando as categorias (classes) não têm a mesma quantidade de exemplos.
- Exemplo: Em um dataset de classificação, 95% das instâncias pertencem à Classe A e apenas 5% à Classe B.

Principais técnicas utilizadas:

- Redefinir o tamanho do conjunto de dados

P. ex: imputação de dados

- Induzir um modelo para uma classe.

Como Lidar?

Imputação de dados:

- Utilizada para completar valores ausentes e reduzir o impacto de dados faltantes, especialmente na classe minoritária. Isso ajuda o modelo a lidar com inconsistências.

Induzir um modelo para uma classe:

- Em alguns casos, o foco pode ser apenas na classe minoritária, como em sistemas de detecção de fraudes (onde o interesse é identificar transações fraudulentas, que são raras).

Como Lidar ?

Redefinir o tamanho do conjunto de dados: Ajustar a quantidade de dados das classes para evitar desbalanceamento.

- **Oversampling** (superamostragem): Aumentar o número de instâncias da classe minoritária, duplicando exemplos ou gerando novos exemplos sintéticos (como o SMOTE).
- **Undersampling** (subamostragem): Reduzir o número de instâncias da classe majoritária.

Principais problemas

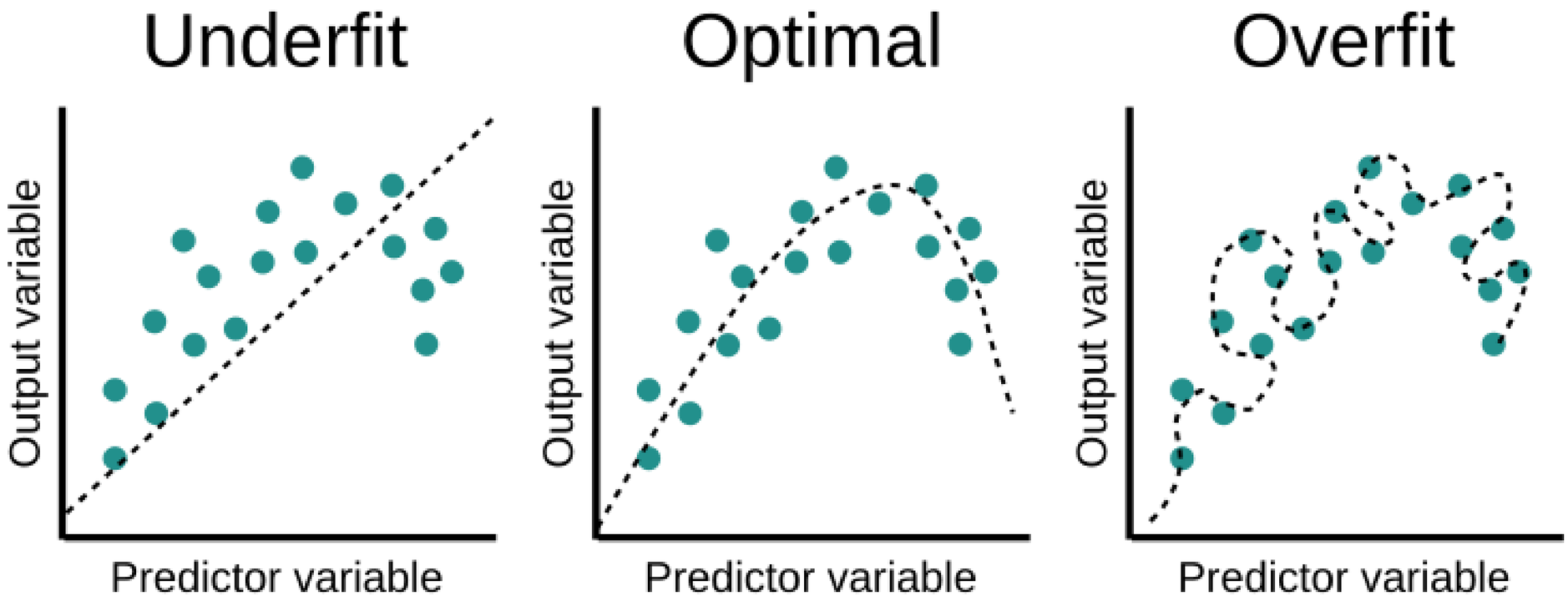
Overfitting (Superajuste):

- O modelo fica muito especializado nos dados de treino e não generaliza bem. Isso pode acontecer se o modelo memoriza o comportamento dos exemplos em vez de aprender padrões generalizáveis.

Underfitting (Subajuste):

- Ocorre quando o modelo não consegue ajustar-se bem aos dados, seja porque ele não capturou a complexidade dos dados de treino ou porque recebeu dados desbalanceados e não aprendeu padrões relevantes.

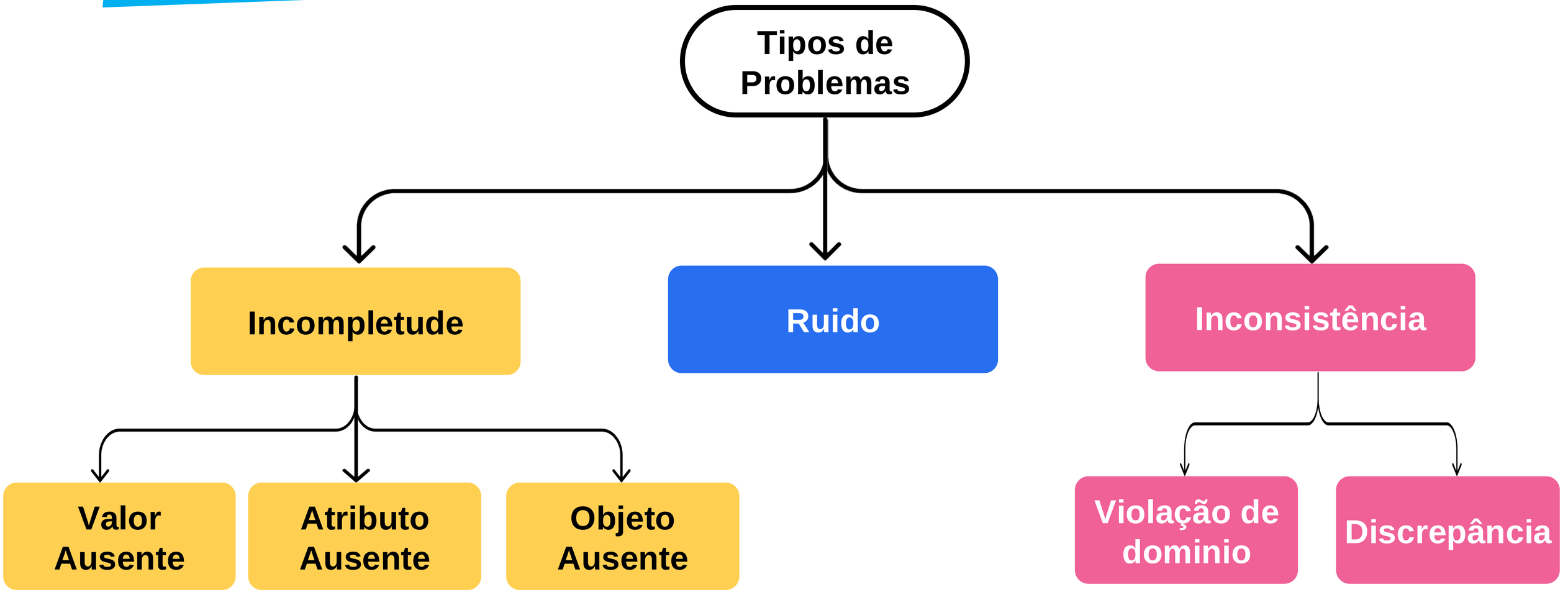
Overfit vs. Underfit



Limpeza De Dados

TIC em trilhas

Principais Problemas



Dados Incompletos

- Falta de valores de atributos, certos atributos de interesse ou contendo apenas dados agregados

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

Dados Incompletos

Alternativas utilizadas para tratar dados incompletos

- Eliminar as instâncias com valores ausentes. Geralmente empregada quando o atributo ausente é o que indica a classe
- Definir e preencher manualmente valores para os atributos com valores ausentes
- Empregar algoritmos de AM que lidam com valores ausentes

Dados Incompletos

Alternativas utilizadas para tratar dados incompletos

- Utilizar algum método para definir automaticamente ausentes, seguindo as seguintes abordagens:

  Criar um novo valor que indique que o atributo possuía um valor desconhecido.

  Problema: o algoritmo de AM pode assumir que o valor desconhecido representa um conceito importante.

  Utilizar a média, moda ou mediana dos valores conhecidos..

  Empregar um indutor para estimar o valor do atributo

Dados com Ruídos

- Dados incorretos ou que contenham **outliers**

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	F	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	**567**	Uniformes	38,4	2	Saudável

Dados com ruídos

Técnicas utilizadas para tratar dados com ruídos

- Baseadas em agrupamento

  Detectar outliers em atributos

- Baseadas em distâncias

  Identificar borderlines: mesmo em quantidade pequena pode movê-los para o lado incorreto da fronteira

- Baseadas em regressão (valor contínuo) ou classificação (valor categórico)

  A partir de um AM predizer o valor com ruído

Dados com ruídos

Técnicas utilizadas para tratar dados com ruídos

- Encestamento

  Valores são divididos em faixas ou cestas, cada uma com o mesmo número de valores

  Os valores são substituídos pela média ou mediana dos valores presentes

Dados com Ruídos

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	F	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
22	**F**	**72**	**Inexistentes**	**38,0**	**3**	**Saudável**

- Dados contendo discrepâncias e que possuem valores conflitantes em seus atributos

Dados Redundantes

- Problemas na coleta, na entrada, no armazenamento, na integração ou na transmissão de dados

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	F	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
22	**F**	**72**	**Inexistentes**	**38,0**	**3**	**Doente**

Conversões/

Transformação de Dados

TIC em trilhas

Conversões/Transformação de Dados

Muitos algoritmos de AM trabalham apenas com variáveis numéricas

- Redes Neurais, SVM, etc.
- Variáveis categóricas precisam ser convertidas
- Conversão depende da existência de ordem
- Variáveis são nominais ou ordinais?

Conversões/Transformação de Dados

Possíveis conversões:

- Conversão de valores categóricos para numéricos
- Conversão de valores numéricos para categóricos (discretização)
- Normalização de valores numéricos

Conversão de Valores Ordinais (Encoding)

Para variáveis ordinais, a ordem dos valores deve ser mantida de alguma maneira

- Estratégia comum: associar valores inteiros crescentes

  Ex: {frio, morno, quente} = {1,2,3}

- Tal estratégia pode inserir distorções relativas entre os conceitos (qualquer política de peso também insere!)
- Diferenças entre símbolos são subjetivas

Conversão de Valores Nominais

- Conversão é feita por binarização
- **Codificação inteira-binária**
- cada valor é provisoriamente convertido para inteiro e, em seguida, para binário

Valor Nominal	Valor Inteiro	A1	A2	A3
amarelo	0	0	0	0
vermelho	1	0	0	1
verde	2	0	1	0
azul	3	0	1	1
branco	4	1	0	0

Conversão de Valores Nominais

Codificação inteira-binária

- **Vantagem**: codificação que demanda menor número de atributos binários
- **Desvantagens**:

Diferença entre valores não é a mesma (nem segundo a representação binária, nem segundo a inteira). Ver (azul x branco) e (azul x vermelho)

Introduz correlação entre atributos. Péssimo para vários algoritmos de AM

Valor Nominal	Valor Inteiro	A1	A2	A3
amarelo	0	0	0	0
vermelho	1	0	0	1
verde	2	0	1	0
azul	3	0	1	1
branco	4	1	0	0

Conversão de Valores Nominais

Codificação 1-de-n

- Um atributo binário associado a cada valor nominal
- Conhecida como one-hot ou dummy encoding.
- **Exemplo**: Codificar {amarelo, vermelho, verde, azul, laranja, branco}

100000 - amarelo
010000 - vermelho
001000 - verde
000100 - azul
000010 - laranja
000001 - branco

Conversão de Valores Nominais

Vantagens:

- Mantém equidistantes quaisquer dois vetores binários
- Atributos binários são descorrelacionados
- Atributos binários são assimétricos (obrigatório no caso de alguns algoritmos de AM)
- Moda do atributo nominal corresponde ao atributo binário com maior números de 1s

Conversão de Valores Nominais

Desvantagens:

- Pode gerar um número enorme de atributos
- Pode levar a dados muito esparsos quando n é grande

Ex: atributo = nome de país

Existem 193 países membros da ONU

Codificação 1-de-n demandaria 192 atributos adicionais

Dados esparsos/maldição da dimensionalidade!

Conversão de Valores Nominais

Desvantagens:

- Possível solução:

Utilizar 1 atributo nominal com apenas 7 valores (continentes)

Tentar discriminar entre os países com um conjunto menor de pseudo-atributos numéricos

Funcionamento satisfatório: depende da aplicação

Conversão de Valores Numéricos

Aplicado a uma parcela dos algoritmos de classificação e de associação

- Atributos quantitativos

  Se for do tipo discreto e binário, com dois valores, associa-se um nome a cada valor.

  Se for formado por sequências binárias sem uma relação de ordem entre si, substitui-se por um nome ou categoria

Demais casos:

- Técnicas de discretização

Trilha de Análise de dados

Módulo 1: Terminar até dia 15/10

Módulo 2: Terminar até dia 22/10

Módulo 3: Terminar até dia 29/10

TIC em trilhas

Trilha de Intro Inteligência Artificial

Módulo 1: Terminar até dia 25/10

Módulo 2: Terminar até dia 01/11

Módulo 3: Terminar até dia 08/11

Módulo 4: Terminar até dia 15/11

Módulo 5: Terminar até dia 22/11

TIC em trilhas

Trilha de Intro Machine Learning

Módulo 1: Terminar até dia

Módulo 2: Terminar até dia

Módulo 3: Terminar até dia

Módulo 4: terminar até dia

TIC em trilhas

Até a próxima aula...

18/10/24

TIC em trilhas