



Capacitação em Inteligência Artificial e Aplicações

Classificadores Supervisionados: Classificação Bayesiana

- Prof. Gerson Vieira Albuquerque Neto
- Prof. Rodrigo Carvalho Souza Costa
- Prof. Yves Augusto Romero



UNIVERSIDADE
ESTADUAL DO CEARÁ



MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA
E INOVAÇÃO





IA

Planejamento da Disciplina

D	S	T	Q	Q	S	S
26	27 Introdução ao curso	28 Áreas e aplicações de IA	29 Tipos e definições de Inteligência artificial	30 Revisão de álgebra e probabilidade	31 Laboratório Python 1	1
2	3 Introdução aos classificadores supervisionados	4 Aula teórica Naive Bayes	5 Aula prática Naive Bayes + KNN	6 Feriado Semana Santa	7 Feriado Semana Santa	8
9	10 Aula Teórica de Árvore de Decisão	11 Aula Prática de Árvore de Decisão	12 Introdução à clusterização / Aula teórica Kmeans	13 Prática Kmeans	14 Introdução ao PCA / prática com classificadores já implementados	15
16	17 Introdução ao Perceptron Simples - Prática	18 Teoria MLP / Aplicação scilearn	19 Introdução ao DeepLearning	20 Uso de biblioteca DeepLearning	21 Feriado Tiradentes	28
23	24 Introdução ao TensorFlow / Keras	25 Introdução ao Pytorch	26 Tensorflow for android	27	28	29



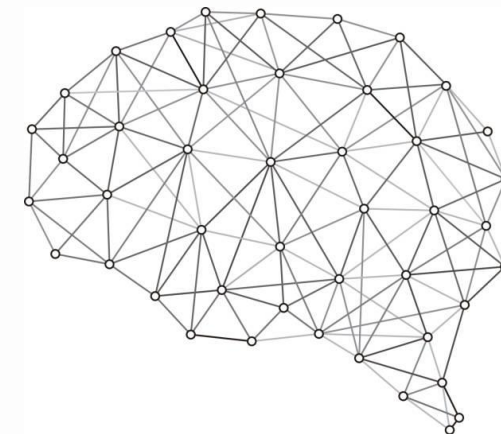
- Após a conclusão deste módulo, você será capaz de:
 - Compreender os detalhes das etapas dos algoritmos de aprendizado de máquina
 - Compreensão de algoritmos de aprendizado baseado em probabilidade.





Revisão da Aula passada

- Exemplos de técnicas de aprendizado de máquina



UNIVERSIDADE
ESTADUAL DO CEARÁ



MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA
E INOVAÇÃO



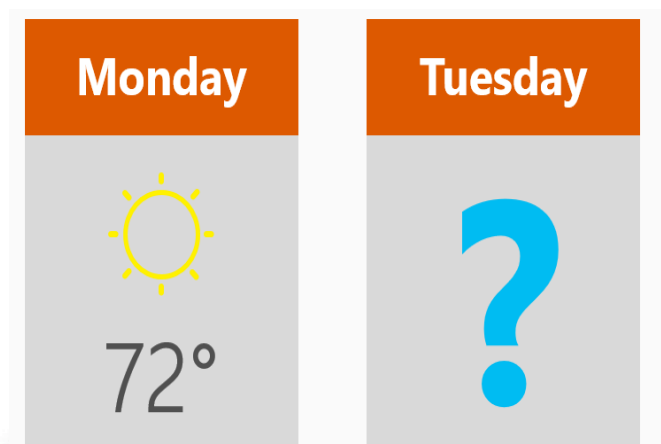


IA

Aprendizagem Supervisionada

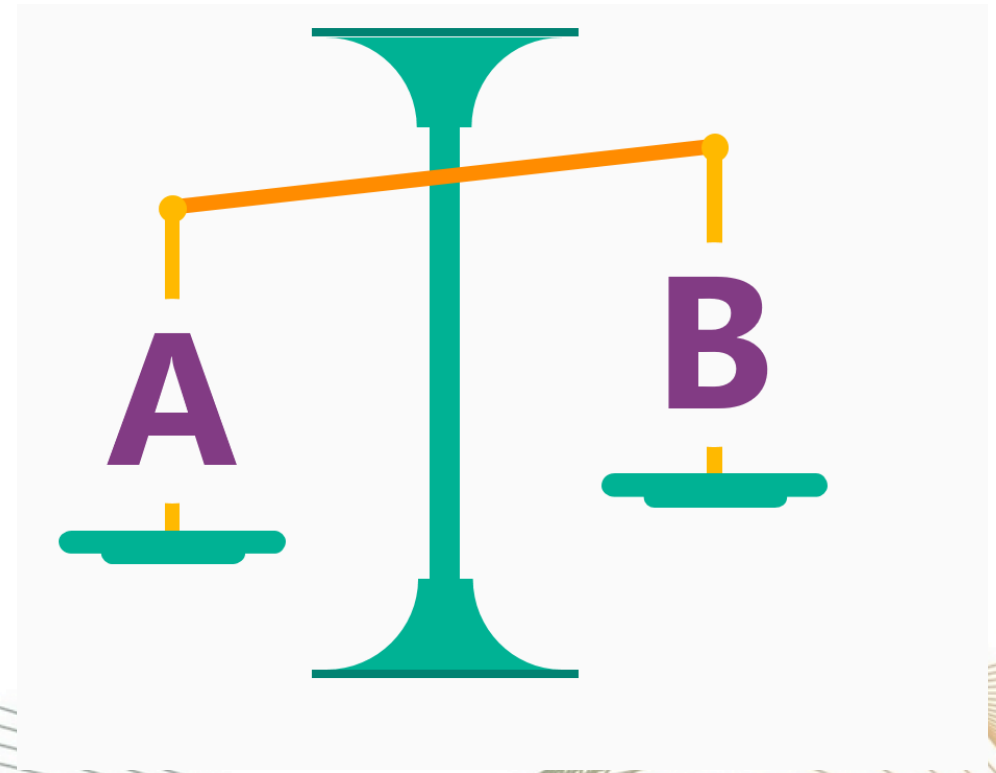
Questões de Regressão

- **Regressão:** reflete as características dos valores de atributo de amostras em um conjunto de dados de amostra. A dependência entre os valores de atributo é descoberta expressando a relação do mapeamento de amostra por meio de funções.
 - Quanto vou me beneficiar do estoque na próxima semana?
 - Qual é a temperatura nesta terça-feira?





- **Classificação:** mapeia amostras em um conjunto de dados de amostra para uma categoria especificada usando um modelo de classificação.
 - Haverá um engarrafamento na estrada XX durante a hora do rush da manhã de amanhã?
 - Qual método é mais atraente para os clientes: Voucher de 5 yuan ou 25% de desconto?



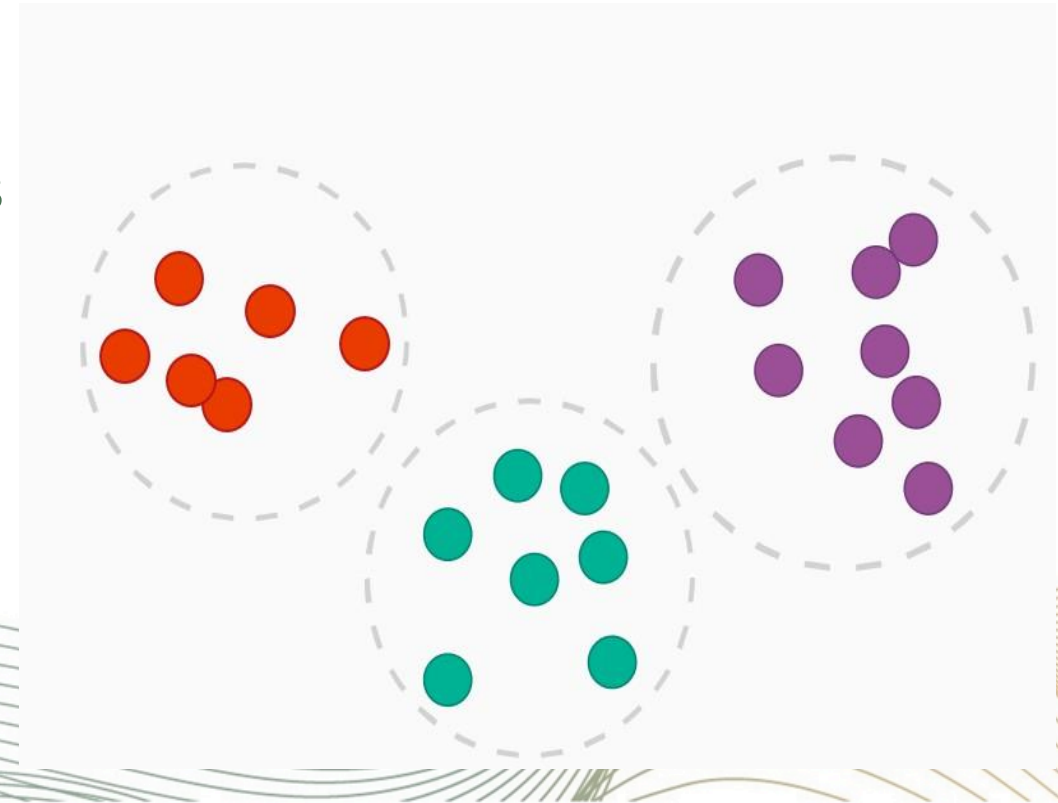


IA

Aprendizagem não-supervisionada

Questões de Clusterização

- **Clustering:** classifica amostras em um conjunto de dados de amostra em várias categorias com base no modelo de clustering. A semelhança de amostras pertencentes à mesma categoria é alta.
 - Qual público gosta de assistir filmes do mesmo assunto?
 - Quais desses componentes estão danificados de maneira semelhante?

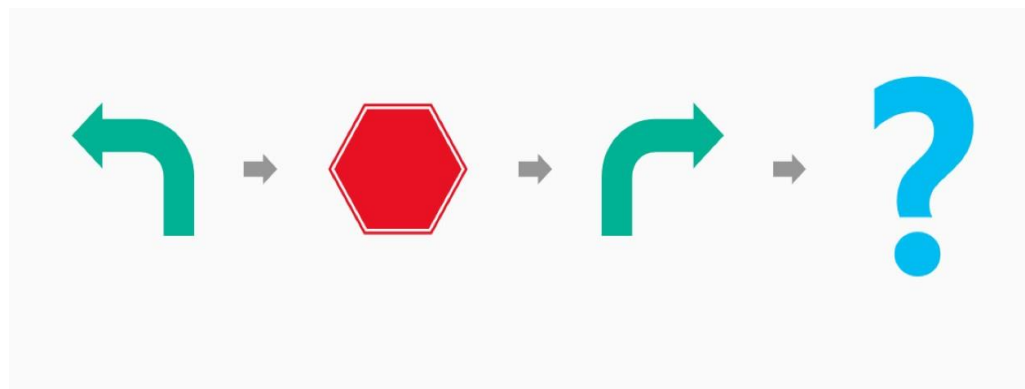


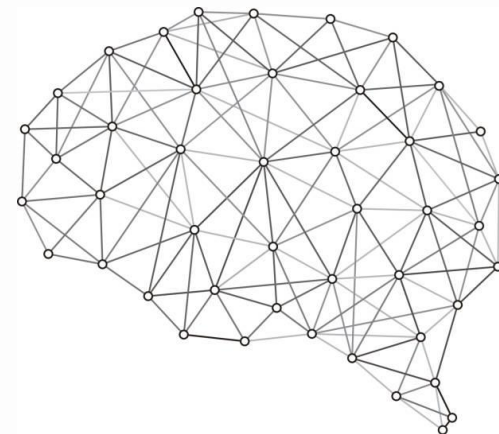


IA

Aprendizagem por reforço - Melhor Comportamento

- **Aprendizado por reforço:** sempre busca os melhores comportamentos. O aprendizado por reforço é direcionado a máquinas ou robôs.
 - Piloto automático: Deve frear ou acelerar quando a luz amarela começa a piscar?
 - Robô de limpeza: Deve continuar trabalhando ou voltar para carregar?





Classificadores Supervisionados: Classificação Bayesiana



- Etapas do Processo de Aprendizado de Máquina
- Exemplo de Aprendizado Supervisionado
- Introdução aos classificadores por probabilidade
- Classificador Bayesiano e Naive Bayes



UNIVERSIDADE
ESTADUAL DO CEARÁ



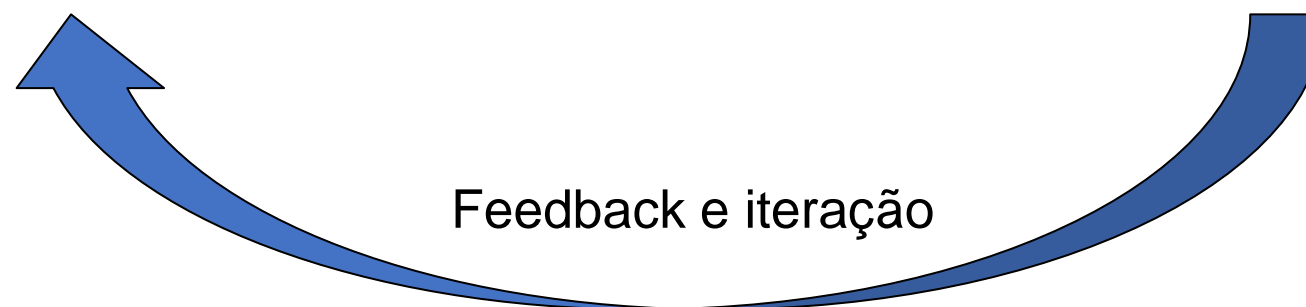
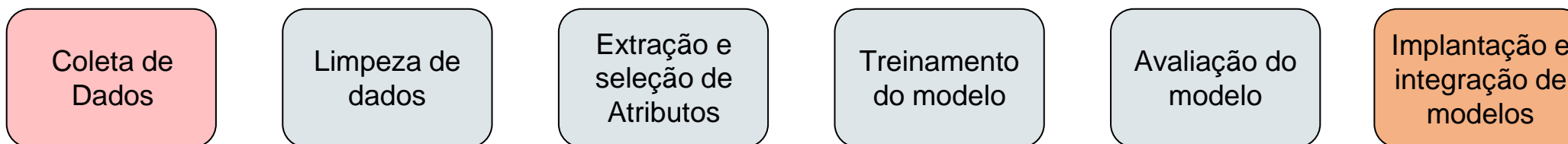
MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA
E INOVAÇÃO





IA

Processo de Aprendizado de Máquina





IA

Obtenção de Dados: Conceitos

Conjunto de Dados (Dataset)

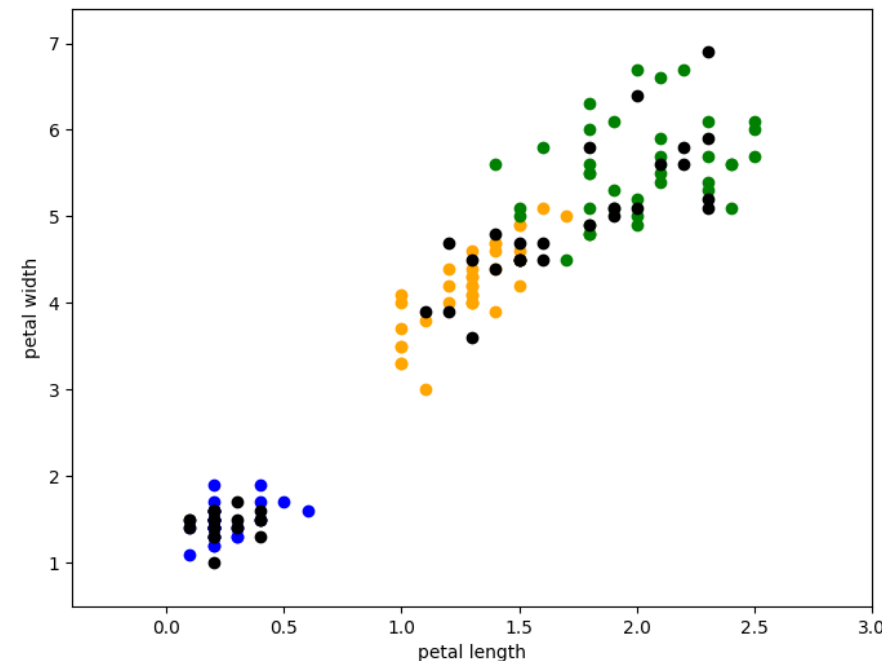
uma coleção de dados usados em tarefas de aprendizado de máquina. Cada registro de dados é chamado de amostra. Eventos ou atributos que refletem o desempenho ou a natureza de uma amostra em um aspecto específico são chamados de recursos.

Conjunto de Treinamento

um conjunto de dados utilizado no processo de formação, em que cada amostra é referida como uma amostra de formação. O processo de criação de um modelo a partir de dados é chamado de aprendizado (treinamento).

Conjunto de Testes

O teste refere-se ao processo de utilização do modelo obtido após o aprendizado para previsão. O conjunto de dados usado é chamado de conjunto de teste, e cada amostra é chamada de amostra de teste.



Dataset Testes



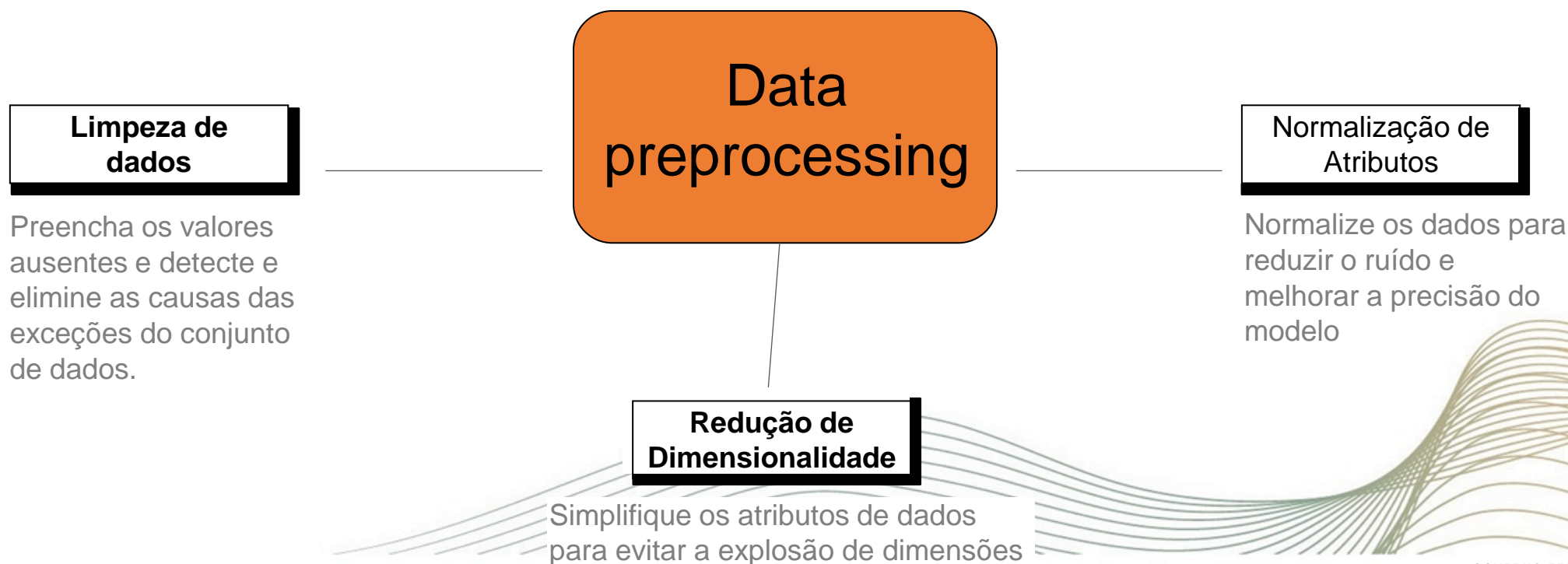
IA

Obtenção de Dados: Típico Formato de um Dataset

		Atributo 1	Atributo 2	Atributo 3	Rótulo	
		No.	Área	Distritos das Escolas	Direção	Preço da Casa
Conjunto de Treinamento	1	100	8	Sul	1000	
	2	120	9	Sudoeste	1300	
	3	60	6	Norte	700	
	4	80	9	Sudeste	1100	
Conjunto de Testes	5	95	3	Sul	850	



- Os dados são cruciais para os modelos.
- Pode se tornar o limite da capacidade de aprendizado do modelo.
- Ou seja, sem bons dados, não há um bom modelo.





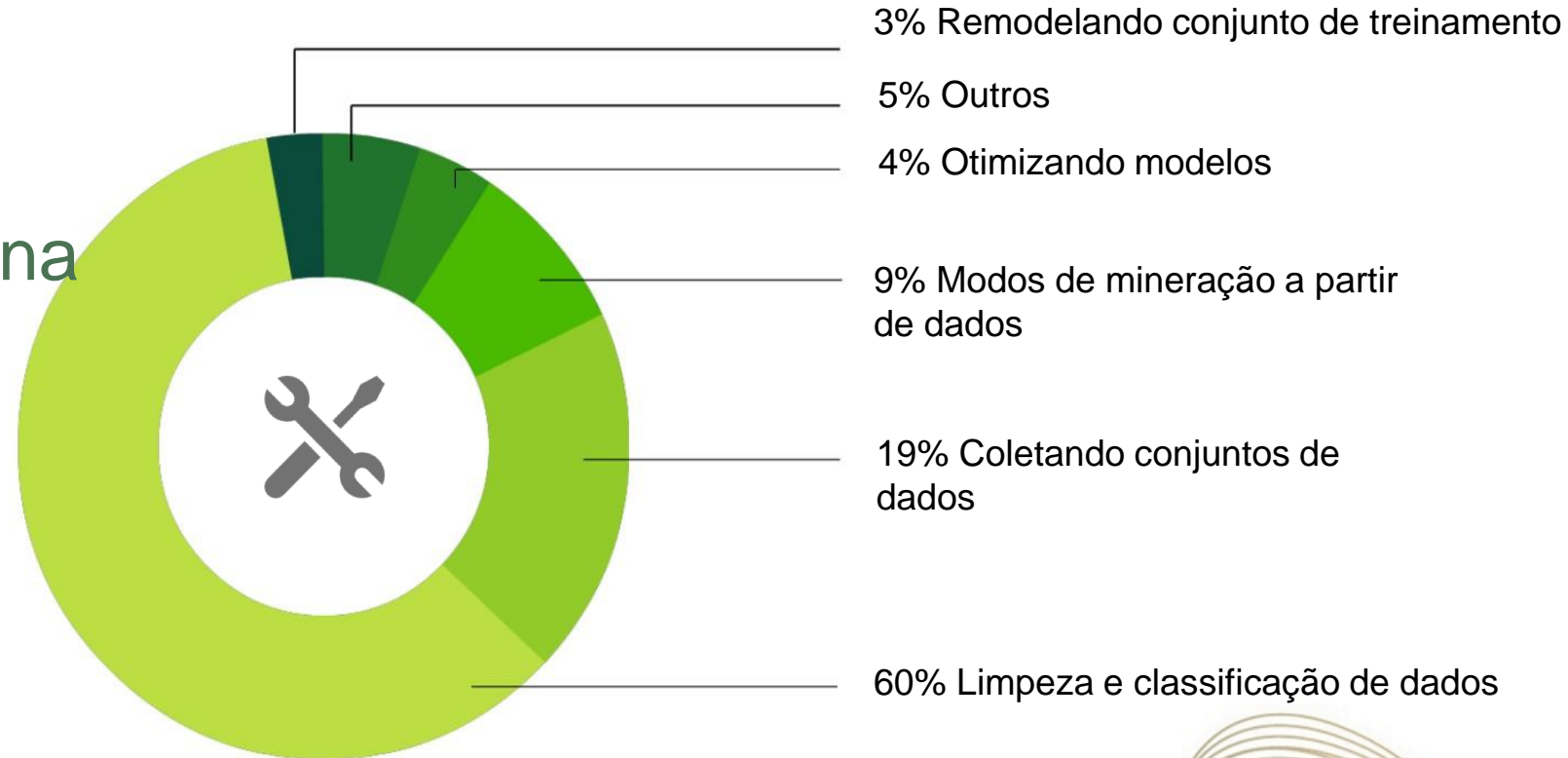
- A maioria dos modelos de aprendizado de máquina processa recursos, que geralmente são representações numéricas de variáveis de entrada que podem ser usadas no modelo.
- Na maioria dos casos, os dados coletados podem ser usados por algoritmos somente após serem pré-processados. As operações de pré-processamento incluem o seguinte:
 - Filtragem de dados
 - Processamento de dados perdidos
 - Processamento de possíveis exceções, erros ou valores anormais
 - Combinação de dados de várias fontes de dados
 - Consolidação de dados



IA

Esforço no trabalho de Limpeza de Dados

- Estatísticas sobre o trabalho dos cientistas de dados em aprendizado de máquina



CrowdFlower Data Science Report 2016



IA

Dados Sujos

- Geralmente, os dados reais possuem problemas:

- Incompletude
- Ruído
- Inconsistência

#	Id	Name	Birthday	Gender	IsTeacher	#Students	Country	City
1	111	John	31/12/1990	M	0	0	Ireland	Dublin
2	222	Mery	15/10/1978	F	1	15	Iceland	
3	333	Alice	19/04/2000	F	0	0	Spain	Madrid
4	444	Mark	01/11/1997	M	0	0	France	Paris
5	555	Alex	15/03/2000	A	1	23	Germany	Berlin
6	555	Peter	1983-12-01	M	1	10	Italy	Rome
7	777	Calvin	05/05/1995	M	0	0	Italy	Italy
8	888	Roxane	03/08/1948	F	0	0	Portugal	Lisbon
9	999	Anne	05/09/1992	F	0	5	Switzerland	Geneva
10	101010	Paul	14/11/1992	M	1	26	Ytali	Rome

Valor Faltando

Valor Inconsistente

Valor na coluna errada

Erro de Escrita

Item duplicado

Inconsistência De formato

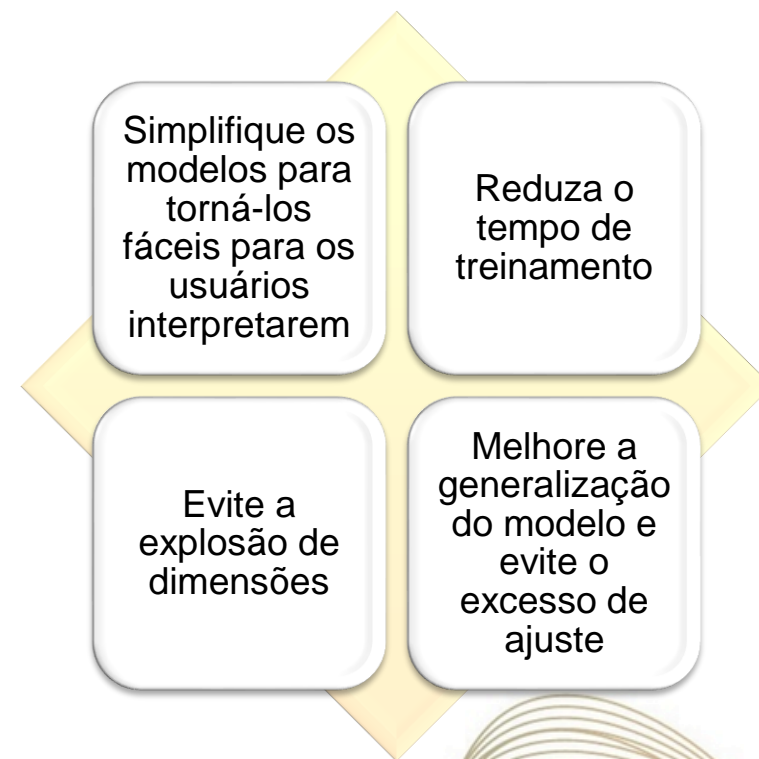
Dependência de Atributo



- Depois de serem pré-processados, os dados precisam ser convertidos em um formulário de representação adequado para o modelo de aprendizado de máquina. Formulários comuns de conversão de dados incluem o seguinte:
 - Com relação à classificação, os dados da categoria são codificados em uma representação numérica correspondente.
 - Os dados de valor são convertidos em dados de categoria para reduzir o valor das variáveis (para segmentação etária).
 - Outros dados
 - No texto, a palavra é convertida em um vetor de palavras através da incorporação de palavras (geralmente usando o modelo word2vec, modelo BERT, etc.).
 - Processar dados de imagem (espaço de cores, escala de cinza, alteração geométrica, recurso Haar e aprimoramento de imagem)
 - Engenharia de características:
 - Normalize os recursos para garantir os mesmos intervalos de valores para variáveis de entrada do mesmo modelo.
 - Expansão de recursos: combine ou converta variáveis existentes para gerar novos recursos, como a média.

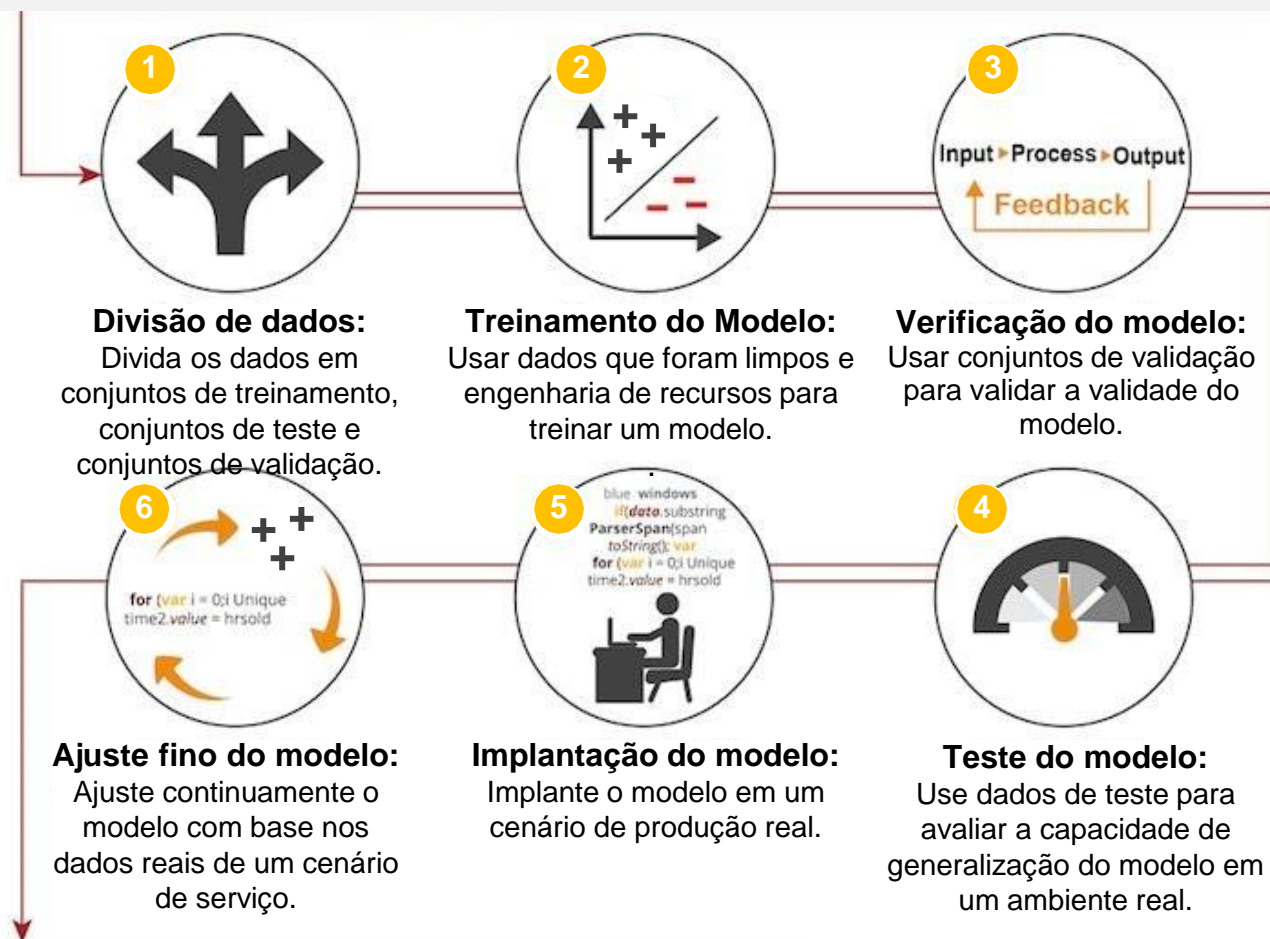


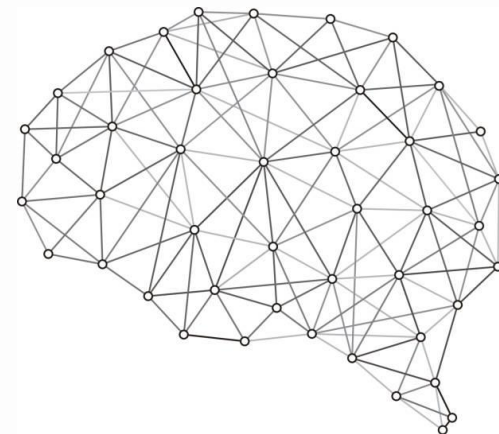
- Geralmente, um conjunto de dados tem muitos recursos, alguns dos quais podem ser redundantes ou irrelevantes para o valor a ser previsto.
- A seleção de recursos é necessária nos seguintes aspectos:
 - Simplificação de modelos
 - Redução de tempo de processamento
 - Evitar explosão de dimensionalidade
 - Melhorar o poder de aprendizado





Procedimento geral de construção de um modelo





Classificadores Supervisionados: Classificação Bayesiana



- Etapas do Processo de Aprendizado de Máquina
- **Exemplo de Aprendizado Supervisionado**
- Introdução aos classificadores por Probabilidade
- Classificador Bayesiano e Naive Bayes



UNIVERSIDADE
ESTADUAL DO CEARÁ



MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA
E INOVAÇÃO

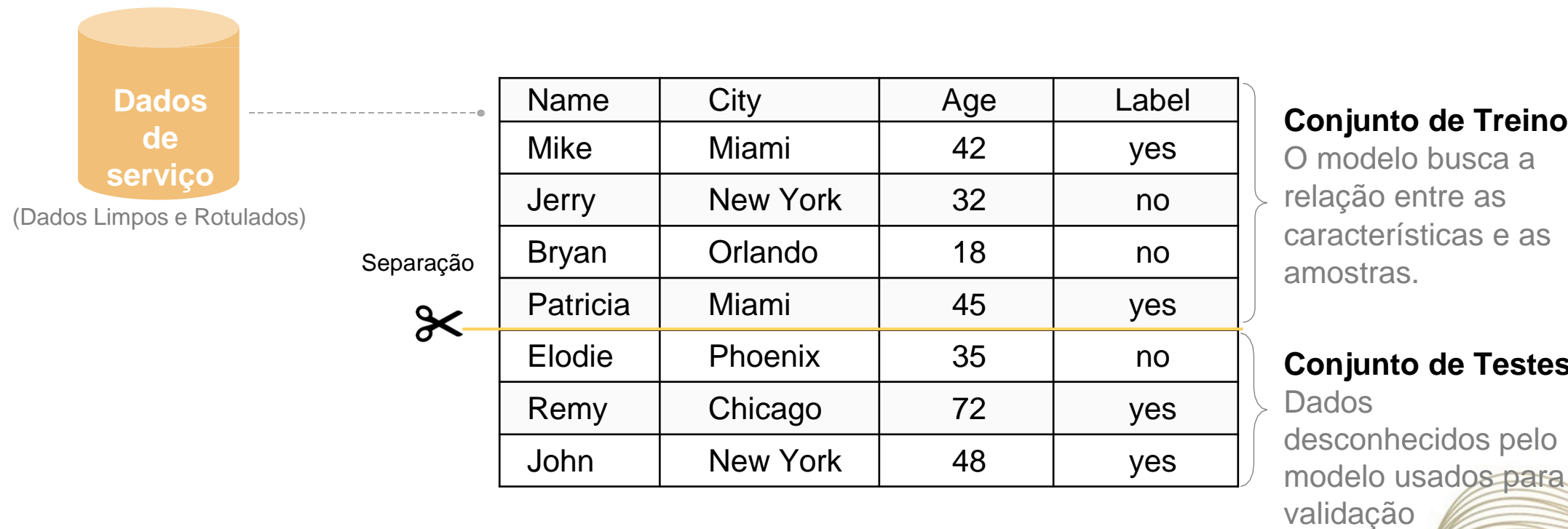




IA

Exemplo de Etapas do Aprendizado Supervisionado:

- 1. Divisão de Dados

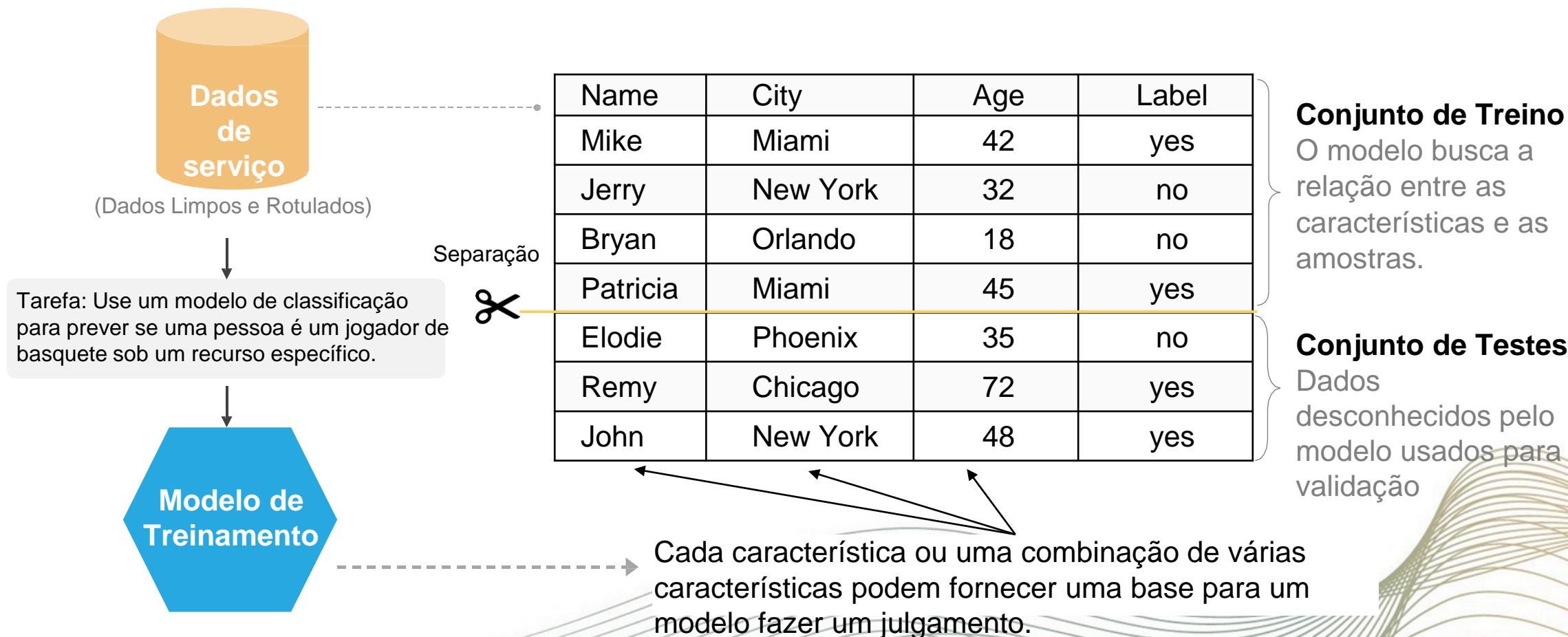




IA

Exemplo de Etapas do Aprendizado Supervisionado:

- 2. Treinamento do Modelo

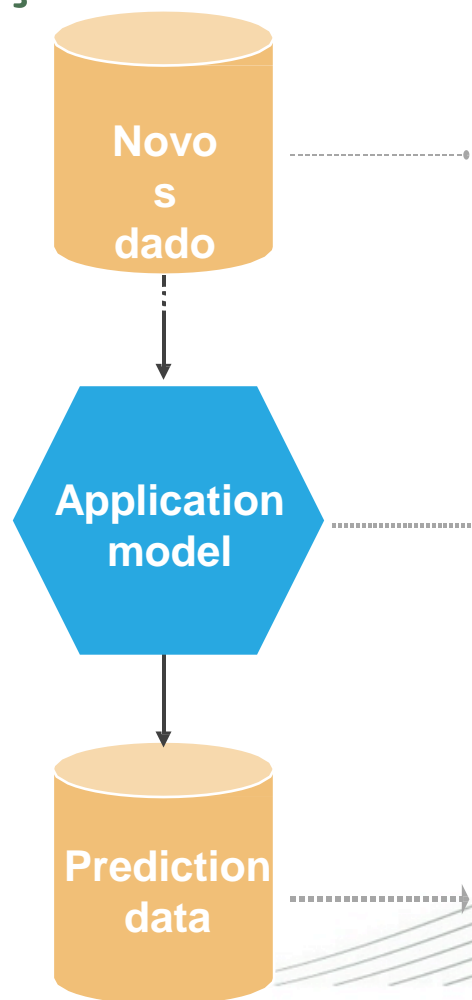




IA

Exemplo de Etapas do Aprendizado Supervisionado:

- 3. Verificação do Modelo



Name	City	Age	Label
Marine	Miami	45	?
Julien	Miami	52	?
Fred	Orlando	20	?
Michelle	Boston	34	?
Nicolas	Phoenix	90	?

Unknown data

Recent data, it is not known whether the people are basketball players.

IF city = Miami → Probability = +0.7
IF city = Orlando → Probability = +0.2
IF age > 42 → Probability = +0.05*age + 0.06
IF age ≤ 42 → Probability = +0.01*age + 0.02

Name	City	Age	Prediction
Marine	Miami	45	0.3
Julien	Miami	52	0.9
Fred	Orlando	20	0.6
Michelle	Boston	34	0.5
Nicolas	Phoenix	90	0.4

Possibility prediction

Apply the model to the new data to predict whether the customer will change the supply.



IA

Exemplo de Etapas do Aprendizado Supervisionado:

- 4 – Teste do Modelo
 - Capacidade de generalização
 - Ele pode prever com precisão os dados reais do serviço?
 - Interpretabilidade
 - O resultado da previsão é fácil de interpretar?
 - Velocidade de previsão
 - Quanto tempo leva para prever cada dado?
 - Exequibilidade
 - A taxa de previsão ainda é aceitável quando o volume de serviço aumenta com um enorme volume de dados?





Classificadores Supervisionados: Classificação Bayesiana



- Etapas do Processo de Aprendizado de Máquina
- Exemplo de Aprendizado Supervisionado
- **Introdução aos classificadores por probabilidade**
- Classificador Bayesiano e Naive Bayes



UNIVERSIDADE
ESTADUAL DO CEARÁ



MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA
E INOVAÇÃO

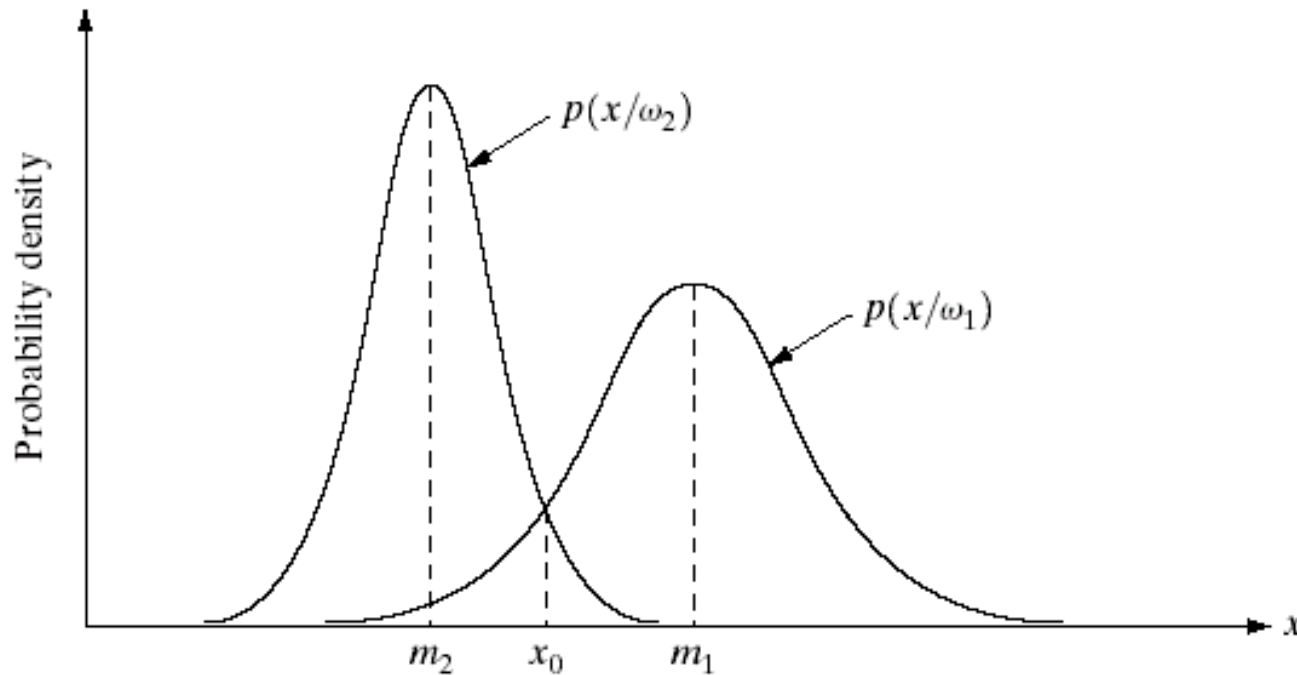




IA

Classificadores estatísticos

- Reconhece a natureza probabilística da informação que procuramos para processar e da forma pela qual devemos expressar os resultados.



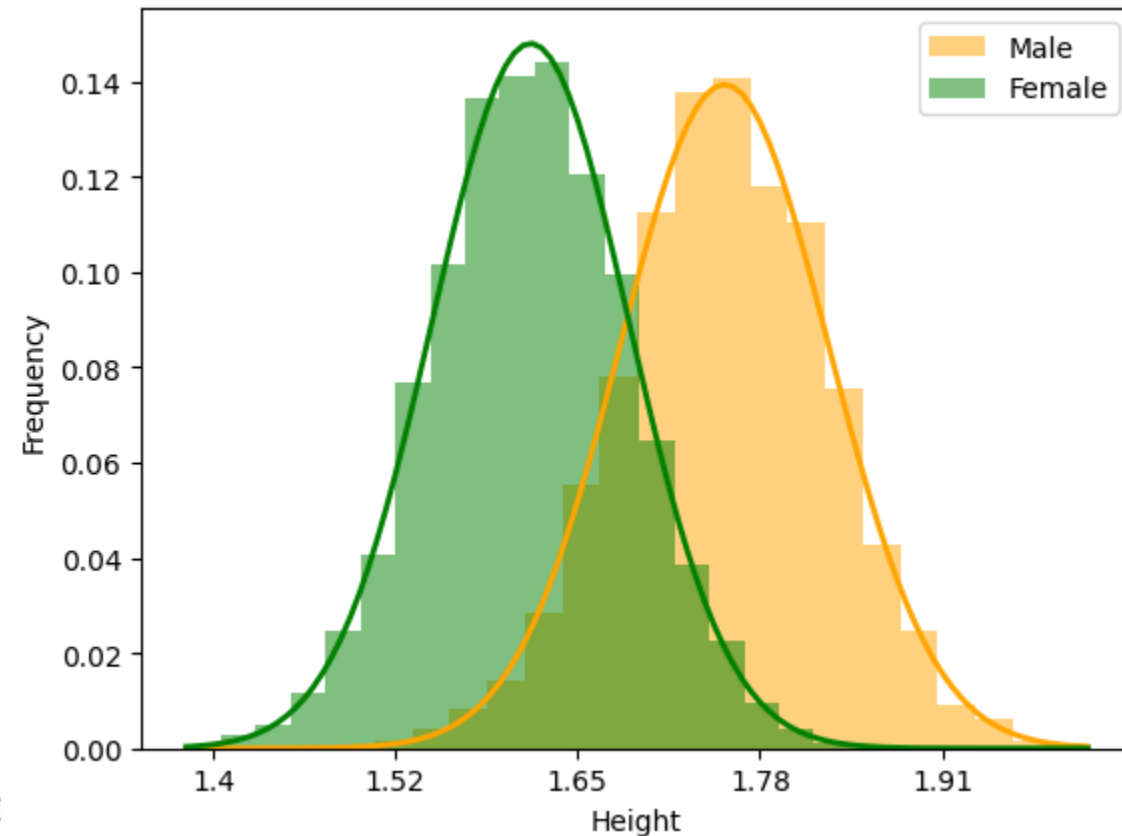
- Supondo um problema composto por duas classes $W=2$ em que as amostras possuem apenas 1 característica
- Se aproximarmos as distribuições dos padrões por gaussianas, teremos:
 - Classe 1: média μ_1 e desvio σ_1
 - Classe 2: média μ_2 e desvio σ_2



IA

Exemplo: Classificação estatística

- Poderíamos realizar uma pesquisa de uma população utilizando como característica a altura
- Ao adquirir o dataset, poderíamos calcular a média e desvio padrão de cada classe e aproximar cada uma das classes por uma distribuição normal.
- Então, uma pessoa com altura de 1,55 tem maior probabilidade de ser do sexo feminino, já uma pessoa com altura superior à 1,75 tem maior probabilidade de ser do sexo masculino





Classificadores Supervisionados: Classificação Bayesiana



- Etapas do Processo de Aprendizado de Máquina
- Exemplo de Aprendizado Supervisionado
- Introdução aos classificadores por probabilidade
- **Classificador Bayesiano e Naive Bayes**



UNIVERSIDADE
ESTADUAL DO CEARÁ



MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA
E INOVAÇÃO





IA

Classificador Bayesiano

- Para uma determinada característica de amostra X , a probabilidade a *posteriori* de uma amostra pertencer a uma categoria H é:

$$P(C_k|\mathbf{x}) = \frac{P(\mathbf{x}|C_k)P(C_k)}{P(\mathbf{x})}$$

- A classificação de uma amostra é feita a partir da escolha da classe que apresenta a maior probabilidade a posteriori, ou seja, a classe que maximiza a expressão $P(C_k|\mathbf{x})$
- Em que :
 - $P(C_k|\mathbf{x}) = P(C_k|X_1, \dots, X_n)$ é uma probabilidade a posteriori, ou uma probabilidade da amostra X pertencer a classe C_k .
 - $\mathbf{x} = X_1, \dots, X_n$ são características de dados, que geralmente são descritas por valores de medição de um conjunto de n atributos.
 - Por exemplo, o recurso de cor pode ter três atributos: vermelho, amarelo e azul.
 - C_k indica que os dados pertencem a uma classe específica C
 - $P(C_k)$ é uma probabilidade a priori.
 - $P(\mathbf{x}) = P(X_1, \dots, X_n)$ é a probabilidade a priori de X .





- um algoritmo de classificação multiclasse simples baseado no teorema de Bayes. Ele assume que os recursos são independentes uns dos outros, simplificando o cálculo da probabilidade a posteriori para:

$$P(C_k|\mathbf{x}) = P(C_k) \times \prod_{i=0}^n P(X_i|C_k)$$

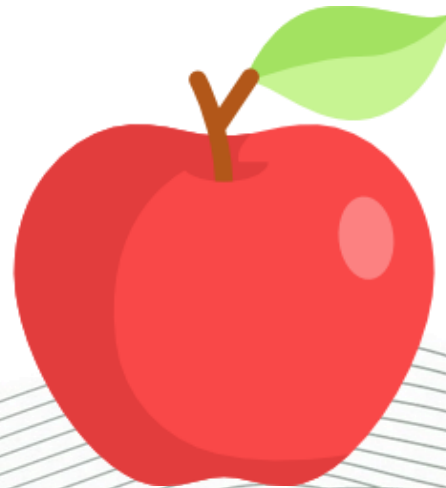
- Tipos de classificadores naive bayes:
 - Naive Bayes Gaussiano: assume que os dados de cada classe seguem uma distribuição Gaussiana (normal) e calcula a média e o desvio padrão para cada classe.
 - Naive Bayes Multinomial: é usado para dados discretos, como contagens de palavras em um texto. Ele calcula a frequência de ocorrência de cada palavra em cada classe.
 - Naive Bayes Bernoulli: é usado para dados binários, onde cada recurso pode estar presente (1) ou ausente (0).



IA

Exemplo Naive Bayes

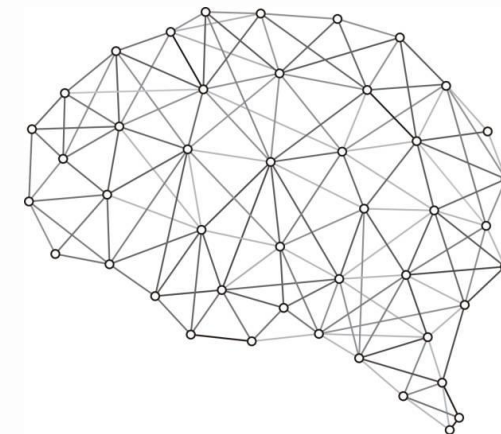
- Por exemplo, se uma fruta é vermelha, redonda e com cerca de 10 cm (3,94 in.) de diâmetro, ela pode ser considerada uma maçã.
- Um classificador Naive Bayes considera que cada característica contribui independentemente para a probabilidade de que o fruto seja uma maçã, independentemente de qualquer correlação possível entre a cor, a arredondamento e o diâmetro.





Na próxima aula

- Vamos praticar o Naive Bayes e vamos apresentar o KNN
- Também vamos comentar as formas de se avaliar os classificadores



UNIVERSIDADE
ESTADUAL DO CEARÁ

Dúvidas?

Módulo de Inteligência Artificial



UNIVERSIDADE
ESTADUAL DO CEARÁ



Instituto Iracema
PESQUISA E INOVAÇÃO



MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA
E INOVAÇÃO

