

Ciência de Dados

WEBINAR 03 – Unidade 02



UNIVERSIDADE
CANDIDO
MENDES

EAD ■

■ Introdução ao Big Data

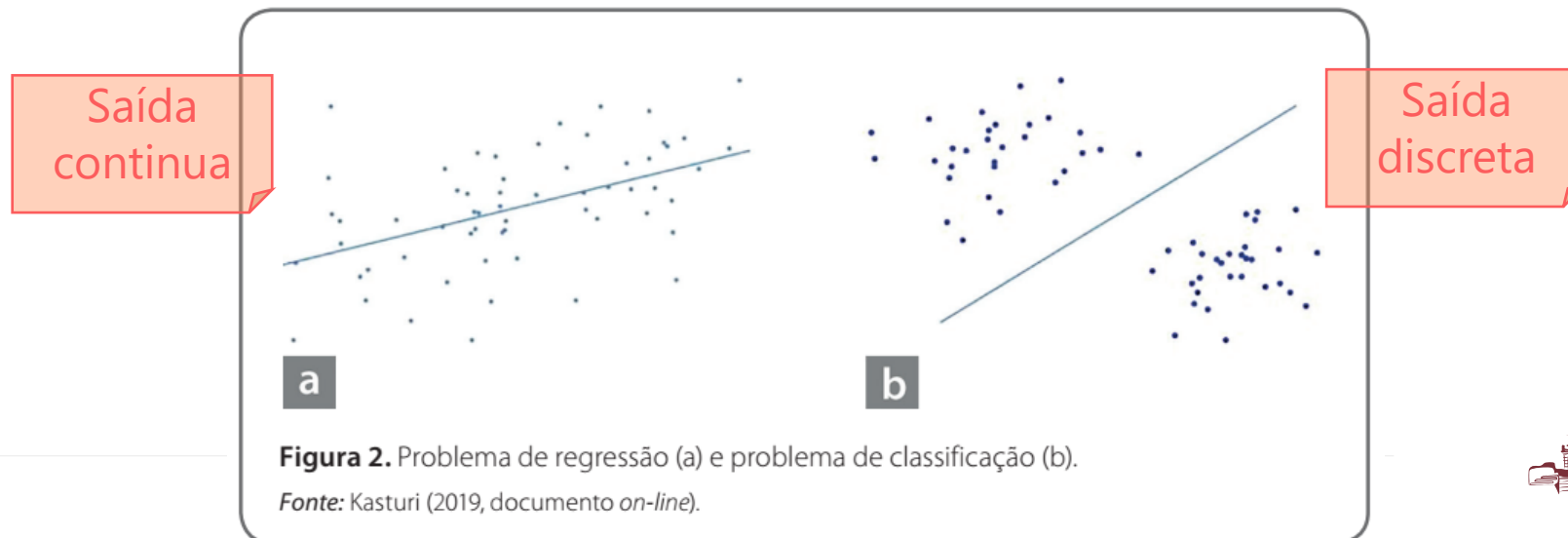
“ Me treine! ”

■ Técnicas de aprendizados em Big Data:

■ Tipos de aprendizados:

■ Aprendizado Supervisionado:

- Obtenção de um modelo generalizado capaz de resolver problemas a partir de dados de entrada;
- Necessita de um conjunto de dados de treinamento com variáveis independentes e dependentes associadas:
 - Dado um conjunto de x variáveis independentes $\rightarrow y$ é conhecido;
 - Ou seja, no aprendizado supervisionado, o modelo será ensinado sobre o que deve ser feito;
- A partir do aprendizado inicial é possível inserir novos dados e obter os resultados;



UNIVERSIDADE
CANDIDO
MENDES

EAD

■ Introdução ao Big Data

■ Técnicas de aprendizados em Big Data:

■ Tipos de aprendizados:

■ Aprendizado Supervisionado:

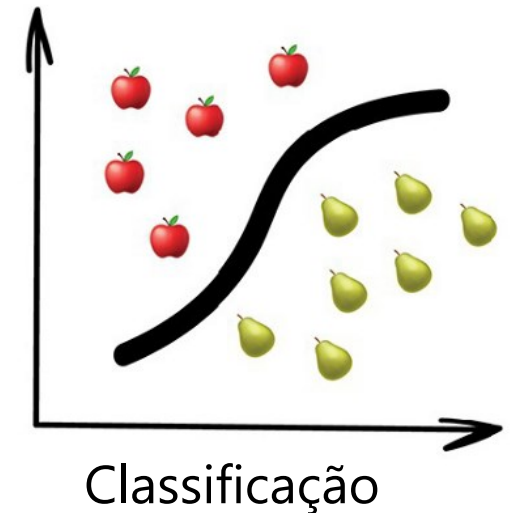
- Deve-se escolher o algoritmo pelo tipo de saída que deseja obter;
- Predição de classe categórica ou discreta:

- Problemas de classificação:

- Detecção de Spam: -> Spam / Não Spam
- Detecção de Fraudes: -> Fraude / Não Fraude
- Diagnóstico de doenças: -> Positivo / Negativo
- Classificação de clientes: -> Classes A, B ou C

- Algoritmos de classificação:

- KNN, Árvores de Decisão, SVM, **Regressão Logística**
- Obs.: Regressão Logística é utilizada em situações que a variável dependente é de natureza dicotômica ou binária;



■ Introdução ao Big Data

■ Técnicas de aprendizados em Big Data:

■ Tipos de aprendizados:

■ Aprendizado Supervisionado:

- Deve-se escolher o algoritmo pelo tipo de saída que deseja obter;

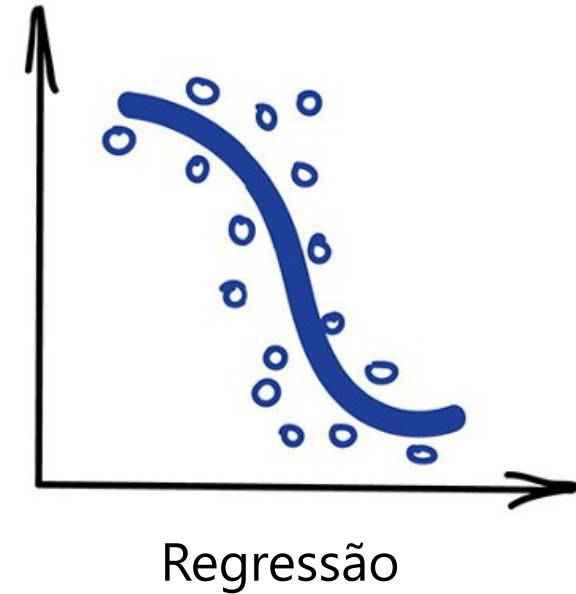
- Predição de classe contínua:

- Problemas de regressão:

- Predição de valores de imóveis;
- Predição de crescimento populacional;
- Predição de valores de ações;
- Avaliação de riscos;

- Algoritmos de regressão:

- Regressão Linear Simples; Regressão Linear Múltipla; Regressão Polinomial; **Árvores de Regressão (média)**;



■ Introdução ao Big Data

“ Sou autodidata,
aprendo sozinho ”

■ Técnicas de aprendizados em Big Data:

■ Tipos de aprendizados:

■ Aprendizado Não Supervisionado:

- Obtenção de um modelo generalizado capaz de resolver problemas a partir de dados de entrada;
- Baseia-se em reconhecimento de padrões;
- Não necessita de um conjunto de dados de treinamento com variáveis independentes e dependentes associadas:
 - Dado um conjunto de x variáveis independentes $\rightarrow y$ é desconhecido;
 - Ou seja, no aprendizado não supervisionado, não sabemos o que o modelo irá nos trazer como respostas;
- A dificuldade nesse modelo é observar as saídas após a análise e identificar se essas fazem sentido ou não;
- O aprendizado não supervisionado, muitas das vezes, pode servir de ponto de partida para implementação de um modelo supervisionado;
 - Exemplo: Classificação de clientes de e-commerce;
 - Identifica padrões, classifica, verifica se classificação faz sentido, utiliza os rótulos obtidos para treinar um algoritmo supervisionado para entrada de novos dados;



■ Introdução ao Big Data

■ Técnicas de aprendizados em Big Data:

■ Tipos de aprendizados:

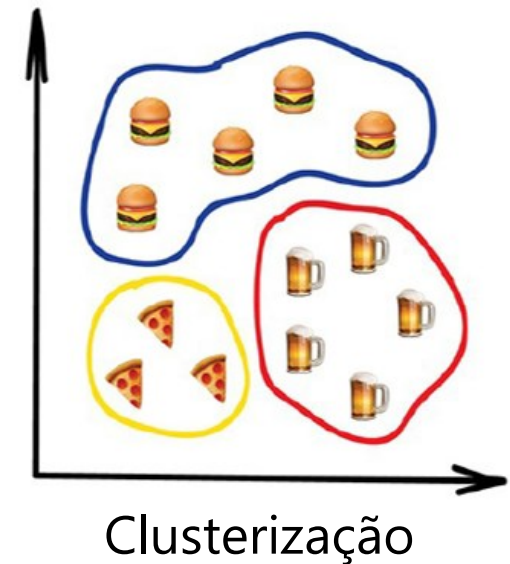
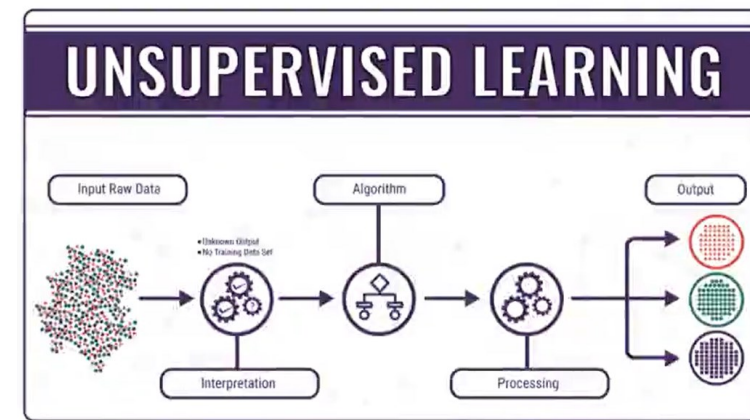
■ Aprendizado Não Supervisionado:

- Deve-se escolher o algoritmo pelo tipo de saída que deseja obter;
 - Clusterização; Associação; Redução de dimensão

■ Clusterização (tipos de classes discretas):

■ Problemas de clusterização:

- Segmentação de clientes: -> Tipos de clientes de acordo com seu perfil de compras;
- Segmentação de produtos: -> Tipos de produtos de acordo com suas características;
- Algoritmos de clusterização:
 - K-means; DBSCAN;



■ Introdução ao Big Data

■ Técnicas de aprendizados em Big Data:

■ Tipos de aprendizados:

■ Aprendizado Não Supervisionado:

- Deve-se escolher o algoritmo pelo tipo de saída que deseja obter;
 - Clusterização; Associação; Redução de dimensão

■ Associação (encontrar padrões):

■ Problemas de associação:

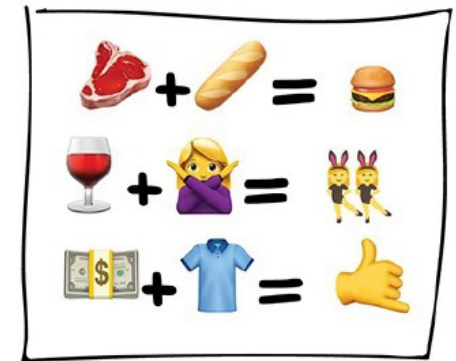
- Diagnóstico médico: -> Sintomas A e B -> Sintomas C;
- Cesta de compras: -> Clientes que compra cerveja na sexta-feira costumam comprar fraldas;
- Filmes: -> Pessoas que assistem o filme X -> Irão gostar do filme Y;

■ Algoritmos de associação:

- Apriori; FP-Growth;

• Exemplo de regras

- {Milk, Diaper} → {Beer} (s=0.4, c=0.67)
- {Milk, Beer} → {Diaper} (s=0.4, c=1.0)
- {Diaper, Beer} → {Milk} (s=0.4, c=0.67)
- {Beer} → {Milk, Diaper} (s=0.4, c=0.67)
- {Diaper} → {Milk, Beer} (s=0.4, c=0.5)
- {Milk} → {Diaper, Beer} (s=0.4, c=0.5)



Associação



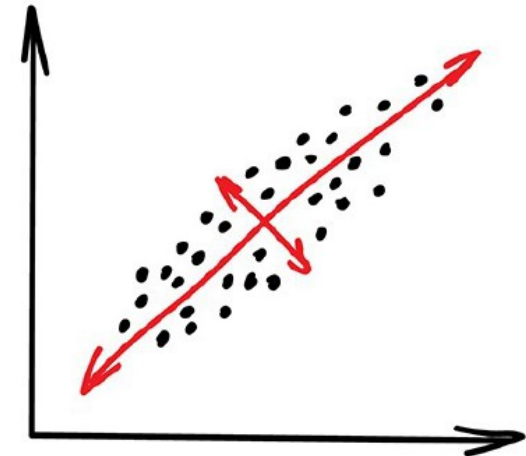
■ Introdução ao Big Data

■ Técnicas de aprendizados em Big Data:

■ Tipos de aprendizados:

■ Aprendizado Não Supervisionado:

- Deve-se escolher o algoritmo pelo tipo de saída que deseja obter;
 - Clusterização; Associação; Redução de dimensão
- Redução de Dimensionalidade (diminuição do número de variáveis):
 - Esses algoritmos geralmente antecedem outros algoritmos de Machine Learning;
 - O objetivo é diminuir a quantidade de variáveis a serem analisadas em um modelo posterior;
 - Algoritmos de clusterização:
 - Principal Component Analysis (PCA);
 - Singular Value Decomposition (SVD);



Redução de
Dimensão



■ Introdução ao Big Data

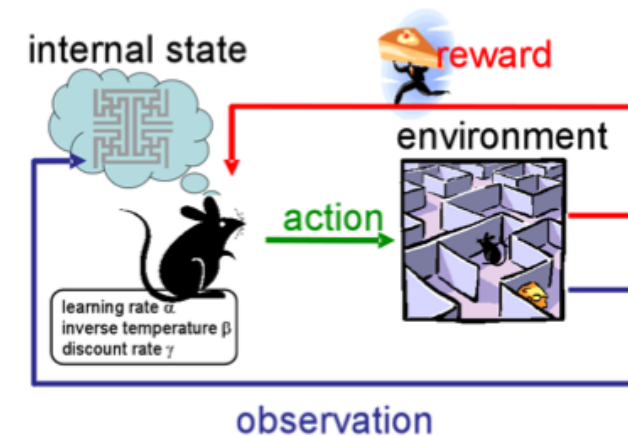
■ Técnicas de aprendizados em Big Data:

■ Tipos de aprendizados:

■ Aprendizado por Reforço:

- Não possuem dados de entrada, seja rotulados ou não;
- Mas dependem de um ambiente propício para realização do treinamento;
- Jogo de labirinto:
 - Se entro em um lugar sem saída, sou punido;
 - Se entro em um lugar que tenho saída, sou recompensado;
 - Imaginem o treinamento com um ratinho de laboratório, caso ele ache a saída irá receber um queijo. Após várias tentativas ele irá aprender o caminho;
- Características do aprendizado por reforço:
 - Não possui um supervisor, somente valores de recompensa e punição;
 - Tomada de decisão sequencial;
 - Número de repetições é importante para o aprendizado;
 - O retorno pode demorar (não instantâneo);
 - Deve-se primeiro realizar uma ação para então obter um retorno;

“ Aprendo por tentativa e erro ”



■ Introdução ao Big Data

■ Técnicas de aprendizados em Big Data:

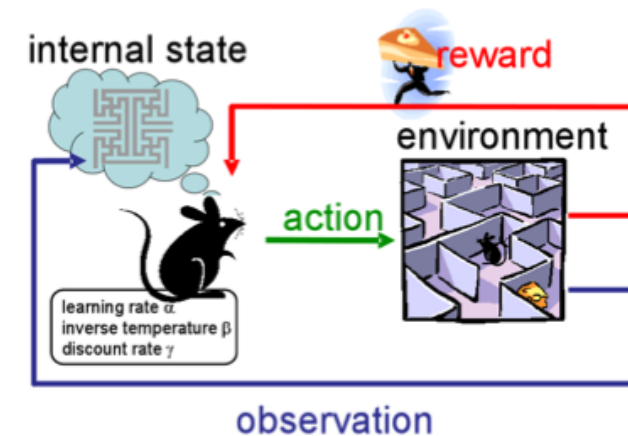
■ Tipos de aprendizados:

■ Aprendizado por Reforço:

■ Algoritmos de aprendizado por reforço:

- Cadeia de Markov;
- Q Learning;
- Problemas de aprendizado por reforço:
 - Construção de jogos;
 - Automação;

“ Aprendo por tentativa e erro ”



■ Introdução ao Big Data

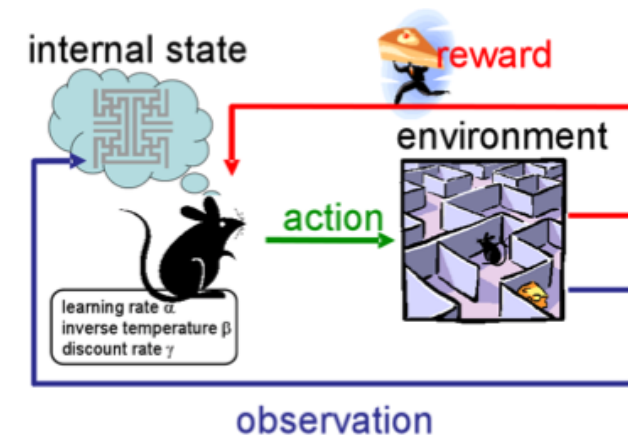
■ Técnicas de aprendizados em Big Data:

■ Tipos de aprendizados:

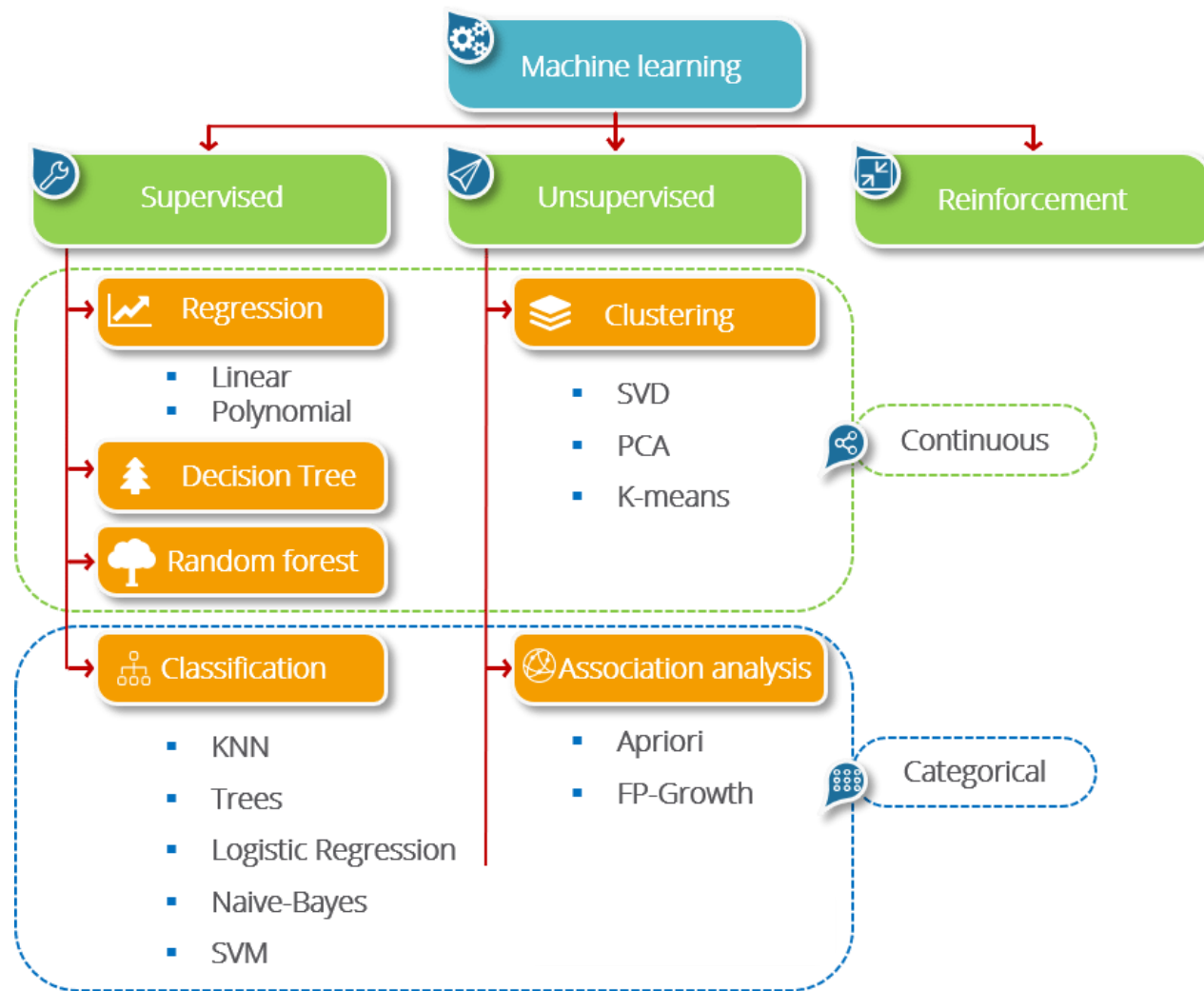
■ Aprendizado por Reforço:

- Não possuem dados de entrada, seja rotulados ou não;
- Mas dependem de um ambiente propício para realização do treinamento;
- Jogo de labirinto:
 - Se entro em um lugar sem saída, sou punido;
 - Se entro em um lugar que tenho saída, sou recompensado;
 - Imaginem o treinamento com um ratinho de laboratório, caso ele ache a saída irá receber um queijo. Após várias tentativas ele irá aprender o caminho;
- Características do aprendizado por reforço:
 - Não possui um supervisor, somente valores de recompensa e punição;
 - Tomada de decisão sequencial;
 - Número de repetições é importante para o aprendizado;
 - O retorno pode demorar (não instantâneo);
 - Deve-se primeiro realizar uma ação para então obter um retorno;

“ Aprendo por tentativa e erro ”



■ Introdução ao Big Data



Quadro 1. Comparativo das principais características entre os tipos de aprendizado

Características	Tipos de aprendizado		
	Supervisionado	Não supervisionado	Reforço
Conjunto de dados	Valores para atributo previsor e alvo.	Dados não rotulados.	Sem atributo-alvo.
Aprimoramento	Treinamento do modelo com base nas instâncias rotuladas.	Análise intrínseca.	Recompensas e punições.
Tarefa	Prever a resposta ou o rótulo correto.	Agrupar instâncias com características similares.	Buscar novas hipóteses no sentido de tentar reduzir as punições e aumentar as recompensas.

Temos ainda o Aprendizado Semissupervisionado, mas voltaremos a tratar do assunto na Unidade 03.

Treinamento com Redes Neurais não será abordado nessa disciplina devido sua complexidade.



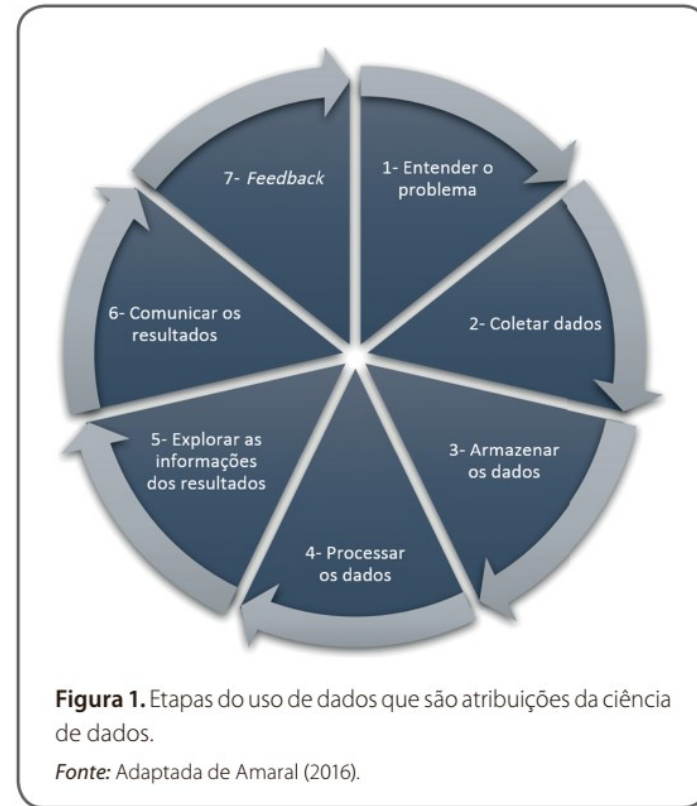
UNIVERSIDADE
CANDIDO
MENDES

EAD

■ Introdução ao Big Data

■ Etapas do Uso dos Dados:

- Entender o problema: determinar o tipo de informação desejada e as variáveis que fazem parte do processo;
- Coletar os dados: Dados podem ser comprados, produzidos ou coletados;
- Armazenamento: garantir a recuperação e a duplicação dos dados;
- Processamento e Exploração dos dados: buscar e aplicar metodologias adequadas para encontrar padrões, extrair informações e interpretar os resultados obtidos a partir da análise dos dados
- Comunicar os resultados: Relatórios, gráficos, sistemas, etc.
- Feedback: Os dados comunicados foram importantes? Levou ao Valor?



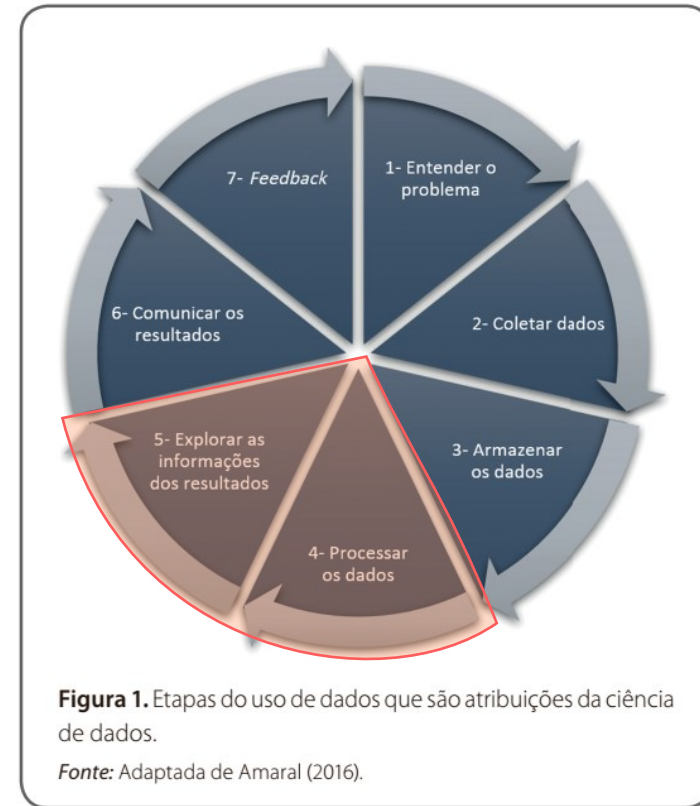
■ Introdução ao Big Data

■ Etapas do Uso dos Dados:

■ Entendendo os dados:

■ Através de estatística:

- Descritiva: responsável por descrever e resumir os dados por meio de gráficos, tabelas e números;
- Inferencial: inferir eventos prováveis, fundamentados pelas características dos dados;
- Probabilística: probabilidade de um evento ocorrer;



■ Introdução ao Big Data

■ Etapas do Uso dos Dados:

■ Estatística Descritiva:

■ Variáveis quantitativas:

- Variáveis contínuas: Números reais, escala continua;
 - Ex.: Peso, altura, tamanho, valor de produtos, tempo, etc.
- Variáveis discretas: Números inteiros;
 - Ex.: Número de pessoas, idade, quantidade de quartos, etc.

■ Variáveis qualitativas: remete a categorias

- Variáveis ordinais: Ordem entre as categorias;
 - Ex.: Meses do ano (jan, fev, mar,...), estágio da doença (inicial, intermediário, terminal), faixa etária; faixa de tamanho (pequeno, médio, grande);
- Variáveis nominais: Não existe ordem entre as categorias;
 - Ex.: Gênero, religião, raça, tipo sanguíneo, etc.



- **Introdução ao Big Data**
- Etapas do Uso dos Dados:
 - Estatística Descritiva na prática:
 - Jupyter Notebook



UNIVERSIDADE
CANDIDO
MENDES

EAD ■

