Ciência de Dados

WEBINAR 04

Revisão Unidade 01 e 02

UNIVERSIDADE CANDIDO MENDES | EAD

Revisão Unidade 01 e Unidade 02

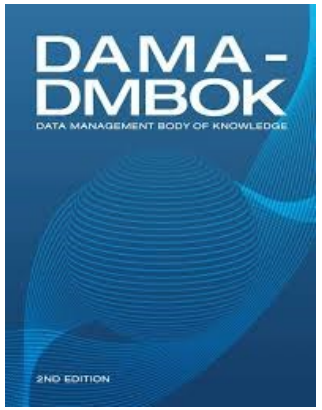
- Introdução a Big Data e Ciência de Dados:
  - O que é Big Data:
    - É uma área do conhecimento cujo objetivo é construir formas de **tratar, analisar e obter informações (insights)** oriundos de uma **grande quantidade de dados**;
    - Grande quantidade de dados não é só em volume!
    - Os 5 (ou 6, 7...) V's do Big Data:
      - Volume;
      - Velocidade;
      - Variedade;
      - Veracidade;
      - Valor

Revisão Unidade 01 e Unidade 02

- Introdução a Big Data e Ciência de Dados:
  - O que é Big Data:
    - Volume: Refere-se ao tamanho dos dados produzidos e à necessidade de serem armazenados;
      - Dados de vendas de produtos em uma empresa de varejo (ERP); Banco de imagens 3D de obras de artes da Itália;
    - Velocidade: A velocidade de produção de dados é muito alta e, muitas das vezes, sua interpretação deve ser tão veloz quanto sua geração;
      - Aquisição e interpretação de dados da disseminação da COVID-19; Controle de sinais de trânsito; Controle de Vendas e-commerce;
    - Variedade: Se temos um volume enorme de dados, também obtemos a variedade dos mesmos!
      - As mídias sociais geram, diariamente, um grande volume de dados não estruturados, como: e-mails, fotos, vídeos, áudios; Dados que precisam ser interpretados para entender o perfil do cliente;

Revisão Unidade 01 e Unidade 02

- Introdução a Big Data e Ciência de Dados:
  - O que é Big Data:
    - Veracidade: Trata-se de planejar e projetar o saneamento do dado, provendo qualidade ao mesmo, para que este dado possa gerar informações confiáveis para suportar a tomada de decisão
    - Ótica dupla:
      - Qualidade dos dados: dados completos, sem erros ou falta de dados parciais;
      - Qualidade dos dados: Dados que representam a realidade, sem viés
    - Valor: O armazenamento, limpeza, transformação e análise deve gerar valor agregado que compense os custos financeiros envolvidos (TAURION, 2013);
      - Big Data alinhado às redes sociais elevaram a Nestlé da 16ª para a 12ª posição entre as empresas com melhor reputação do mundo no mesmo ano de 2011. Atualmente a Nestlé está entre as 10 empresas com a melhor reputação do mundo.

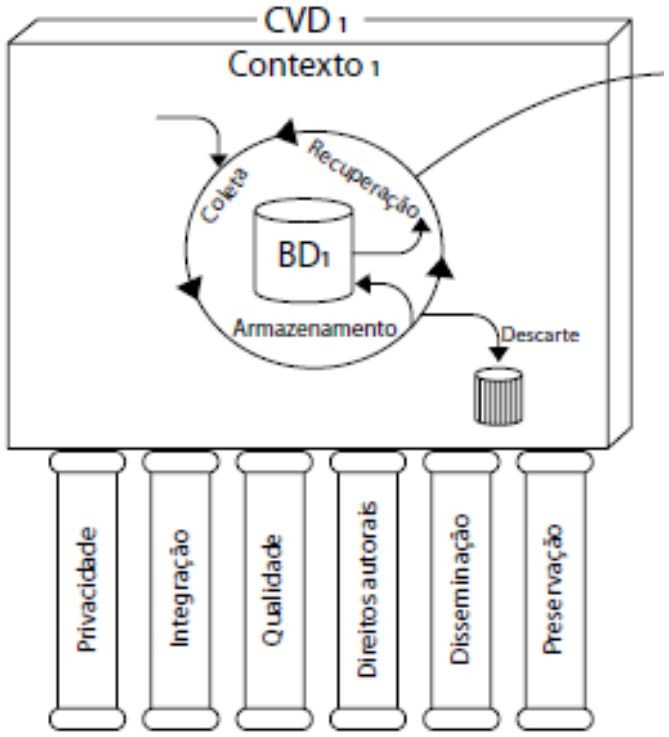


Revisão Unidade 01 e Unidade 02

- Introdução a Big Data e Ciência de Dados:
  - O que é Big Data:
    - Tipos de dados:
      - **Os dados estruturados** são aqueles organizados e representados com uma estrutura rígida.
        - Tabelas em Banco de Dados (MySQL, SQL Server, etc.) e planilhas eletrônicas;
      - **Os dados não-estruturados** são aqueles que não têm estrutura definida;
        - Posts nas Redes Sociais; Textos no Word ou Bloco de Notas; Vídeos; Áudios; Página da internet;
      - **Os dados semiestruturados** são aqueles que estão no meio termo. Possui certa estrutura, como o JSON ou XML, mas são flexíveis, pois é possível criar quantos campos de diferentes formatos quiser;
      - E os Bancos de Dados NoSQL (Mongo, Cassandra, etc.) representam dados estruturados ou não?!

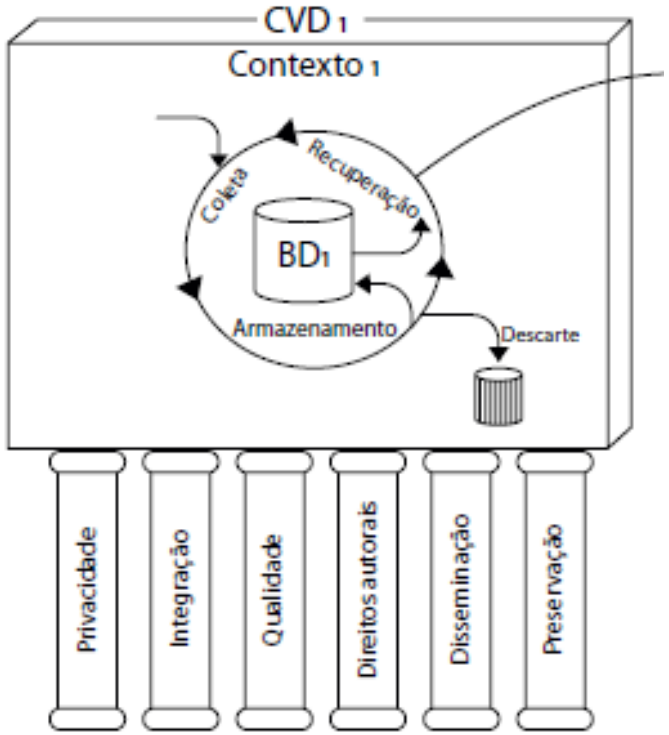
Revisão Unidade 01 e Unidade 02

- Introdução a Big Data e Ciência de Dados:
  - Ciclo de Vida dos Dados:
    - Segundo Sant'Ana (2016), o ciclo de vida dos dados é composto por quatro fases: coleta, armazenamento, recuperação e descarte.
    - Pilares de apoio ao projeto de Big Data:
      - **Qualidade dos dados:** A confiança (Veracidade) no contexto dos dados está intimamente ligada à sua qualidade;
      - **Privacidade:** Princípio da confidencialidade;
        - Lei nº 13.709/18 - Lei Geral de Proteção de Dados (BRASIL, 2018): regula as atividades de tratamento de dados pessoais;
      - **Disseminação:** Deve-se haver um plano bem definido de como a informação será disseminada para os usuários;
        - Sistemas Web; APP's; BI; Planilhas analíticas;
        - Não adiantar gerar uma informação que não seja utilizada para dar valor a empresa ou usuários



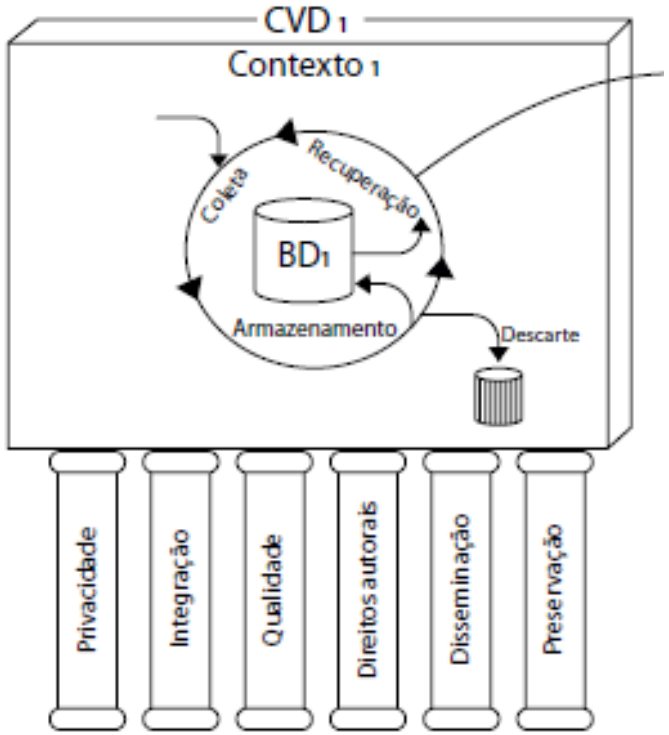
Revisão Unidade 01 e Unidade 02

- Introdução a Big Data e Ciência de Dados:
  - Ciclo de Vida dos Dados:
    - Pilares de apoio ao projeto de Big Data:
      - **Direitos autorais:** Dados protegidos por direitos autorais devem ser consumidos somente através de canais permitidos ou autorizados pelo detentor dos direitos, ou outorgados formalmente;
        - Caso Napster;
      - **Preservação:** Identificação dos dados necessários para construção de análise e resolução de problemas, armazenamentos e preservação para elaboração do projeto de Big Data;
        - **Os metadados são informações sobre os dados.**
          - Fotos: data, hora, localização, etc;



Revisão Unidade 01 e Unidade 02

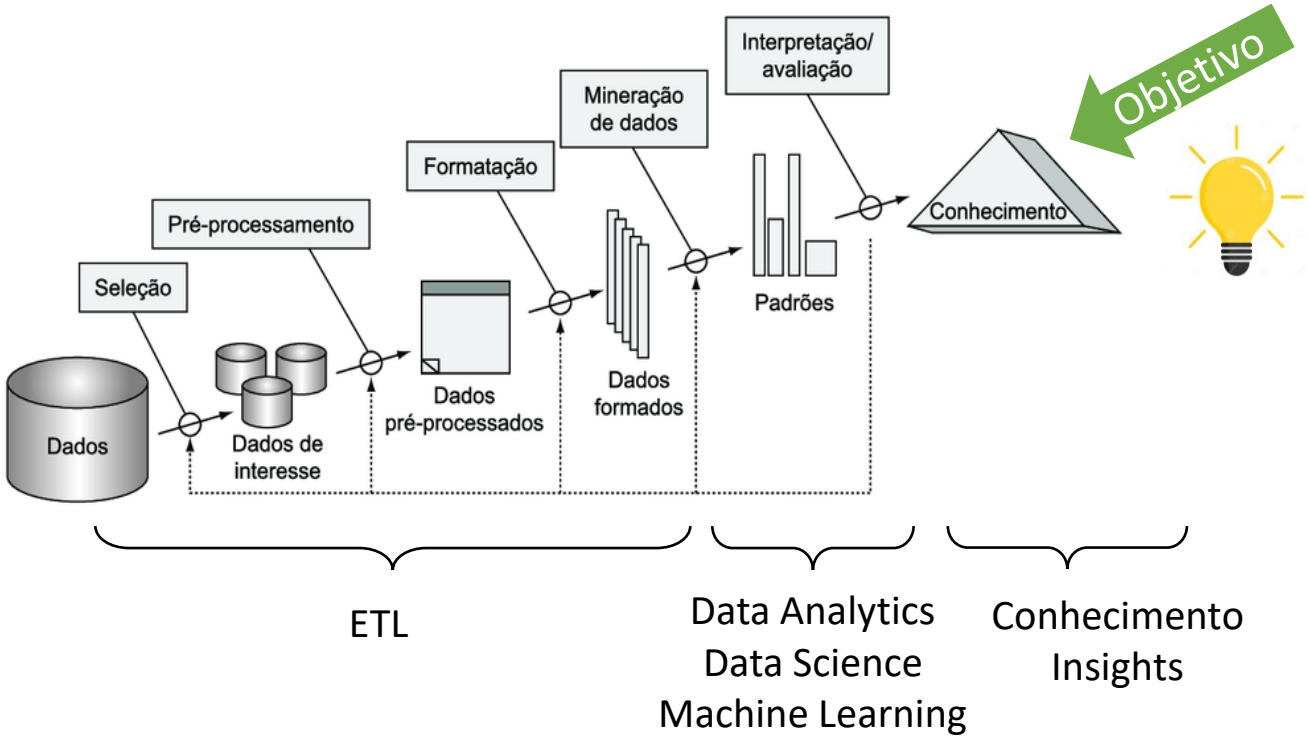
- Introdução a Big Data e Ciência de Dados:
  - Ciclo de Vida dos Dados:
    - Pilares de apoio ao projeto de Big Data:
      - **Integração:** Utilização de várias fontes de dados diferentes
        - Enriquecimento de Dados;
        - Exemplo:
          - Avaliação de Imóveis:
            - Dados de ofertas e vendas de imóveis;
            - Dados socioeconômicos:
              - Renda per capita;
              - IDH – Índice de Desenvolvimento Humano;
              - Taxa de Desemprego;
              - Oferta de serviços públicos;



Pilares

Introdução ao Big Data

- Introdução a Big Data e Ciência de Dados:
  - Não confunda! Ciclo de Vida de um projeto de Big Data:



Etapas do processo KDD (FAYYAD et al., 1996)

- **Seleção**: selecionar um conjunto ou subconjunto de dados que farão parte da análise.
- **Processamento**: esta etapa consiste em fazer a verificação da qualidade dos dados armazenados. Processo de limpar, corrigir ou remover dados inconsistentes, verificar dados ausentes ou incompletos, identificar anomalias (outliers).
- **Transformação** : esta etapa consiste em aplicar técnicas de transformação como: normalização, agregação, criação de novos atributos, redução e sintetização dos dados.
- **Mineração de Dados**: esta etapa consiste em construir modelos ou aplicar técnicas de mineração de dados. Essas técnicas têm por objetivo (1) verificar uma hipótese, (2) descobrir novos padrões de forma autônoma. Além disso, a descoberta pode ser dividida em: preditiva e descritiva. Esses modelos geralmente são aplicados e refeitos inúmeras vezes dependendo do objetivo do projeto.
- **Interpretação e Avaliação**: esta etapa consiste em avaliar o desempenho do modelo, aplicando em cima de dados que não foram utilizados na fase de treinamento ou mineração. A validação pode ser feita de diversas formas, algumas delas são: utilizar medidas estatísticas, passar pela avaliação dos profissionais de negócio.

Introdução ao Big Data

- Introdução a Big Data e Ciência de Dados:
  - Papéis técnicos nos processos de ciência de dados:

Administradores de Bancos de Dados (DBA's):

- Responsáveis por criar e manter bancos de dados;
- manutenção do servidor físico do banco de dados;
- recuperação de desastres;
- melhoria no desempenho de consultas ao banco de dados feito por aplicações da empresa;
- controle de acesso aos dados;
- criação de objetos (tabelas, funções, procedimentos, visualizações, etc.) no banco de dados;
- uso dos dados da empresa para alimentar os processos de negócios.

Engenheiro de dados:

- Responsáveis por criar a infraestrutura necessária para disponibilizar os dados necessários para consumo;
- Esse papel, muitas das vezes, são realizados pelos próprios DBA's, já que estes possuem o conhecimento da infra e da estrutura dos dados armazenados;
- Primeiro filtro de dados;
- Responsável pela criação de Data Warehouse e Data Lakes;
- Responsável pela migração de dados on-premisses para cloud;

Cientista de dados:

- Extrair conhecimento através dos dados brutos;
- Conhecimento em programação e softwares utilizados para manipular dados;
- Conhecimento matemático e estatístico;
- Trabalha em conjunto;
- Realiza a limpeza, tratamento, transformação, carga e análise de dados;
- Conhecimento do negócio;
- Responsável pela construção ou utilização de algoritmos de Aprendizado de Máquina;
- Geralmente, a construção de algoritmos de Machine Learning está migrando para um Engenheiro de Machine Learning;
- Conhecimento avançado em cálculo, álgebra, programação matemática;

Introdução ao Big Data

- Introdução a Big Data e Ciência de Dados:
  - Papéis técnicos nos processos de ciência de dados:

Aprendizado Supervisionado
"Me treine"

- Dados rotulados: variável dependente (y) é conhecida;

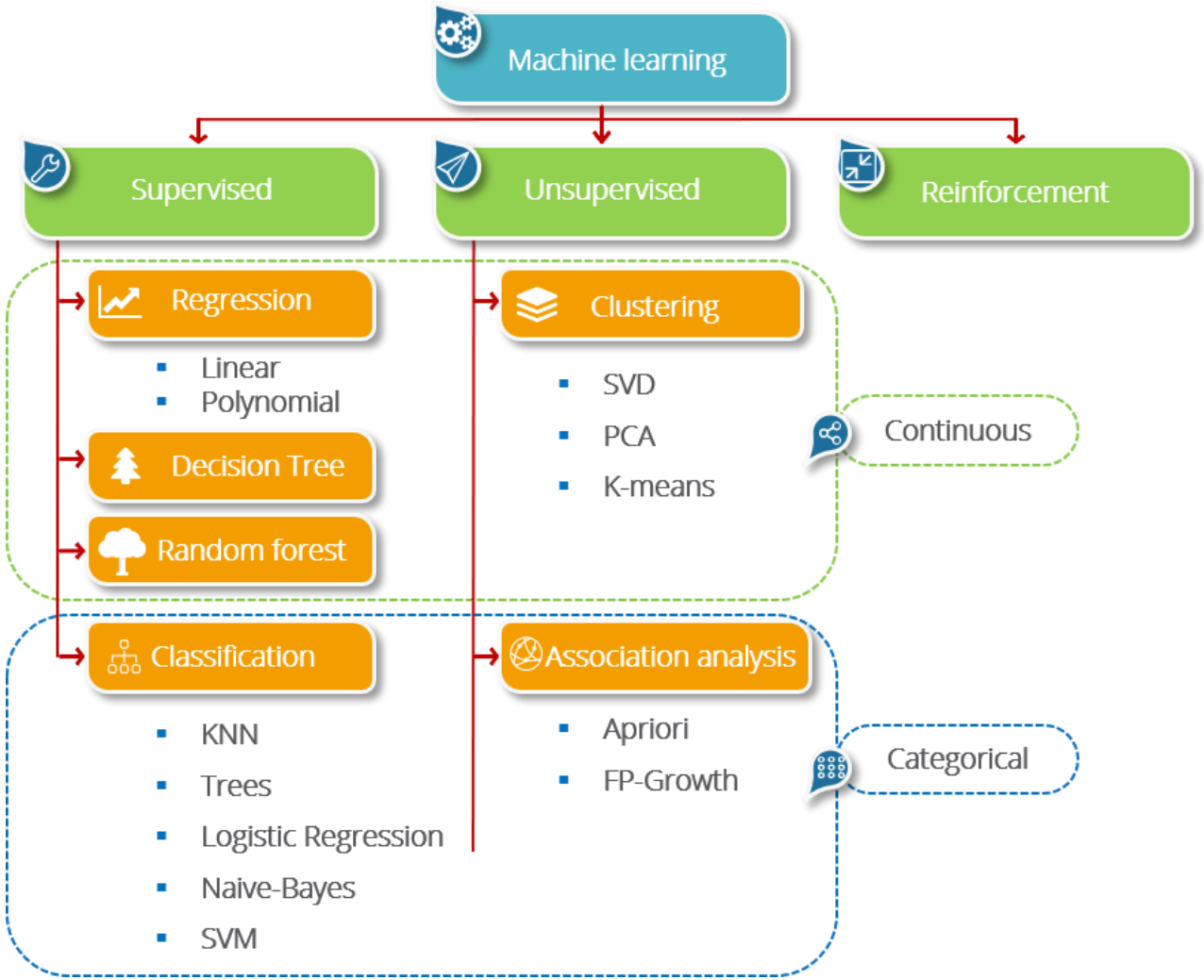
Aprendizado Não Supervisionado
"Sou autodidata, aprendo sozinho"

- Dados não rotulados: variável independente (y) não é conhecida;

Aprendizado por Reforço
"Aprendo por tentativa e erro"

- Depende de uma ambiente propício para realização do treinamento;
- Baseado em punição e recompensa;

Introdução ao Big Data



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

Características	Tipos de aprendizado		
	Semissupervisionado		Reforço
	Supervisionado	Não supervisionado	
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.

Introdução ao Big Data

- Etapas do Uso dos Dados:
  - 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.

UNIVERSIDADE
CANDIDO
MENDES

EAD