
ser
educacional
gente criando o futuro

ser educacional
gente criando o futuro

Inteligência Analítica
unidade 2

ANDRÉ TIBA (andre.tiba@sereducacional.com)

Sumário do curso

- Unidade 1 – Introdução à Estatística
- **Unidade 2 – Introdução à Mineração de Dados**
- Unidade 3 – Introdução à Modelos de Agrupamento e Predição
- Unidade 4 – Aplicação e Persistência ddo Conhecimento

Objetivos da unidade 2

- Entender conceitos de arquitetura tecnológica para ambientes analíticos
- Debater sobre extração de conhecimentos
- Primeiros conceitos de técnicas de mineração de dados
- Discutir sobre preparação de dados

Sumário da unidade 2

1) Arquitetura Tecnológica para Ambientes Analíticos
2) Descoberta do Conhecimento
3) Técnicas de Mineração de Dados
4) Preparação dos Dados

1) Arquitetura Tecnológica para Ambientes Analíticos

- Arquitetura Corporativa de Tecnologia da Informação (ACTI):
  - Método organizacional e sistemático, para construção de um ambiente de Tecnologia de Informação direcionado à empresas.
  - Integra processos e serviços ao ambiente de TI, dando agilidade à geração de negócios.

1) Arquitetura Tecnológica para Ambientes Analíticos

- Arquitetura Corporativa de Tecnologia da Informação (ACTI):
  - É definido como um processo (voltados à TI) que atua na organização interna de um negócio.
    - Conexões entre equipamentos e pessoas
    - Políticas de padronização de funcionamento operacional
    - Segurança e privacidade.

1) Arquitetura Tecnológica para Ambientes Analíticos

- No processo de evolução de uma empresa, ela precisará de atributos importantes, definidos pela ACTI:
  - Escalabilidade, flexibilidade, confiança e disponibilidade.

1) Introdução à Big Data

- Atualmente, geramos e coletamos uma quantidade absurda de dados:
  - pessoais, de trabalho, de negócios,
  - gerados por pessoas, por sensores, por processos automatizados.

1) Introdução à Big Data

- Dentro desta massa de dados há muita informação e conhecimento.
- As empresas precisam destes ativos para se manterem competitivas e sobreviverem no mercado.

1) Introdução à Big Data

- Big Data Analytics
  - Softwares especializados em coleta e tratamento de dados, para torná-los úteis (informação ou conhecimento) para empresas.
  - dados estruturados e dados não estruturados.

1) Infraestrutura para Big Data

- Nos primórdios, o custo da infraestrutura para ralizar Big Data era muito alto:
  - Construção de locais próprios (datacenters) onde ficavam os servidores.

1) Infraestrutura para Big Data

- Fatores possibilitaram expansão do uso do big data pelas empresas:
  - Barateamento dos hardwares
  - A criação da computação em nuvem
  - Evolução das Data Warehouses

1) Infraestrutura para Big Data

- Quatro fatores que devem ser considerados na infraestrutura do big data de uma empresa:

1) Coleta de dados:

    - Composto por: clientes, fornecedores, finança da empresa, banco de dados, redes sociais, etc ...

1) Infraestrutura para Big Data

- Quatro fatores que devem ser considerados na infraestrutura do big data de uma empresa:

2) Armazenamento dos dados:

  - Em geral são armazenados em um data warehouse (dados estruturados) ou em data lakes (dados não estruturados).

  - Para a maioria das empresas a solução está no armazenamento em nuvem.

1) Infraestrutura para Big Data

Data lake	Data warehouse
Armazenamento de dados desestruturados, Dados semi-estruturados e estruturados	Dados estruturados
Esquema definido na leitura	Esquema definido na escrita
Ciência de dados, análise preditivas, BI	BI baseado em SQL
Armazenamento de dados detalhados, brutos e também processados	Armazenamento de dados frequentemente acessados, assim como dados agregados e sumarizados
Separação entre o armazenamento e o processamento	Acoplamento entre o armazenamento e o processamento

1) Infraestrutura para Big Data

- Quatro fatores que devem ser considerados na infraestrutura do big data de uma empresa:

3) Análise dos dados:

  - Etapa1: tratamento ou pré processamento dos dados → limpeza, formatação, etc...

  - Etapa 2: construção dos modelos analíticos → grandes provedores de computação em nuvem (google, amazon, microsoft) possuem suas próprias ferramentas.

1) Infraestrutura para Big Data

- Quatro fatores que devem ser considerados na infraestrutura do big data de uma empresa:

4) Visualização e saída dos dados:

- Os resultados da análise de dados são em geral visualizados por meio de gráficos, relatórios, recomendações-chaves, dashboards (ferramentas simples capazes de prover diversos tipos de visualizações de dados).

1) Tecnologias fundamentais para Big Data e Inteligência Analítica

- Apache Hadoop
  - Projeto de código aberto escrito em Java
  - Baseado em computação distribuída
  - Dados são arquivados de forma redundante
  - Adaptado para suportar falhas e se manter funcionado

1) Tecnologias fundamentais para Big Data e Inteligência Analítica

- Apache Hive
  - Solução de data warehouse sobre o Haddop
  - Trabalha com linguagem declarativa semelhante ao SQL (HiveQL)Baseado em computação distribuída
  - Os dados são organizados em: tabelas, partições e buckets (baldes).

1) Tecnologias fundamentais para Big Data e Inteligência Analítica

- Apache Spark
  - É considerado um framework com capacidade de fluxo.
  - Foi desenvolvido para operar com grandes volumes de dados em alta performance de velocidade.
  - Por isso possibilita análise tem tempo real
  - Trabalha direto na memória, reduzindo o tempo de leitura e escrita em disco.

2) Descoberta de Conhecimento

- As mídias sociais como Facebook, Twitter, e aplicativos como WhatsApp, Instagram, dentre outros, tem gerado uma massa de dados, que só cresce
  - existe MUITO conhecimento "escondido" e que pode ser extraído destes dados.
- Boa parte destes dados pertencem a empresas, portanto são privados!

2) Descoberta de Conhecimento

- As mídias sociais criaram um “novo” formato de divulgação de noticias e comunicação através do compartilhamento de conteúdo.
  - Antes da Big Data, informações e conhecimentos eram divulgadas apenas por sites.

2) Descoberta de Conhecimento

- Hoje temos um leque de mídias digitais, compostas por plataformas de streaming e redes sociais diversas, que criam, divulgam e compartilham conteúdos e conhecimentos.
  - Possibilita socializar conhecimento e informação além das questões pessoais.
- A grande massa de informações disponíveis na rede existe porque ferramentas associadas a Big Data e computação em nuvem viabilizaram esta façanha!

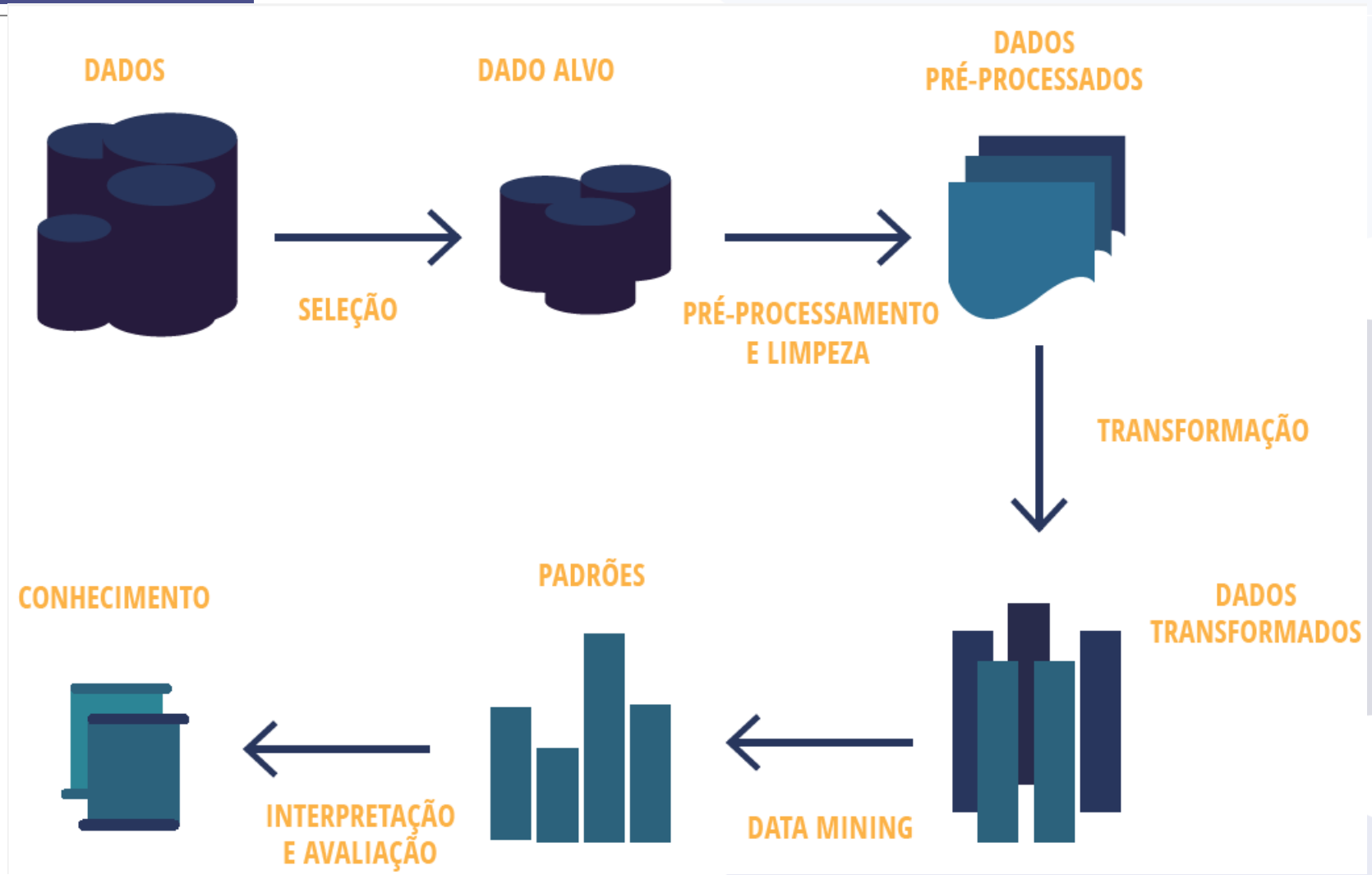
2) Descoberta de Conhecimento

- Estudos mostram que existe uma correlação entre a capacidade de acesso aos dado e a qualidade dos conhecimento gerado e compartilhado.
  - Países com melhor infraestrutura possibilitam à sua população menor custo para acessar dados em quantidade e variedade.
  - Isso reflete diretamente na criação de conhecimentos mais relevantes.

2) Descoberta de Conhecimento em Banco de Dados (Knowledge Discovery in Databases, KDD)

- A Big Data possibilitou que enormes massas de dados estruturados sejam armazenados.
  - Transações diversas no formato eletrônico
  - Destes dados são mais “fáceis” de se extrair informações e conhecimento que dos dados não estruturados.

2) Descoberta de Conhecimento em Banco de Dados (Knowledge Discovery in Databases, KDD)



2) Descoberta de Conhecimento em Banco de Dados (Knowledge Discovery in Databases, KDD)

- Tipos de conhecimento obtidos no KDD:
  - Superficial: recuperação de informação através de uma simples consulta.
  - Multidimensional: análise de dados através de rápidas consultas, por exemplo o SQL.

2) Descoberta de Conhecimento em Banco de Dados (Knowledge Discovery in Databases, KDD)

- Tipos de conhecimento obtidos no KDD:
  - Oculto: necessita de algoritmos de aprendizagem de máquinas para reconhecimento e identificação de padrões.
  - Profundo: necessita de algoritmos de AM ainda mais sofisticados, pois conhecimento está difuso. O conhecimento está contido implicitamente na massa sobre toda a massa de dados BD.

2) Descoberta de Conhecimento em Banco de Dados (Knowledge Discovery in Databases, KDD)

- Dificuldades e Limitações para extração de conhecimentos em BD:
  - Visualização de informação em bases de dados cada vez maiores, com dezenas de atributos.
  - Quanto maior a base de dados, mas será necessário realizar pré processamento para se utilizar algoritmos de ML
    - Envolve 70% do tempo gasto, em média. Os outros 30% são de fato para encontrar uma abordagem de ML capaz de apresentar uma boa solução.

3) Mineração de Dados (Data Mining, DM)

- Conjunto de técnicas capazes de tratar grandes volumes de dados, extraindo informações sob o ponto de vista estatístico.
  - Estabelece relações entre os dados (atributos) sem aparente conexão lógica.
  - Dentro de um banco de dados, fornece ferramentas para reconhecimento de padrões.

3) Mineração de Dados (Data Mining, DM)

- Tarefas mais comuns da mineração de dados:
  - **descrição**: estabelece características gerais de um conjunto. Define o que se caracteriza por ser daquele padrão. Possui caráter exploratório.
    - Ex: descrever o perfil de um consumidor vegetariano
  - **classificação**: tem como objetivo identificar a qual classe pertence o dado.
    - Ex: classificar um tomador de empréstico entre bom ou mal pagador

3) Mineração de Dados (Data Mining, DM)

- Tarefas mais comuns da mineração de dados:
  - **estimação** ou **regressão**: tem como objetivo estimar um resultado numérico.
    - Ex: estimar as condições físicas usando atributos como peso, altura, pressão, taxas diversas, etc ..
  - **predição**: tem como objetivo prever o valor de um futuro de um atributo.
    - Ex: prever o valor de um imóvel daqui a 6 meses

3) Mineração de Dados (Data Mining, DM)

- Tarefas mais comuns da mineração de dados:
  - **agrupamento**: agrupar dados similares
    - Ex: criação de perfis de consumidores para indicação de produtos.
  - **associação**: associar atributos (nem sempre há uma relação direta entre os atributos).
    - Ex: lista de compras. Homens que vão a noite comprar fraldas, também compram bebidas.

3) Mineração de Dados - Visão Geral

- Data Mining e Data Warehouse estão relacionados.
  - Data Warehouse → tomada de decisão em uma base de dados.
  - Data mining → aplicado a conjuntos específicos de dados entre os atributos).

3) Mineração de Dados - Visão Geral

- A Descoberta de Conhecimento em BD (KDD) segue seis etapas:
  - Seleção de dados
  - Limpeza de dados
  - Enriquecimento dos dados
  - Passagem pela codificação
  - Relatórios
  - Apresentação dos conhecimentos

3) Mineração de Dados - Visão Geral

- Tipos conhecimentos que podem ser extraídos:
  - **Regras Associativas**
    - Ex: ao adquirir um produto, compra-se outro associado ao primeiro.
    - Ex: Compra-se uma televisão e em seguida um home-theater

3) Mineração de Dados - Visão Geral

- Tipos conhecimentos que podem ser extraídos:
  - **Padrões sequenciais**
    - Um cliente adquire um produto uma sequencia de produtos, ao longo do tempo.
    - O cliente compra uma geladeira, depois um fogão, então o sistema entende que o cliente está trocando sua cozinha e oferece uma oferta de um microondas.

3) Mineração de Dados - Visão Geral

- Tipos conhecimentos que podem ser extraídos:
  - **Árvores de decisão (ou classificação)**
    - Sistema capaz de estabelecer regras claras para classificação de padrões (algo raro em ML).

3) Mineração de Dados - Visão Geral

- Principais **metas** desejadas em Mineração de Dados:
  - **Predição**
    - Deseja-se prever um valor numérico de um atributo com base em um conjunto de atributos.

3) Mineração de Dados - Visão Geral

- Principais **metas** desejadas em Mineração de Dados :
  - **Identificação**
    - Deseja-se classificar/rotular um dado, dentro de um conjunto rótulos possíveis.

3) Mineração de Dados - Visão Geral

- Principais **metas** desejadas em Mineração de Dados :
  - **Otimização**
    - Deseja-se encontrar um conjunto de ações que otimize um processo ou um produto.

3) Mineração de Dados - Técnicas usadas em Data Mining

- Técnicas de **Classificação**:
  - Tem como objetivo classificar um dado (chamado também de padrão).
  - O classificador é treinado para reconhecer e classificar um conjunto de dados (conjunto de treino).
  - Espera-se que ele seja **capaz** de realizar uma boa classificação para um novo dado (que não pertence ao conjunto de treino).

3) Mineração de Dados - Técnicas usadas em Data Mining

- Técnicas que procuram por **Padrões Sequenciais**:
  - Extrair informações a partir de sequencias de ações.
    - Ex: um consumidor faz suas compras regularmente em um dado supermercado. A partir da analise das listas de compras que ele fez ao de um ano, tentar estabelecer padrões de consumo a partir destes consumos anteriores.

3) Mineração de Dados - Técnicas usadas em Data Mining

- Técnicas que analisam **Padrões em Série Temporal**
  - Predizer o valor futuro a partir dos valores passados.
    - Uma série temporal é uma sequencia de eventos, que ocorre com uma regularidade estabelecida.
      - Ex: cotação do dólar (série diária)
      - Ex: índice da inflação (série mensal ou anual)

3) Mineração de Dados - Técnicas usadas em Data Mining

- Técnicas que tratam de problemas de **Regressão**
  - Estimar o valor de um atributo a partir de um conjunto de outros atributos.
    - Pode ser usada em problemas sequenciais (séries temporais)
    - Pode ser usada para problemas não sequenciais

3) Mineração de Dados - Técnicas usadas em Data Mining

- Redes Neurais
  - Podem ser usadas para tratar problemas de classificação, de regressão, de séries temporais, de agrupamento.
  - Existem muitos modelos distintos de redes neurais.
  - São eficientes para resolver problemas, mas não é capaz de descrever bem como consegue resolver. Apenas resolve!

3) Mineração de Dados - Técnicas usadas em Data Mining

- Algoritmos Genéticos (Computação Evolucionária)
  - Usados para problemas de otimização.
  - É um algoritmo de busca baseado em princípios de genética.

3) Mineração de Dados - Aplicações

- Em finanças: análise de crédito, de financiamento, de risco, predição de ativos, etc ...
- Em marketing: entender padrões de consumo, estabelecer perfis e comportamento de clientes.
- Em saúde: realizar diagnósticos estatísticos a partir de imagens, de dados clínicos.
- Em produção: otimização de recursos e de processos.

4) Preparação dos Dados

- É uma parte MUITO importante do processo de análise de dados.
- Se não for feita com cuidado, pode INVIABILIZAR a análise dos dados.
- Pode ser trabalhosa e demorada, se os dados estiverem sem algum tratamento prévio (dados “crus”).

4) Preparação dos Dados

- Em linhas gerais, consiste nas seguintes etapas:
  - Coleta
  - Limpeza
  - Combinação/Redução
  - Estruturação e organização (reescala/mudança de tipo, etc ...)

4) Preparação dos Dados

- Dados organizados e estruturados podem ser analisados em diferentes escalas.
- Também podem ser analisados de forma descentralizada.
  - Ex: processamento em paralelo (o que diminui o tempo de processamento)
- A confiabilidade dos resultados obtidos pela análise de dados passa por uma preparação de dados metódica e consistente.

4) Preparação dos Dados

- **Coleta dos dados**:
  - Se possível, planejar que tipo de dado será coletado, em que formato será coletado
    - Isso facilita as etapas seguintes

4) Preparação dos Dados

- **Tratamento e Limpeza**:
  - Necessário para descartar inconsistências
    - Ex: um local onde deve-se indicar uma peça de roupa tem-se “teia” ao invés “meia”
  - Dados ausentes: um grande problema!
  - Valores discrepantes
    - Ex: no atributo idade tem-se marcado 142 anos

4) Preparação dos Dados

- **Transformação/Reduções**
  - Retirada de atributos desnecessários
  - Transformação dos atributos originais
    - Ex: é categórico e precisa ser transformado em numérico
  - Reduz da dimensionalidade da base de dados usando uma combinação de atributos

4) Preparação dos Dados

- **Produção de nova base**
  - Em problemas de classificação, as bases podem estar desbalanceadas, sendo necessário balanceá-las (mesma quantidade de padrões de cada classe).
  - Em alguns casos, a quantidade de dados é muito baixa, sendo necessário criar novos dados a partir dos dados originais existentes.

4) Preparação dos Dados

- **Tratamento e Limpeza**:
  - Trata dados com valores discrepantes
    - Ex: no atributo idade tem-se marcado 142 anos

OBRIGADO(A)

ser educacional
gente criando o futuro

UNINASSAU.DIGITAL UNINABUCO.DIGITAL UNAMA.DIGITAL UNG.DIGITAL UNIVERITAS.DIGITAL UNINORTE.DIGITAL UNIFACIMED.DIGITAL UNIFAEL