MINERAÇÃO DE DADOS

Prof. Me. Napoleão Póvoa Ribeiro Filho

UNITINS

GOVERNO DO TOCANTINS
TRABALHANDO E CUIDANDO DE TODOS

PRÉ-PROCESSAMENTO DE DADOS

- Responsável pela preparação dos dados brutos para que possam ser utilizados de forma eficiente por algoritmos de análise e modelagem.
- Inclui um conjunto de técnicas para: limpeza, integração, redução, transformação e discretização dos dados.

PROBLEMA COM OS DADOS

Nessa base de dados brutos (fonte ou atômicos), é possível notar algo “estranho”?

Nome	Idade	Nível Educacional	Estado Civil	Gênero	Cartão de Crédito	Renda Mensal($)
Roberto Felix	42	Especialização	Divorciado	M	Sim	5000
Joana Pereira	10	Doutorado	Viúva	F	Sim	6500
Isabela Assis	33	Graduação	Casada	F		3900
Marco Araújo	29	Graduação	89 kg	M	Não	3100

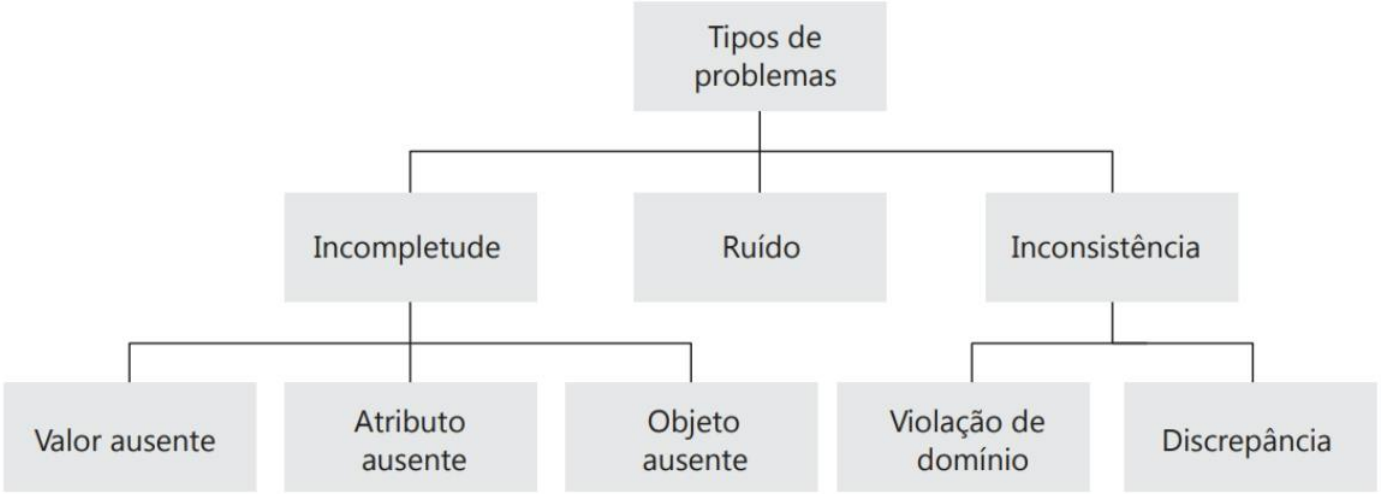
UNITINS
UNIVERSIDADE ESTADUAL DO TOCANTINS

PROBLEMA COM OS DADOS

- **Incompletude**: podem faltar valores de um atributo, como no caso do cartão de crédito de Isabela
- **Inconsistência**: ocorre quando diferentes e conflitantes versões do mesmo dado aparecem em locais variados. Por exemplo, a idade de Joana e o estado civil de Marco
- **Ruído**: aquele dado que apresenta alguma variação em relação ao seu valor sem ruído, podendo levar a inconsistências. Por exemplo, erro de digitação ou a pessoa mentir a idade ou salário

UNITINS
UNIVERSIDADE ESTADUAL DO TOCANTINS

PROBLEMA COM OS DADOS



UNITINS
UNIVERSIDADE ESTADUAL DO TOCANTINS

PROBLEMA COM OS DADOS

- Conhecer e preparar de forma adequada os dados para análise é uma etapa chamada de pré-processamento de dados e que pode tornar todo o processo de mineração muito mais eficiente e eficaz.
- Contudo, dados mal ou não pré-processados podem inviabilizar uma análise ou invalidar um resultado.

QUESTÕES IMPORTANTES

- Se existem dados ausentes, inconsistentes ou ruidosos, como tratá-los?
- É possível resumir a base de dados de forma que sejam obtidos resultados melhores no processo de mineração?
- Existem atributos que são mais relevantes que outros, ou até irrelevantes, para uma dada análise?
- Quais são os tipos de atributos da base? É preciso padronizá-los?
- Há atributos naturalmente inter-relacionados?

RESPONDENDO ESSAS QUESTÕES

- Ferramentas como histogramas, gráficos com distribuição de valores de um atributo, gráficos comparando atributos ou classes
- Especialistas de domínio podem esclarecer questões como ausências, inconsistências, significado de valores entre outras questões.

TIPOS DE DADOS

- Dados são valores quantitativos ou qualitativos associados a alguns atributos. Eles podem ser classificados em:
  - Estruturados
  - Semiestruturados
  - Não estruturados

UNITINS
UNIVERSIDADE ESTADUAL DO TOCANTINS

DADOS ESTRUTURADOS

- São organizados em um formato fixo e seguem um esquema bem definido, como tabelas em bancos de dados relacionais.
- Exemplo:
  - Banco de dados de um e-commerce, onde cada linha representa um pedido e contém colunas como ID do Cliente, Valor da Compra, Data e Forma de Pagamento.

DADOS SEMIESTRUTURADOS

- Possuem alguma organização, mas não seguem um modelo rígido como os dados estruturados.
- Geralmente, são armazenados em formatos como **JSON, XML, CSV** e podem ter elementos aninhados ou com estrutura variável.
- Exemplo:
  - Arquivos JSON contendo registros de transações financeiras, onde diferentes clientes podem ter campos distintos dependendo do tipo de transação

DADOS NÃO ESTRUTURADOS

- Não seguem um formato predefinido e geralmente são armazenados em grandes volumes.
- Eles exigem técnicas avançadas de processamento, como Processamento de Linguagem Natural (PLN) para textos e Visão Computacional para imagens e vídeos.

UNITINS
UNIVERSIDADE ESTADUAL DO TOCANTINS

DADOS NÃO ESTRUTURADOS

- Exemplo:
  - Postagens em redes sociais, contendo textos, imagens e vídeos sem uma estrutura padronizada.
  - Imagens médicas (raios X) utilizadas para diagnóstico automático, onde a informação está presente em pixels e não em tabelas estruturadas.

ATRIBUTOS

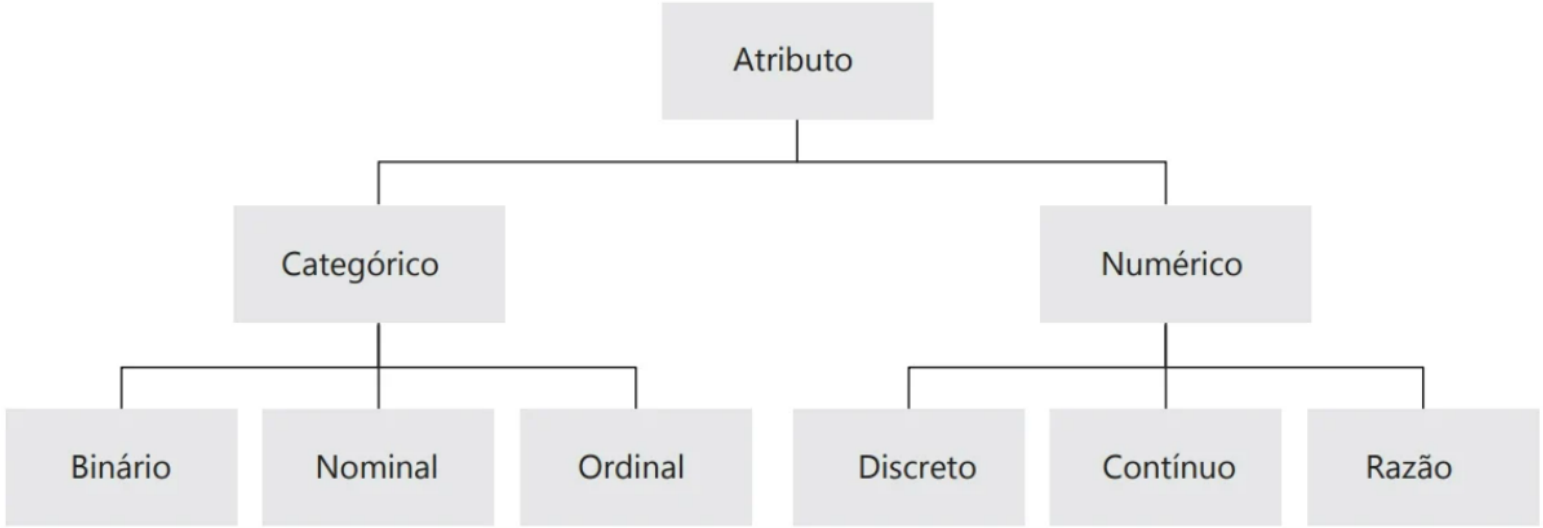
- Representa uma característica e, em mineração de dados, está organizada em colunas
- Podem ser classificados como independentes ou dependentes, com base no papel que desempenham na análise e modelagem preditiva.

ATRIBUTOS

Tipo de atributo	Definição	Exemplo
Independente	Variável de entrada, usada para prever um resultado.	Idade, Renda, Histórico de Crédito
Dependente	Variável de saída, que depende dos atributos independentes.	Status do Empréstimo (Aprovado/Negado)

Os atributos independentes são usados como preditores para estimar o atributo dependente, permitindo a construção de modelos de classificação ou regressão.

TIPOS DE ATRIBUTOS

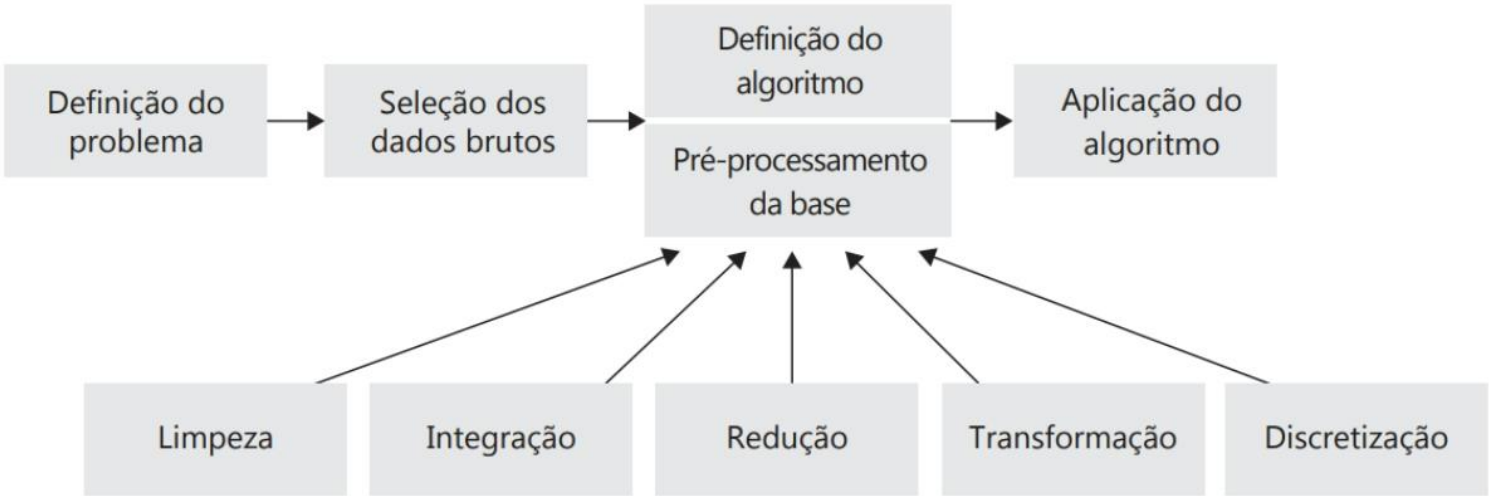


UNITINS
UNIVERSIDADE ESTADUAL DO TOCANTINS

TIPOS DE ATRIBUTOS

Tipo de atributo	Características	Exemplo
Binário	Apenas dois valores possíveis	Sim/Não, Masculino/Feminino
Nominal	Sem ordem definida	Cidade, Profissão
Ordinal	Tem ordem, mas sem distância mensurável	Satisfação do Cliente, Escolaridade
Discreto	Apenas valores inteiros	Número de Filhos, Chamadas Atendidas
Contínuo	Valores em intervalos contínuos	Altura, Peso
Razão	Possui um zero absoluto significativo	Salário, Distância

ETAPAS DO PROCESSO DE PREPARAÇÃO DA BASE DE DADOS



UNITINS
UNIVERSIDADE ESTADUAL DO TOCANTINS

LIMPEZA

Valores ausentes

- Podem ocorrer devido a falhas no registro, erros na coleta ou problemas técnicos.
- O valor a ser imputado não deve somar nem subtrair informação à base, ou seja, ele não deve enviesar a base
- A forma de tratá-los depende do contexto do problema e da natureza dos dados.

LIMPEZA

Estratégia	Descrição	Exemplo
Exclusão	Remove registros ou colunas com valores ausentes	Se uma coluna tem 80% de valores faltantes, podemos removê-la
Imputação Estatística	Substitui valores ausentes por média, mediana ou moda	Preencher Salário ausente com a média ou mediana dos salários
Criação de Categoria "Desconhecido"	Para variáveis categóricas, cria uma nova categoria para valores ausentes	Substituir Cidade ausente por "Desconhecido"
Interpolação ou Modelos Preditivos	Estima valores ausentes usando aprendizado de máquina ou interpolação	Prever Idade ausente com base em atributos como Renda
Algoritmos que Lidam com Ausências	Utiliza modelos como Random Forest e XGBoost, que podem trabalhar com dados incompletos	Treinar um modelo XGBoost sem preencher os valores ausentes

LEITURA

Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações

Capítulo 2: páginas 45 a 56

- 2.1 INTRODUÇÃO
- 2.2 O PROCESSO DE PREPARAÇÃO DA BASE DE DADOS
- 2.3 LIMPEZA DOS DADOS

UNITINS
UNIVERSIDADE ESTADUAL DO TOCANTINS

UNITINS

UNIVERSIDADE ESTADUAL DO TOCANTINS