CENTRO DE PESQUISA E FORMAÇÃO
SESC

Mineração de Dados

AULA 2

Pré-processamento de dados

Sobre o Curso

CENTRO DE PESQUISA E FORMAÇÃO SESC

Slides, Artigos, Materiais...

Quem leu?

MINERAÇÃO DE DADOS: UM OLHAR INSTIGANTE DE POSSIBILIDADES E APLICAÇÕES PARA ÓRGÃOS DA ADMINISTRAÇÃO PÚBLICA FEDERAL

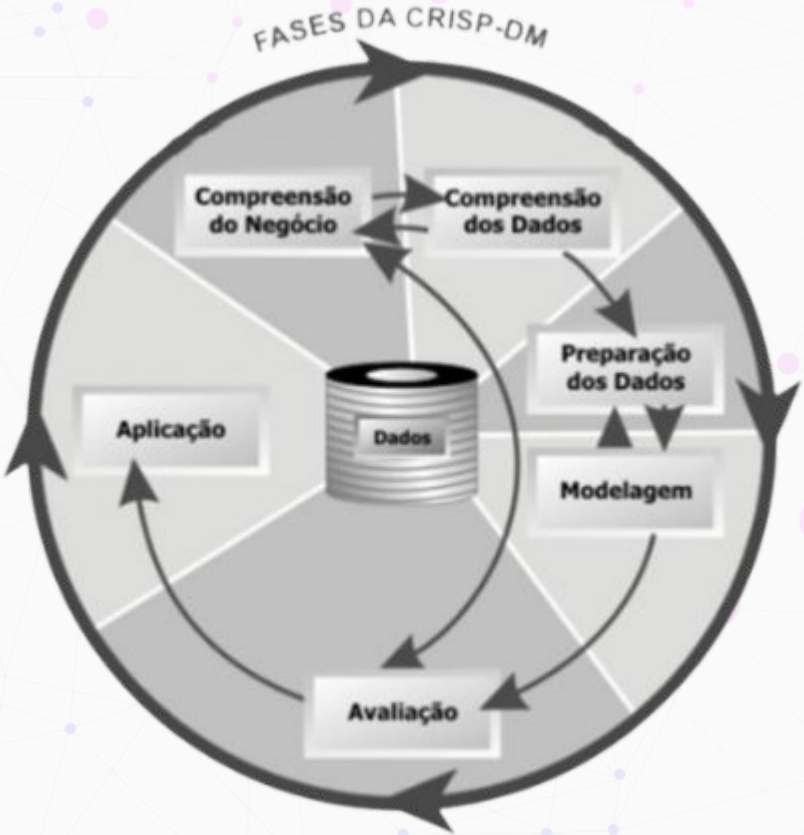
Roberto Rosa da Silveira Junior
Daniel Lins Rodriguez

https://doi.org/10.21874/rsp.v73.i3.5446

Sobre o artigo

- Um especialista em mineração de dados atua dentro de um contexto de negócio no qual geralmente ele não é especialista. Por esse motivo, é relevante a participação de um gestor ou alguém que tenha domínio e entendimento sobre os dados trabalhados.
- **CRISP – DM** (Cross-Industry Standard Process for Data Mining) pode ser considerado o padrão de maior aceitação para fases e atividades.

CRISP – DM



Fonte: Chapman et al. (2000).

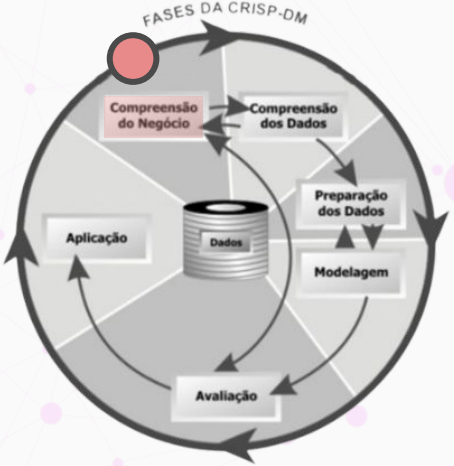
CRISP – DM

CENTRO DE PESQUISA E FORMAÇÃO SESC

Compreensão dos Negócios

Antes de minerar os dados, é essencial definir o que se quer alcançar. Essa etapa orienta todas as outras e garante que o processo esteja alinhado aos objetivos do negócio.

- Define o objetivo da mineração
- Alinha expectativas com metas organizacionais
- Fundamenta decisões nas etapas seguintes
- Envolve gestores e especialistas no problema

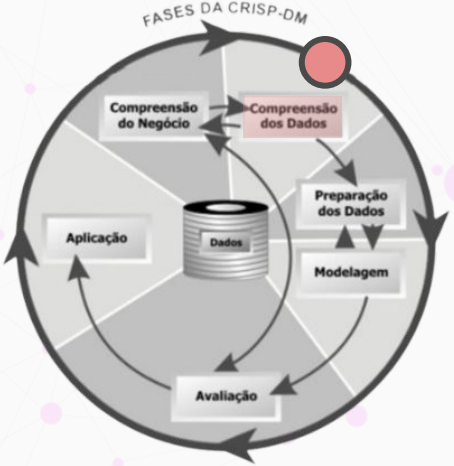


CRISP – DM

Compreensão dos Dados

Nesta fase, observa-se cuidadosamente os dados disponíveis. O foco é entender sua estrutura e identificar as variáveis realmente úteis para o problema.

- Analisa os dados com atenção
- Usa agrupamentos e visualização exploratória
- Seleciona variáveis relevantes
- Verifica se há dependência entre variáveis

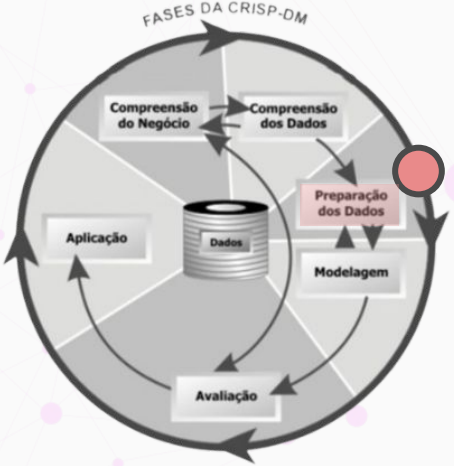


CRISP – DM

Preparação dos Dados

Os dados são tratados, integrados e organizados para garantir qualidade e compatibilidade com os métodos de mineração que serão aplicados depois.

- Integra dados de fontes diversas
- Filtra e limpa inconsistências
- Trata valores ausentes (missing)
- Normaliza e padroniza os dados

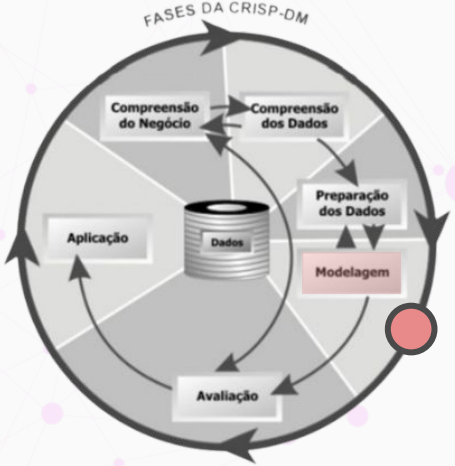


CRISP – DM

Modelagem

Aqui são escolhidos os algoritmos mais adequados. Eles são configurados e aplicados de acordo com os dados e com os objetivos definidos no início.

- Seleciona os algoritmos apropriados
- Ajusta parâmetros conforme os objetivos
- Aplica técnicas de mineração aos dados
- Fase central do processo analítico



CRISP – DM

Avaliação

Avaliamos se o modelo funciona bem. Métricas e testes ajudam a verificar a precisão dos resultados e garantir que o modelo seja útil e confiável.

- Valida o modelo com métricas específicas
- Envolve especialistas do domínio
- Usa testes como *cross-validation*
- Analisa precisão, erro e F-measure

CRISP – DM

CENTRO DE PESQUISA E FORMAÇÃO SESC

Deploy

Os resultados são apresentados de forma útil e aplicável. Essa etapa pode gerar apenas um relatório ou levar à automação de processos com base no modelo.

- Apresenta os resultados de forma clara
- Gera relatórios ou sistemas aplicáveis
- Apoia mudanças em processos e decisões
- Permite replicação em outras áreas

Elementos da mineração de dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

INSTÂNCIA/REGISTRO

ATRIBUTO

DESCRIÇÃO

Instância	Atributo 1	Atributo 2	Atributo 3	Atributo 4	Classe
1	10	0,1	P	1,9	baixo
2	20	?	P	1,8	médio
3	10	0,5	I	1,3	médio
4	30	0,1	I	1,4	?
5	40	0,6	K	1,2	alto
6	50	0,3	K	1,4	?
7	40	0,8	L	1,1	alto
8	30	0,1	P	1,4	médio
9	20	?	K	1,6	?
10	10	0,1	P	1,9	baixo

PREDIÇÃO

Técnicas de mineração de dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Aprendizado supervisionado

- técnicas abrangem um conjunto de dados que possuem uma variável alvo pré-definida
- instâncias são categorizadas em relação a essa variável pré-definida

exemplos: classificação e regressão

Aprendizado não supervisionado

- não precisam de uma categorização anterior ou prévia para as instâncias
- utilizam medida de similaridade entre os atributos
- exemplos: agrupamento e associação

Aplicações da Mineração de Dados na Administração Pública

A mineração de dados permite transformar grandes volumes de dados em conhecimento útil. Aplicações práticas já demonstraram seu potencial em diferentes esferas públicas e sociais.

- KDT: **extração de competências** em currículos não estruturados
- Detecção de **exportações fictícias** e lavagem de dinheiro
- **Fiscalização volante** com uso de dados de frotas públicas (PRV)
- Identificação de **alunos com risco de evasão** em AVAs
- Classificação de **risco de vida em pacientes** do SUS
- Análise automatizada com **SVM** em bases públicas
- Detecção de **bots e fraudes** em processos de pregões

Pré-processamento de dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Tudo começa com *raw data* (**dados brutos**)

Nome	Idade	Nível educacional	Estado civil	Gênero	Cartão de crédito	Renda mensal ($)
Roberto Felix	42	Especialização	Divorciado	M	Sim	5.000
Joana Pereira	10	Doutorado	Viúva	F	Sim	6.500
?	?	?	?	?	?	?
Isabela Assis	33	Graduação	Casada	F	?	3.900
Marco Araújo	29	Graduação	89 Kg	M	Não	3.100

Pré-processamento de dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Tudo começa com *raw data* (**dados brutos**)

Nome	Idade	Nível educacional	Estado civil	Gênero	Cartão de crédito	Renda mensal ($)
Roberto Felix	42	Especialização	Divorciado	M	Sim	5.000
Joana Pereira	10	Doutorado	Viúva	F	Sim	6.500
?	?	?	?	?	?	?
Isabela Assis	33	Graduação	Casada	F	?	3.900
Marco Araújo	29	Graduação	89 Kg	M	Não	3.100

Pré-processamento de dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Incompletude

Bases de dados podem estar incompletas por vários motivos. Nem sempre é fácil perceber essas falhas, sendo necessário conhecimento de domínio para identificá-las corretamente.

- Falta de valor em um atributo (ex: “?” no cartão de crédito)
- Ausência de um atributo inteiro (ex: dia da semana)
- Ausência de um objeto (ex: linha em branco na tabela)
- Incompletudes nem sempre são visíveis de imediato
- Especialistas podem detectar lacunas relevantes (ex: aluno ausente da chamada)

Pré-processamento de dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Inconsistência

Ocorrem quando há conflitos ou valores fora do padrão esperado. Podem surgir por erros de entrada, formatos distintos ou dados que não fazem sentido no contexto.

- Versões conflitantes de um mesmo dado em locais distintos
- Valores fora do domínio esperado do atributo
- Discrepâncias entre atributos (ex: idade x grau acadêmico)
- Incompatibilidades de unidade (ex: kg vs £, m vs km)
- Exemplo: idade incompatível com título de Doutorado
- Exemplo: estado civil incoerente com outros dados

Pré-processamento de dados

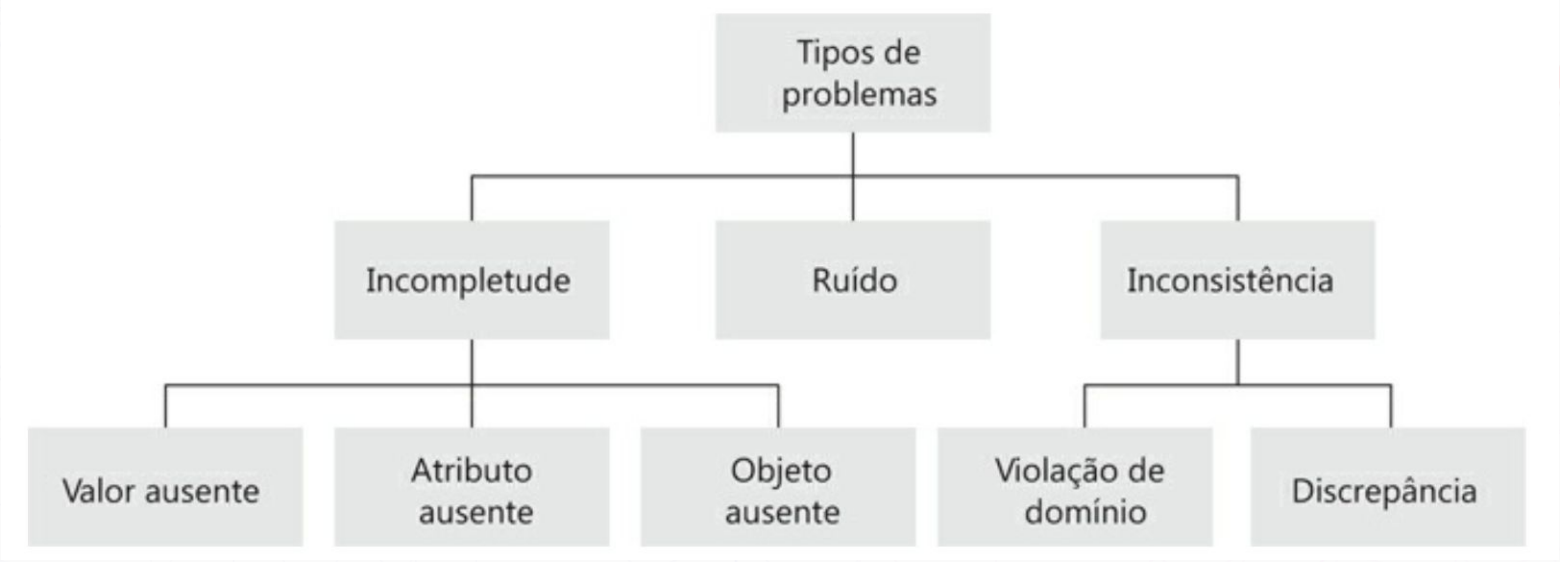
CENTRO DE PESQUISA E FORMAÇÃO SESC

Ruído

Em dados, ruído representa variações inesperadas ou inexplicáveis. Pode distorcer análises e gerar inconsistências, sendo difícil de detectar quando está em níveis baixos.

- Variações indesejadas e sem explicação clara
- Relaciona-se ao ruído estatístico ou de sinais
- Pode afetar a consistência e a qualidade dos dados
- Um dado ruidoso difere do valor real esperado
- Nem sempre é possível identificar ruídos sutis
- Pode comprometer a análise se não for tratado

Pré-processamento de dados



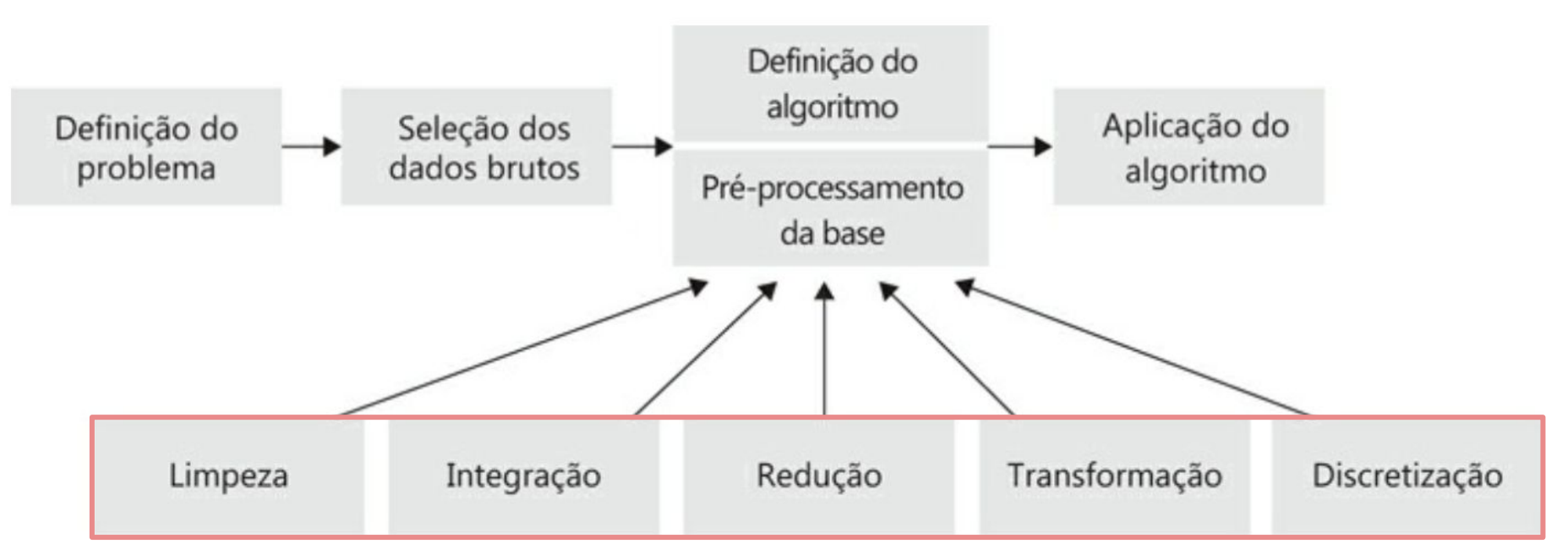
Fonte: de Castro, Ferrari (2016)

Pré-processamento de dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

- 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?

Pré-processamento de dados



Fonte: de Castro, Ferrari (2016)

Limpeza dos Dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Valores ausentes

Valores ausentes são lacunas em atributos ou objetos de uma base. Podem prejudicar a análise e exigem tratamento adequado, como imputação ou remoção.

- Representações comuns: “?”, branco, código específico
- Indicam ausência ou não observação de um valor
- Podem afetar algoritmos que não lidam com lacunas
- Imputação: técnica para estimar e preencher valores
- Algoritmos podem falhar com dados incompletos
- Eliminar registros pode descartar informações valiosas

Limpeza dos Dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Valores ausentes

A ausência de dados pode ocorrer de modo aleatório ou não, o que influencia o método de tratamento mais adequado.

- **MCAR**: ausência totalmente aleatória (ex: erro de digitação)
- **MAR**: ausência depende de dados observados (ex: idade declarada)
- **NMAR**: ausência depende do valor não informado (ex: salário)

Limpeza dos Dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Valores ausentes

Métodos de imputação:

- **Ignorar o objeto:** remove registros incompletos (pode reduzir a base)
- **Imputação manual:** valor escolhido empiricamente (pouco recomendável)
- **Constante global:** usa um valor fixo (risco de enviesar o modelo)
- **Hot-deck:** usa valor de um objeto similar aleatório
- **Última observação:** repete valor anterior (last observation carried forward)
- **Média ou moda:** usa média (numéricos) ou moda (nominais) da base

Limpeza dos Dados

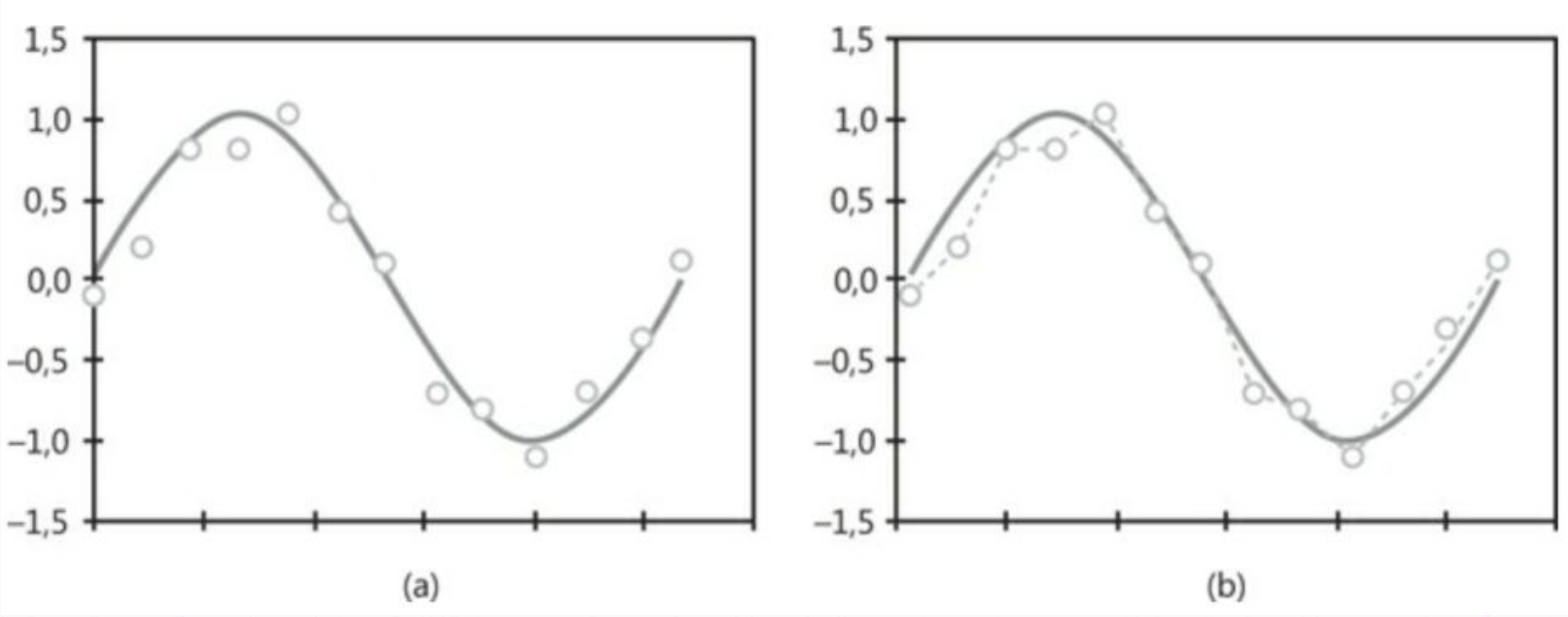
CENTRO DE PESQUISA E FORMAÇÃO SESC

Dados Ruidosos

Erros de entrada e medição geram ruído nos dados. Mesmo com técnicas de correção, parte desse ruído é inevitável e precisa ser tratado com cautela pelos algoritmos.

- Ruído = distorção acumulada nos dados reais
- Origem: erro humano, falha técnica ou aleatoriedade
- Nem sempre é possível detectar ou corrigir o ruído
- Ruídos não seguem um padrão fixo de ocorrência
- Algoritmos devem aprender padrões sem aprender o ruído
- O impacto pode ser apenas minimizado — não eliminado

Limpeza dos Dados



Fonte: de Castro, Ferrari (2016)

Limpeza dos Dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

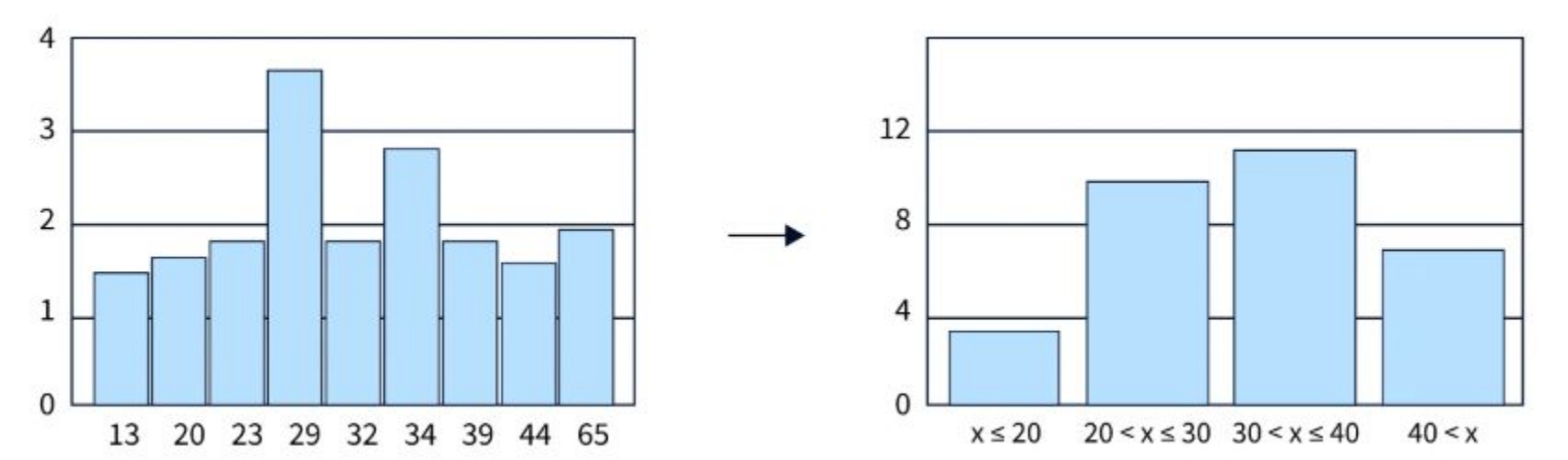
Encaixotamento (binning)

O encaixotamento transforma valores numéricos contínuos em categorias discretas (caixas). Isso reduz a variabilidade e facilita a análise ou visualização de padrões.

- Agrupa valores em faixas chamadas "caixas" (bins)
- Mesma largura: caixas com intervalos iguais
- Mesma frequência: mesma quantidade de valores por caixa
- Passos: ordenar valores, definir nº de caixas, representar a caixa
- Valores podem ser substituídos por:
  - Média ou moda da caixa
  - Extremos mais próximos

Limpeza dos Dados

Encaixotamento (binning)



Limpeza dos Dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Agrupamento

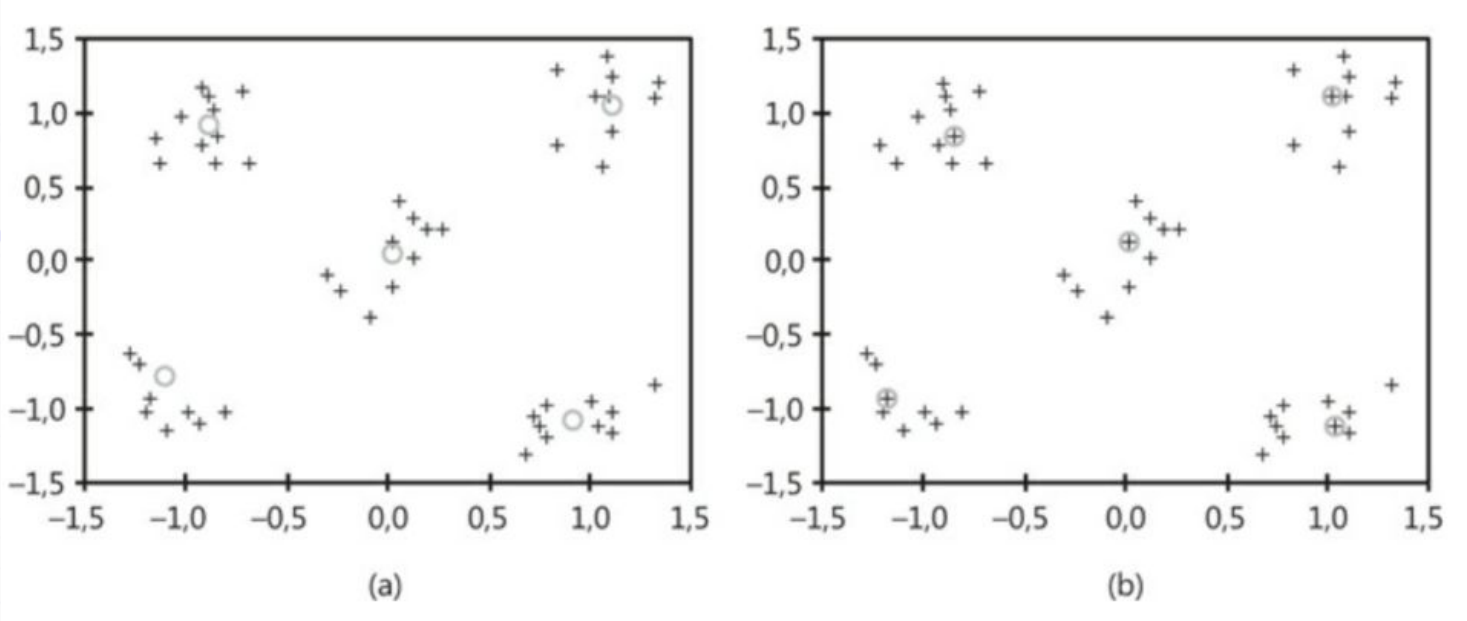
O agrupamento identifica grupos de objetos semelhantes entre si e distintos dos demais. É uma técnica de aprendizado não supervisionado usada para rotular automaticamente os dados.

- Descobre grupos de objetos similares automaticamente
- Objetos em um grupo são mais parecidos entre si
- Cada grupo pode ser representado por:
  - **Centroide**: valor médio dos objetos do grupo (calculado)
  - **Medoide**: objeto mais central do grupo (real)
- Considera todos os atributos da base simultaneamente
- Suaviza os dados como um todo (não atributo a atributo)

Limpeza dos Dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Agrupamento



Centroide **Medoide**

Fonte: de Castro, Ferrari (2016)

Limpeza dos Dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

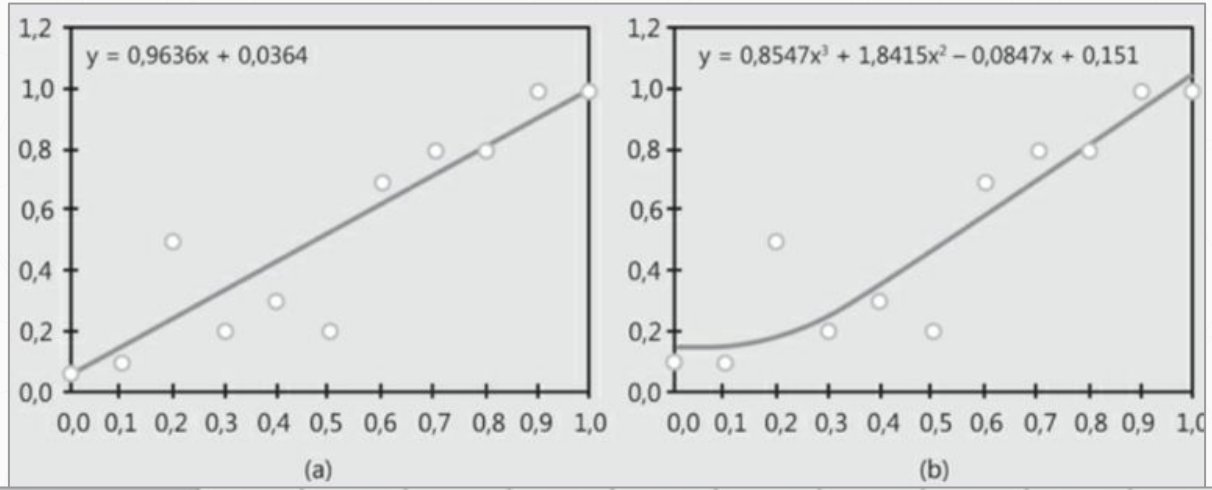
Aproximação

A aproximação suaviza os dados substituindo valores reais por valores gerados por funções matemáticas, como modelos polinomiais. É útil para reduzir ruído e simplificar padrões.

- Usa funções para representar tendências nos dados
- Exemplo: polinômios de grau 1 (reta) ou grau 3
- Modelo paramétrico: definido por seus coeficientes
- Etapas:
  - Escolher o tipo de função (reta, curva, etc.)
  - Ajustar o modelo aos dados
  - Substituir valores reais pelos valores da função
- Reduz a variabilidade e facilita análises preditivas

Limpeza dos Dados

Aproximação



Valor real	0,10	0,10	0,50	0,20	0,30	0,20	0,70	0,80	0,80	1,00	1,00
Valor suavizado (a)	0,04	0,13	0,23	0,33	0,42	0,52	0,61	0,71	0,81	0,90	1,00
Valor suavizado (b)	0,15	0,16	0,20	0,27	0,36	0,46	0,58	0,70	0,82	0,94	1,05

Fonte: de Castro, Ferrari (2016)

Integração dos Dados

Antes da mineração, é necessário integrar dados de diferentes fontes em uma base única. Essa tarefa é essencial, mas pode gerar desafios técnicos e de padronização.

Dados podem vir de sistemas, lojas ou arquivos distintos

Integração reúne tudo em uma base unificada para análise

Desafios comuns:

- Formatos de armazenagem diferentes
- Conflitos em datas e convenções
- Chaves de acesso incompatíveis
- Falta de padronização nos atributos
- Fundamental para garantir consistência na análise

Integração dos Dados

CENTRO DE PESQUISA E FORMAÇÃO
SESC

Redundância de Dados

Redundância ocorre quando a mesma informação aparece mais de uma vez ou pode ser inferida de outras variáveis, gerando complexidade e potencial distorção nas análises.

- Mesmo dado presente em diferentes locais da base
- Atributos podem ser derivados de outros
  - Ex: idade ↔ data de nascimento
  - Ex: total ↔ preço × quantidade
- Nem todos os atributos são necessários para análise
- Pode ser detectada com análise de correlação
- Contribui para a redução da base de dados

Integração dos Dados

CENTRO DE PESQUISA E FORMAÇÃO
SESC

Duplicidade

Duplicidade é um tipo específico de redundância em que objetos ou registros se repetem. Pode distorcer análises se não for tratada, mas em alguns contextos tem utilidade.

- Dados ou registros repetidos na base
- Pode causar anomalias ou distorções
- Prevenção por normalização da base (relacionamento de tabelas)
- Exemplo: mesmo cliente cadastrado duas vezes
- Útil em backups e verificação de consistência

Integração dos Dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Conflitos de Dados

Conflitos ocorrem quando há discrepâncias entre fontes sobre o mesmo dado. Frequentemente causados por diferentes unidades, formatos ou escalas, exigem padronização.

- Mesma entidade com valores diferentes em fontes distintas
- Ex: peso em kg vs. libras, distância em km vs. milhas
- Origem: diferentes escalas, formatos ou codificações
- Causam confusão e comprometem a integração
- Precisam ser tratados na etapa de padronização

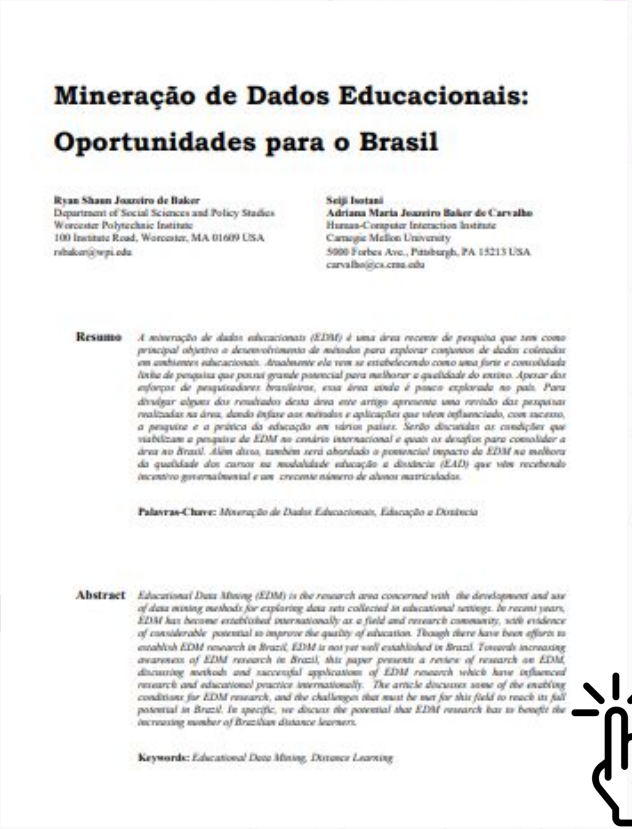
Redução dos Dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Bases muito grandes podem tornar os algoritmos de mineração ineficientes. Reduzir o número de objetos ou atributos ajuda a acelerar o processo, sem comprometer a qualidade da análise.

- Grandes volumes aumentam a complexidade computacional
- Muitos atributos → dados esparsos e instabilidade numérica
- **Redução pode ser feita em:**
  - Número de objetos (amostragem)
  - Número de atributos (redução de dimensionalidade)
- Ex: redes sociais, fraudes em cartões, perfis de clientes
- Técnicas devem manter a integridade dos dados
- Eficiência computacional sem perda significativa de eficácia

Leitura deste módulo



Mineração de Dados Educacionais:
Oportunidades para o Brasil

Ryan Shaun Joazeiro de Baker
Department of Social Sciences and Policy Studies
Worcester Polytechnic Institute
100 Institute Road, Worcester, MA 01609 USA
rsbaker@wpi.edu

Seiji Isotani
Adriana Maria Joazeiro Baker de Carvalho
Human-Computer Interaction Institute
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213 USA
carvalho@cs.cmu.edu

Resumo

A mineração de dados educacionais (EDM) é uma área recente de pesquisa que tem como principal objetivo o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais. Atualmente ela vem se estabelecendo como uma forte e consolidada linha de pesquisa que possui grande potencial para melhorar a qualidade do ensino. Apesar dos esforços de pesquisadores brasileiros, essa área ainda é pouco explorada no país. Para divulgar alguns dos resultados desta área este artigo apresenta uma revisão das pesquisas realizadas na área, dando ênfase aos métodos e aplicações que vêem influenciado, com sucesso, a pesquisa e a prática da educação em vários países. Serão discutidas as condições que viabilizam a pesquisa de EDM no cenário internacional e quais os desafios para consolidar a área no Brasil. Além disso, também será abordado o potencial impacto de EDM na melhoria da qualidade dos cursos na modalidade educação a distância (EAD) que vêm recebendo incentivo governamental e um crescente número de alunos matriculados.

Palavras-Chave: *Mineração de Dados Educacionais, Educação a Distância*

Abstract

Educational Data Mining (EDM) is the research area concerned with the development and use of data mining methods for exploring data sets collected in educational settings. In recent years, EDM has become established internationally as a field and research community, with evidence of considerable potential to improve the quality of education. Though there have been efforts to establish EDM research in Brazil, EDM is not yet well established in Brazil. Towards increasing awareness of EDM research in Brazil, this paper presents a review of research on EDM, discussing methods and successful applications of EDM research which have influenced research and educational practice internationally. The article discusses some of the enabling conditions for EDM research, and the challenges that must be met for this field to reach its full potential in Brazil. In specific, we discuss the potential that EDM research has to benefit the increasing number of Brazilian distance learners.

Keywords: *Educational Data Mining, Distance Learning*

Mineração de Dados Educacionais: Oportunidades para o Brasil

Ryan Shaun Joazeiro de Baker
Seiji Isotani
Adriana Maria Joazeiro Baker de Carvalho

https://repositorio.usp.br/item/002207788

CENTRO DE PESQUISA E FORMAÇÃO
SESC

OBRIGADO!

Até a próxima aula!

CENTRO DE PESQUISA E FORMAÇÃO
SESC

Hora da pausa! Voltamos em:

<<20:00->>