CENTRO DE PESQUISA E FORMAÇÃO
SESC

Mineração de Dados

AULA 3

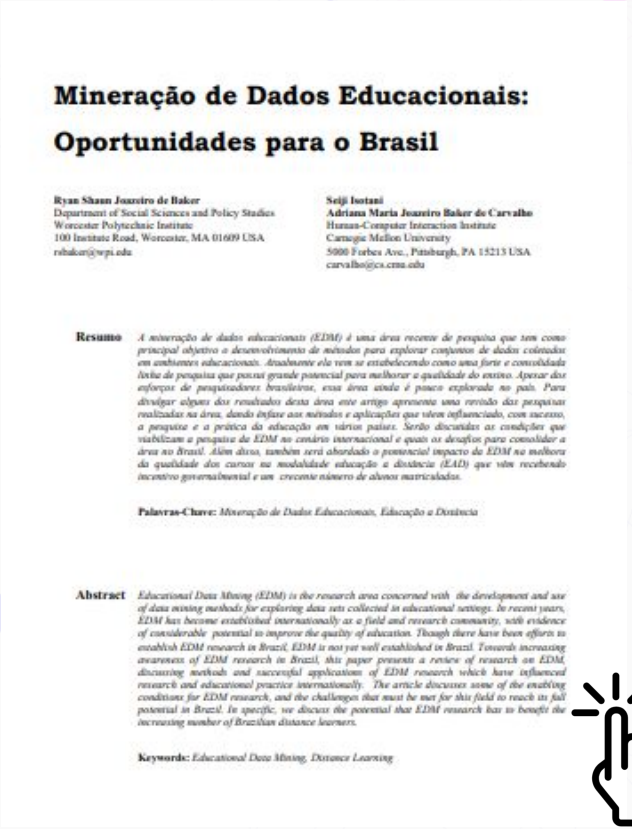
Algoritmos e Validação

Sobre o Curso

CENTRO DE PESQUISA E FORMAÇÃO SESC

Slides, Artigos, Materiais...

Quem leu?



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

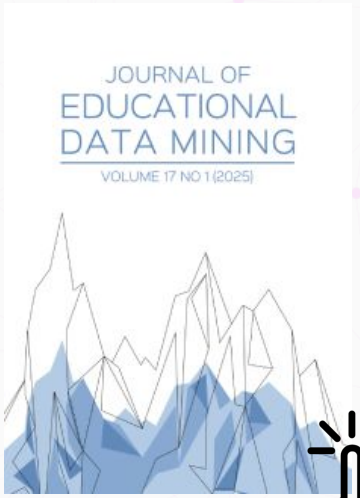
https://repositorio.usp.br/item/002207788

Sobre o artigo

CENTRO DE PESQUISA E FORMAÇÃO SESC

Educational Data Mining Conference

Journal of Educational Data Mining



Sobre o artigo

Principais sub-áreas de pesquisa em EDM:

Predição (Prediction)
- Classificação (Classification)
- Regressão (Regression)
- Estimação de Densidade (Density Estimation)

Agrupamento (Clustering)

Mineração de relações (Relationship Mining)
- Mineração de Regras de associação (Association Rule Mining)
- Mineração de Correlações (Correlation
- Mining)
- Mineração de Padrões Sequenciais (Sequential Pattern Mining)
- Mineração de Causas (Causal Mining)

Destilação de dados para facilitar decisões humanas (Distillation of Data for Human Judgment)

Descobrimento com modelos (Discovery with Models)

Sobre o artigo

🔓 Dados educacionais abertos impulsionam a pesquisa

💡 Redução de barreiras na pesquisa educacional elimina etapas onerosas

🌐 Crescimento da EAD amplia as possibilidades

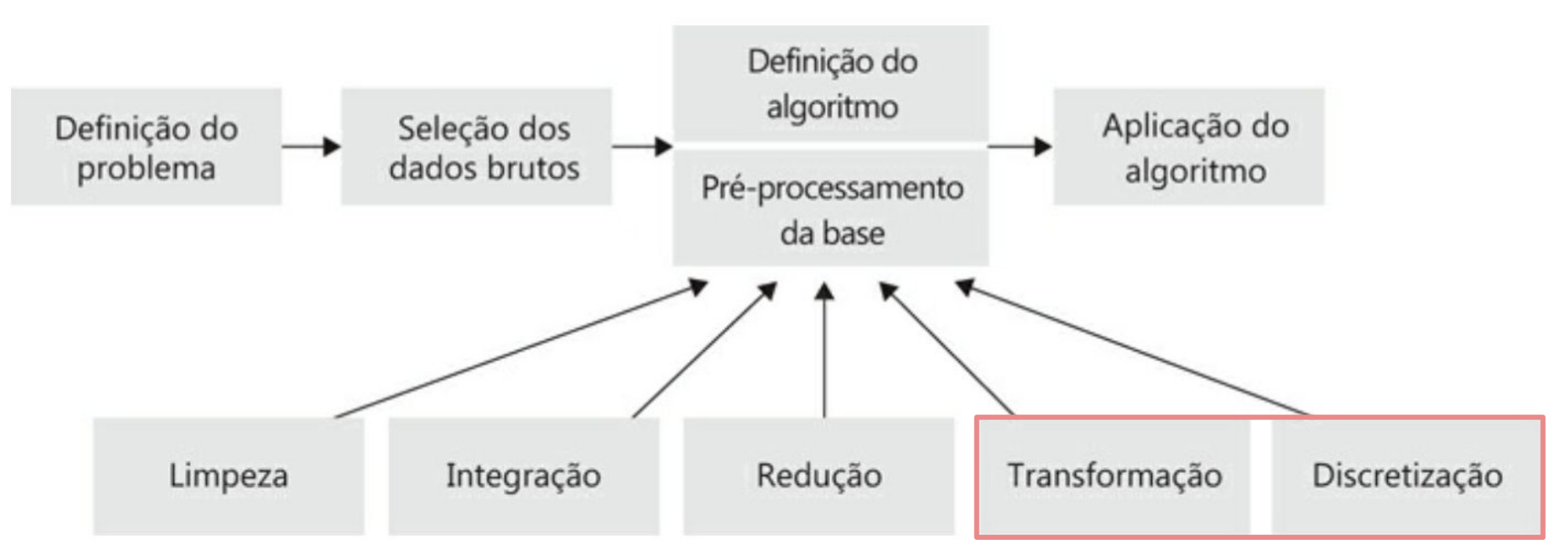
🧪 Potencial para acelerar descobertas

Mais dados = mais oportunidades de identificar padrões, testar hipóteses e melhorar práticas pedagógicas.

Dica de livro!

CENTRO DE PESQUISA E FORMAÇÃO
SESC

Pré-processamento de dados



Fonte: de Castro, Ferrari (2016)

Transformação dos Dados

Bases de dados integradas ou brutas podem ter formatos inconsistentes. A transformação dos dados **prepara os atributos para que se tornem compatíveis com os algoritmos de mineração**.

- Resolve problemas de padronização (maiúsculas, unidades, etc.)
- Trata inconsistências entre dados integrados
- Uniformiza formatos de atributos (ex: numéricos vs. categóricos)
- Reduz ruídos e garante compatibilidade entre variáveis
- Adapta os dados às exigências dos algoritmos de mineração
- Pode envolver normalização, codificação e conversão de formatos

Transformação dos Dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Padronização dos Dados

Padronizar dados é essencial para evitar erros causados por diferenças de formato, capitalização ou unidades. Isso garante consistência para a análise e integração de dados.

- **Capitalização**: padronizar para maiúsculas evita erros com letras
- **Caracteres especiais**: remover ou substituir acentos em atributos nominais
- **Formatos**: unificar padrões de datas, CPFs, documentos etc.
- **Unidades**: converter medidas para uma unidade comum (ex: km ↔ milhas)
- Essencial ao integrar bases de origens diversas
- Previne falhas em ferramentas sensíveis a variações nos dados

Transformação dos Dados

CENTRO DE PESQUISA E FORMAÇÃO
SESC

Normalização dos Dados

A normalização transforma os dados para que todos os atributos fiquem na mesma escala, facilitando o uso de algoritmos como redes neurais e métodos baseados em distância.

- Evita a saturação em redes neurais artificiais
- Garante que atributos tenham o mesmo domínio de valores
- Fundamental para algoritmos sensíveis à escala dos dados
- Exemplos comuns de normalização:
  - **Max-Min**: escala valores para um intervalo fixo (ex: 0 a 1)
  - **Escore-z**: transforma dados para média 0 e desvio padrão 1
  - **Escalonamento decimal**: ajusta pela potência de 10 mais próxima
  - **Range interquartil**: baseia-se na dispersão central dos dados

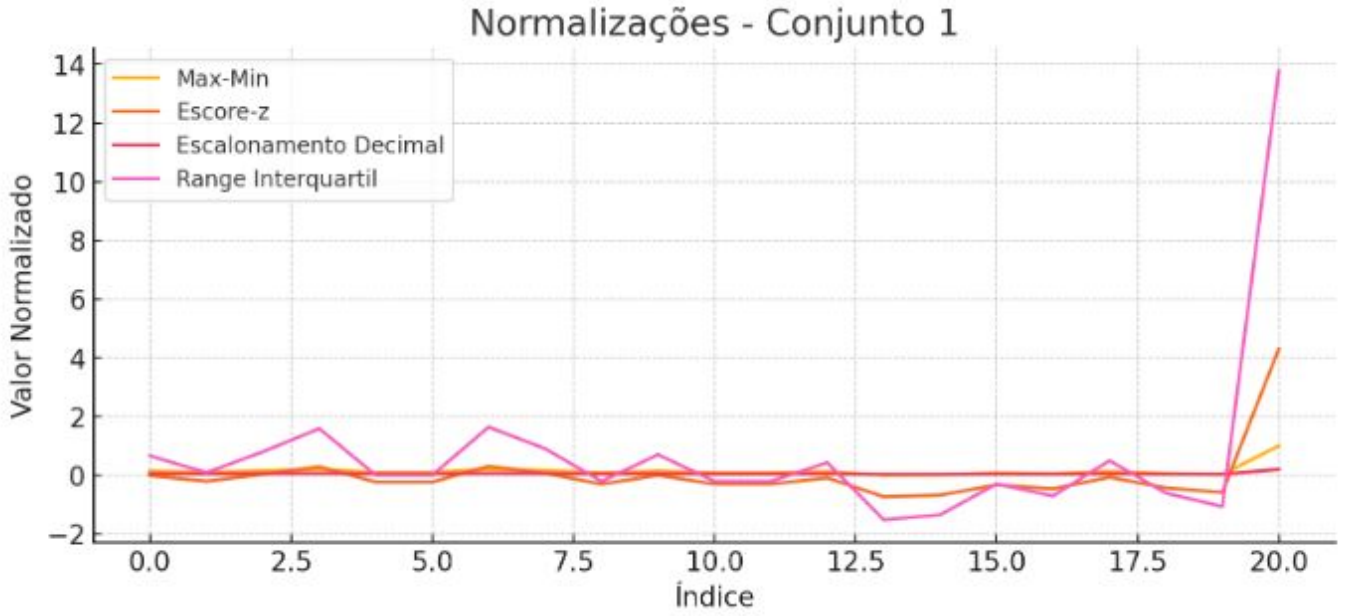
Transformação dos Dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Conjunto 1

python Copiar Editar

```
[55, 49, 56, 65, 48, 48, 66, 58, 45, 55, 45, 45, 52, 31, 33, 44, 40, 53, 41, 36, 200]
```



Transformação dos Dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Max-Min

python Copiar código

```
[0.142, 0.105, 0.151, 0.203, 0.099, 0.099, 0.206, 0.158, 0.085, 0.145,
 0.086, 0.086, 0.127, 0.0, 0.011, 0.08, 0.053, 0.132, 0.059, 0.03, 1.0]
```

Escore-z (Z-score)

python Copiar código

```
[-0.016, -0.205, 0.029, 0.289, -0.234, -0.234, 0.306, 0.064, -0.304,
 -0.003, -0.302, -0.303, -0.092, -0.734, -0.678, -0.332, -0.466, -0.071,
 -0.435, -0.585, 4.304]
```

Escalonamento Decimal

python Copiar código

```
[0.055, 0.049, 0.056, 0.065, 0.048, 0.048, 0.066, 0.058, 0.045, 0.055,
 0.045, 0.045, 0.052, 0.031, 0.033, 0.044, 0.04, 0.053, 0.041, 0.036, 0.2]
```

Range Interquartil (IQR)

python Copiar código

```
[0.661, 0.087, 0.798, 1.59, -0.0, 0.0, 1.641, 0.907, -0.213, 0.703,
 -0.208, -0.21, 0.431, -1.52, -1.349, -0.297, -0.705, 0.496, -0.61,
 -1.066, 13.788]
```

Discretização dos Dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Discretização **converte atributos numéricos contínuos em categorias**. É útil quando algoritmos exigem variáveis categóricas ou para simplificar a análise.

- Torna atributos contínuos compatíveis com algoritmos categóricos
- Reduz a complexidade ao diminuir o número de valores únicos
- Pode ser feita por divisão do domínio em intervalos fixos
- Métodos comuns de discretização:
  - Encaixotamento (binning): substitui valores por médias ou extremos
  - Histograma: usa faixas para definir os grupos
  - Agrupamento: segmenta os valores por similaridade
  - Baseada em entropia: maximiza a pureza dos intervalos

Análise Descritiva de Dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Antes de aplicar técnicas mais complexas, é essencial entender os dados. A análise descritiva permite explorar, resumir e visualizar informações para conhecer a base de dados.

- Permite compreender a estrutura e distribuição dos dados
- É útil especialmente quando se desconhece o domínio dos dados
- Trabalha com análises univariadas, bivariadas e, ocasionalmente, trivariadas
- Usa técnicas de sumarização numérica e visual para interpretar padrões

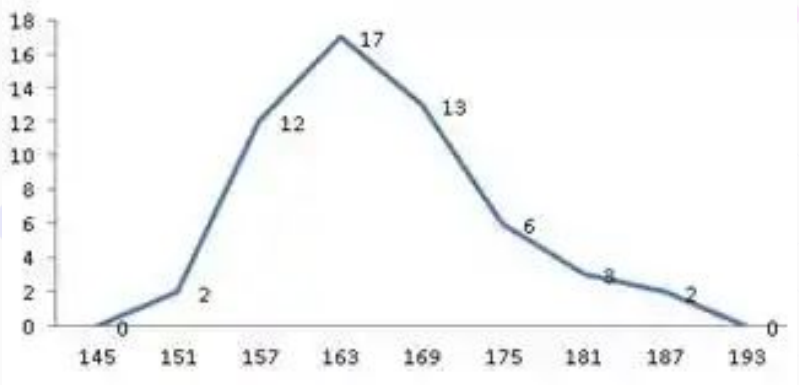
			
DESCRITIVA O que aconteceu?	**DIAGNÓSTICA** Por que aconteceu?	**PREDITIVA** O que é provável de acontecer?	**PRESCRITIVA** Qual decisão tomar?

Análise Descritiva de Dados

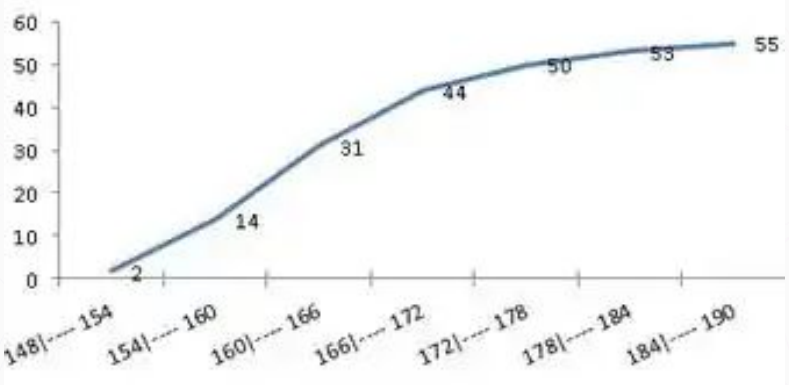
Tipos de Gráficos para Dados Univariados

Gráficos são aliados importantes para visualizar dados de maneira intuitiva, revelando rapidamente padrões e concentrações.

Polígono de frequências



Ogiva

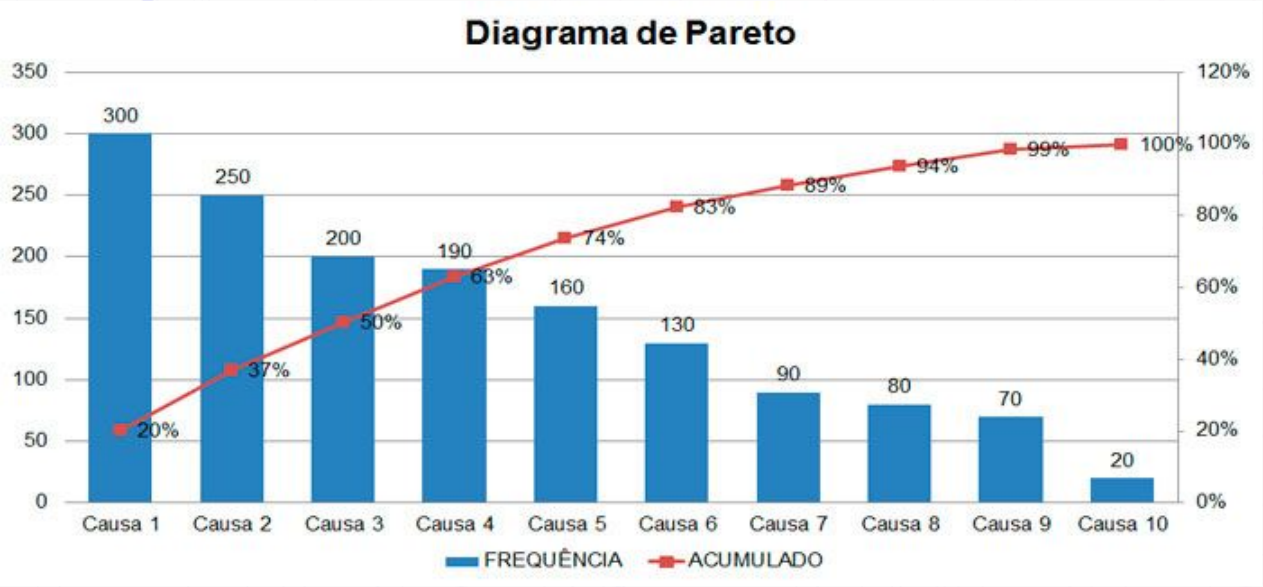


Análise Descritiva de Dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Tipos de Gráficos para Dados Univariados

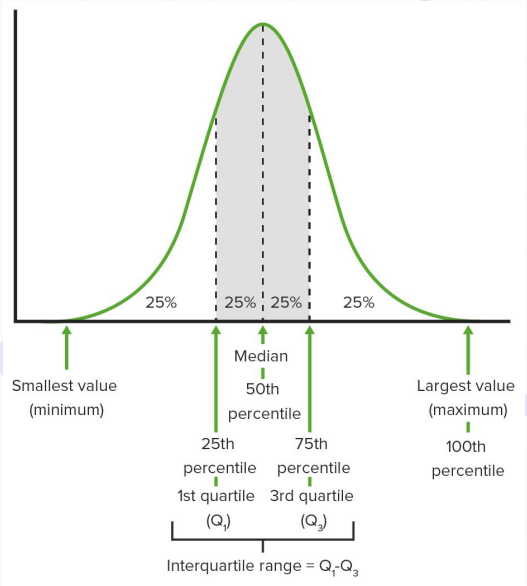
Gráficos são aliados importantes para visualizar dados de maneira intuitiva, revelando rapidamente padrões e concentrações.

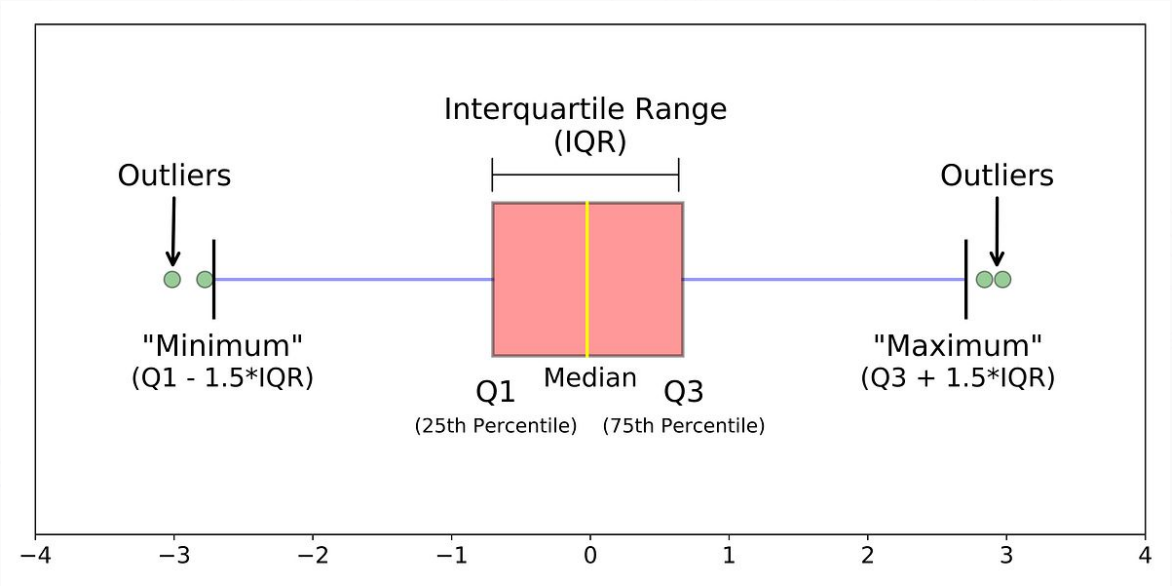


Análise Descritiva de Dados

Medidas de Tendência Central e Dispersão

Essas medidas numéricas ajudam a resumir os dados com valores centrais e entender a variabilidade presente.



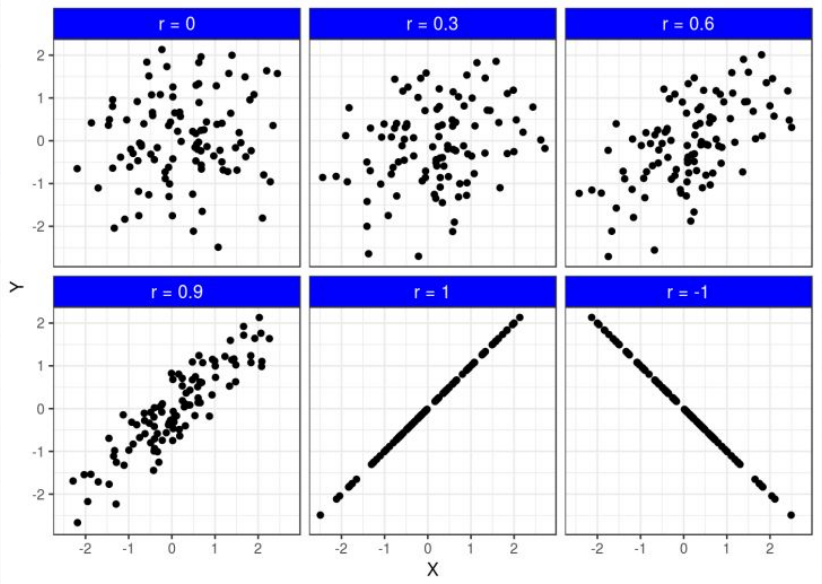


Análise Descritiva de Dados

CENTRO DE PESQUISA E FORMAÇÃO
SESC

Medidas de Associação e Correlação

Ao analisar dois atributos, queremos saber se há uma relação entre eles. Correlações ajudam a identificar padrões conjuntos.



Classificação de Dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Classificação é uma tarefa de aprendizado supervisionado que visa prever a classe ou categoria de um objeto com base em atributos conhecidos.

- Utiliza **dados históricos** com rótulo de classe (atributo alvo)
- Permite **prever categorias** como “adimplente” ou “inadimplente”
- É um tipo de predição para atributos discretos
- Subdivide objetos em subconjuntos com características similares

Classificação de Dados

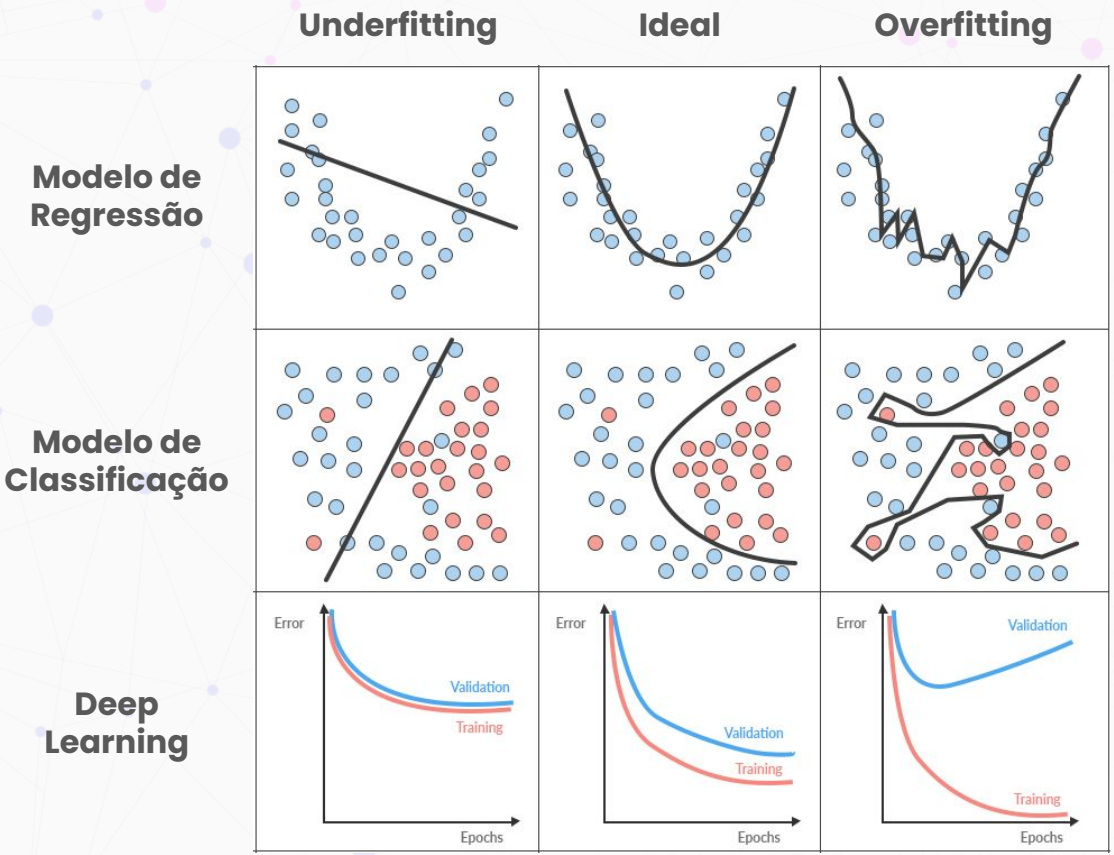
CENTRO DE PESQUISA E FORMAÇÃO SESC

Construção do Modelo Preditivo

O modelo preditivo é criado a partir de dados rotulados e passa por etapas de treino e teste para verificar sua eficácia.

- **Treinamento**: cria o modelo usando dados com classe conhecida
- **Teste**: avalia a capacidade do modelo em dados desconhecidos
- Dois tipos de erro comuns: **viés** (bias) e **variância**
- O ideal é encontrar **equilíbrio entre underfitting e overfitting**

Classificação de Dados



Classificação de Dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Validação Cruzada e Técnicas de Avaliação

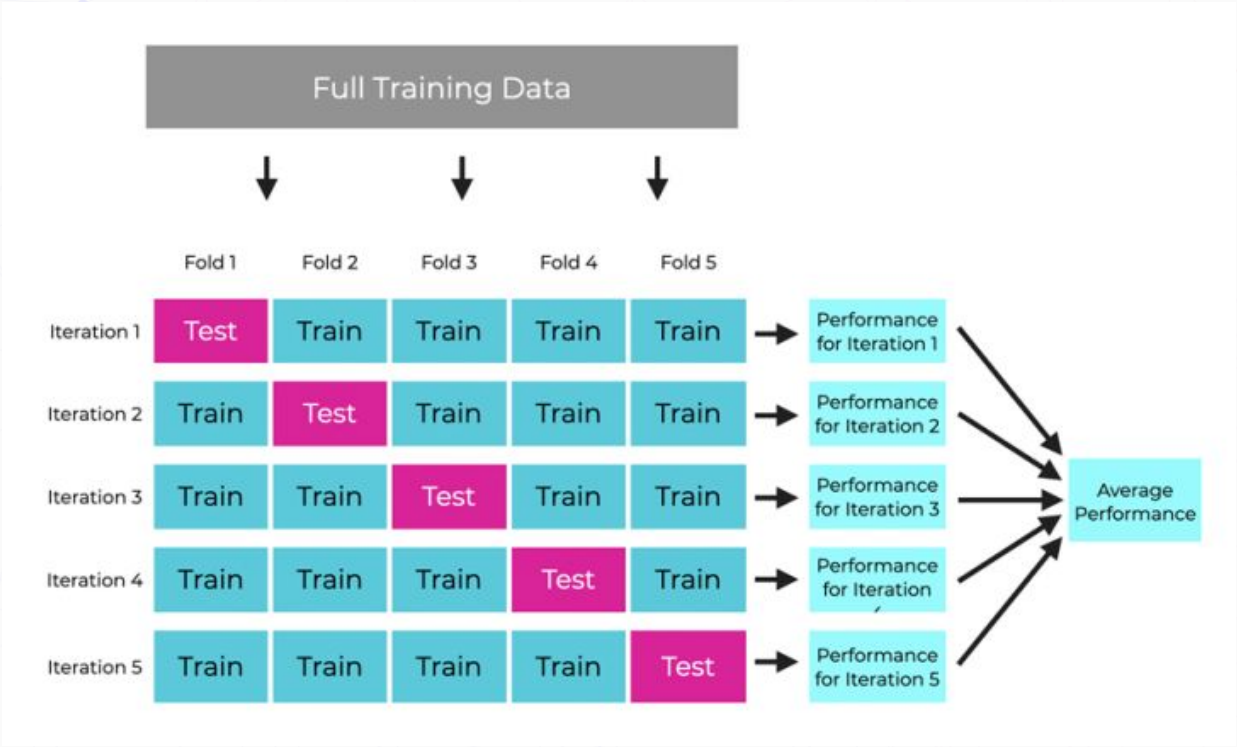
A validação cruzada aumenta a confiabilidade do modelo ao testá-lo diversas vezes em diferentes subconjuntos da base.

- **k-fold cross-validation**: divide os dados em k pastas, alternando treino e teste
- **Leave-one-out**: útil para bases pequenas; testa um objeto por vez
- Ajuda a reduzir variabilidade e evita sobreajuste
- Executa múltiplas iterações para estabilizar os resultados

Classificação de Dados

CENTRO DE PESQUISA E FORMAÇÃO
SESC

K-FOLD



Classificação de Dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Classificação de Dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Medidas de Avaliação do Classificador

A qualidade de um classificador é medida com base nos acertos e erros obtidos durante a fase de teste, resumidos em métricas objetivas.

- **Matriz de confusão:** organiza os acertos e erros por classe
- **Acurácia:** taxa global de acertos
- **Taxa de verdadeiros/falsos positivos:** foco nas decisões críticas
- **Erro:** complemento da acurácia, mede os equívocos do modelo

Classificação de Dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

		Detectada	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Classificação de Dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

previsto / real	gato	rato	cachorro
gato	10	2	3
rato	5	14	1
cachorro	1	2	12

Classificação de Dados

CENTRO DE PESQUISA E FORMAÇÃO SESC

Precisão, Revocação e F1-Score

Em classificações onde o erro pode ter impacto desigual, usamos métricas que consideram relevância e recuperação dos dados.

- **Precisão:** % de acertos entre os classificados como positivos
- **Revocação (recall):** % de positivos que foram corretamente classificados
- **F1-score:** equilíbrio entre precisão e revocação
- Úteis em contextos críticos como detecção de fraudes ou diagnósticos

Algoritmos de Classificação

CENTRO DE PESQUISA E FORMAÇÃO SESC

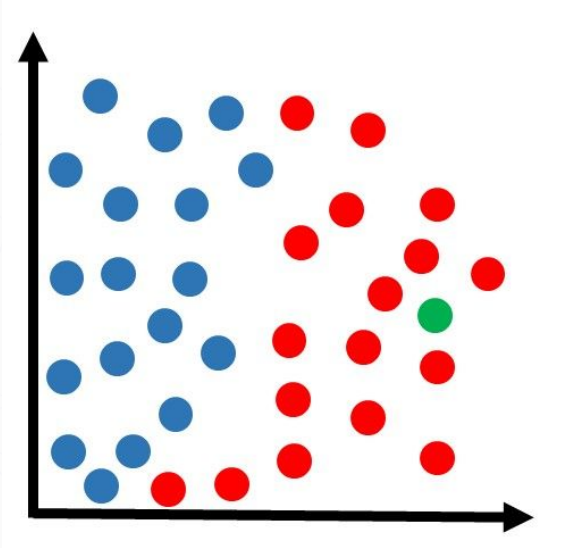
Existem diversas abordagens para construir modelos de classificação. Cada tipo de algoritmo possui lógica e aplicações específicas.

- **Baseados em conhecimento:** usam regras explícitas de especialistas (ex: if-else)
- **Baseados em árvore ou regras:** estruturam decisões hierárquicas
- **Conexionistas:** como redes neurais, imitam o funcionamento do cérebro
- **Baseados em distância e função:** usam medidas matemáticas para separar classes
- **Probabilísticos:** atribuem classes com base em probabilidades

Algoritmos de Classificação

K Vizinhos Mais Próximos (KNN)

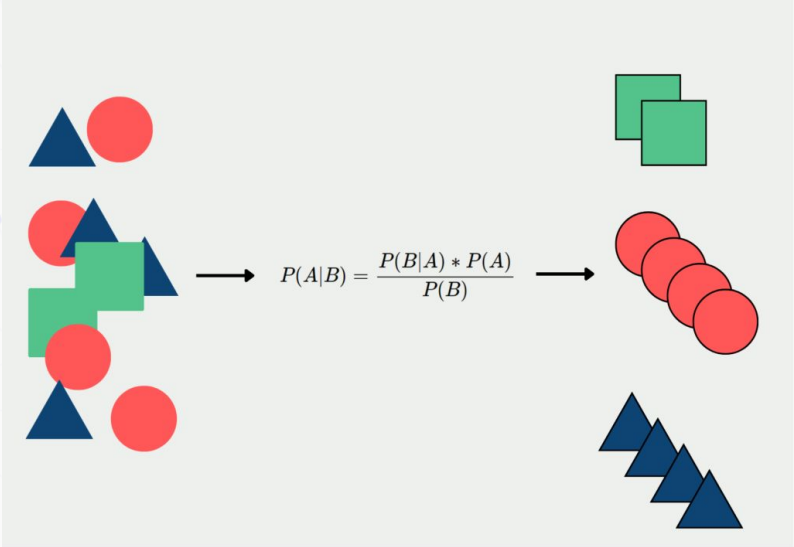
O KNN classifica objetos com base nos seus vizinhos mais próximos, sendo simples e eficaz para muitos tipos de dados.



Algoritmos de Classificação

Classificador Naive Bayes

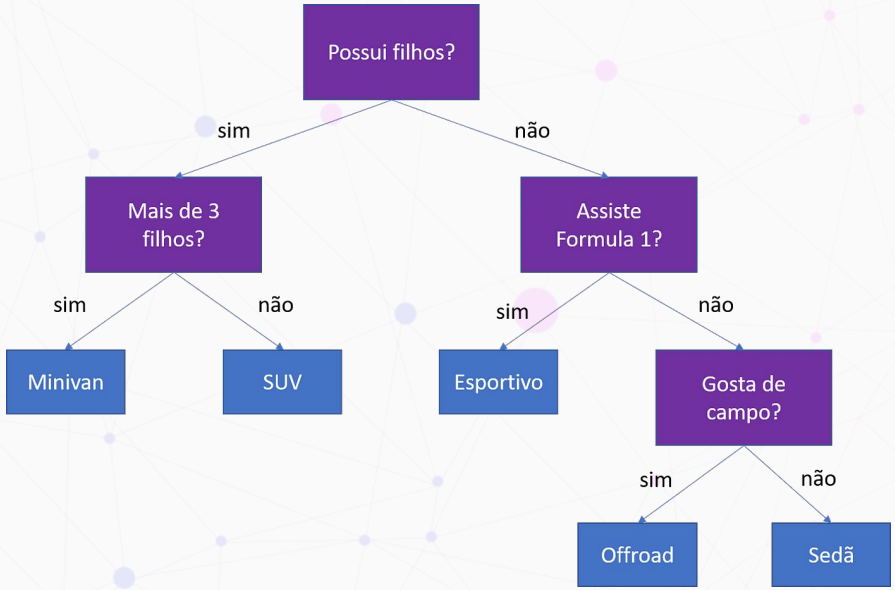
Naive Bayes usa probabilidades e pressupõe independência entre atributos, sendo rápido e eficaz mesmo com dados ruidosos.



Algoritmos de Classificação

Árvores de Decisão

Árvores de decisão criam regras visuais e interpretáveis para classificar dados com base na divisão sucessiva de atributos.



Regras de Associação

Regras de associação são técnicas usadas para identificar padrões em bases de dados transacionais, como compras em supermercados

Buscam **relações entre itens** que aparecem juntos em transações diferentes

Muito usadas em varejo, marketing, estoque e recomendações

Representadas como: X → Y, onde X é o antecedente e Y o consequente

A base precisa ser transformada em formato binário (presença/ausência de itens)

Regras de Associação

CENTRO DE PESQUISA E FORMAÇÃO SESC

Medidas de Interesse: Suporte e Confiança

Para avaliar a relevância das regras geradas, utilizamos medidas quantitativas como suporte e confiança.

- **Suporte**: % de transações que contêm tanto X quanto Y → indica frequência
- **Confiança**: % de transações com X nas quais também ocorre Y → indica precisão
- Regras fortes atendem a limiares mínimos de suporte (min_sup) e confiança (min_conf)
- Ex: se pão → leite tem suporte 30% e confiança 80%, é uma regra forte se superar os limites definidos

Regras de Associação

CENTRO DE PESQUISA E FORMAÇÃO SESC

Além de suporte e confiança, outras métricas ajudam a lidar com grandes volumes de regras e refinar a análise

- **Lift**: avalia o quanto Y é mais frequente com X do que isoladamente
- **Convicção**: mede a confiabilidade da regra considerando o erro
- **Compreensibilidade**: regras com menos itens são mais fáceis de interpretar
- **Grau de interesse**: pondera o suporte da regra com o tamanho de X e Y
- O número de regras cresce exponencialmente com o número de itens: usar filtros é essencial

CENTRO DE PESQUISA E FORMAÇÃO SESC

Para nosso último encontro:

Convidado especial:

Leandro Alvim

É professor Adjunto do magistério superior (2010) do Departamento de Ciência da Computação (DCC) da Universidade Federal Rural do Rio de Janeiro (UFRRJ), Campus Nova Iguaçu. Foi chefe do Departamento de Tecnologias e Linguagens (2013-2014) e foi chefe do Departamento de Ciência da Computação. Doutor em Informática pela Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) na área de concentração Algoritmos, Raciocínio Automático e Otimização. Foi coordenador de diversos projetos de pesquisa relacionando academia e mercado, como exemplo a empresa Buscapé Company. É revisor de periódicos da editora Elsevier e de projetos da do The Fund for Scientific Research-FNRS. Suas principais áreas de interesse são: Aprendizado de Máquina, Processamento de Linguagem Natural, Mercado Financeiro, Otimização, Grafos e Algoritmos.

CENTRO DE PESQUISA E FORMAÇÃO
SESC

OBRIGADO!

Até a próxima aula!

CENTRO DE PESQUISA E FORMAÇÃO
SESC

Hora da pausa! Voltamos em:

<<20:00->>