CIÊNCIA DE DADOS:

A BASE PARA DECISÕES INTELIGENTES

AULA 2: CONTEXTUALIZAÇÃO E QUALIDADE DOS DADOS

CONTEXTUALIZAÇÃO INICIAL

O CORAÇÃO DO PROJETO: QUAL CLIENTE COMPRARÁ?

Nesta aula, revisaremos o problema central do nosso projeto: **"Quais clientes têm maior probabilidade de comprar nos próximos 30 dias?"**

Compreender essa questão é fundamental para direcionar nossas estratégias.

A qualidade dos dados é a espinha dorsal de qualquer análise preditiva. Dados precisos significam decisões mais acertadas, otimizando o marketing e aumentando as conversões.

- Marketing direcionado
- Aumento de conversão
- Redução de desperdício

REVISANDO NOSSO DATASET

Nosso conjunto de dados abrange diversas dimensões do comportamento do cliente, mas a qualidade inicial é um desafio.

DADOS DO CLIENTE

cliente_id, idade, genero, UF

HISTÓRICO DE COMPRAS

recencia_dias, frequencia_12m, valor_total_12m, ticket_medio_12m

COMPORTAMENTO ONLINE

cliques_email_90d, visitas_site_90d, uso_cupom_12m

VARIÁVEIS E QUALIDADE

forma_pagamento, canal_predominante, vai_comprar_30d

Atenção: Dados podem estar incompletos, duplicados ou inconsistentes.

NOSSO CAMINHO NO CRISP-DM

FOCO NA PREPARAÇÃO DE DADOS

"COMO GARANTIR A QUALIDADE DOS DADOS?"

O PODER DA MINERAÇÃO DE DADOS

Definição: A mineração de dados é o processo de explorar grandes conjuntos de dados para identificar padrões significativos e informações úteis.

Objetivo: Transformar dados brutos em conhecimento compreensível e acionável.

Fluxo Simplificado: Coleta → Limpeza → Análise → Interpretação.

MINERAÇÃO DE DADOS NO VAREJO

PREVISÃO DE VENDAS

Estimativa de demanda futura para otimização de estoque e campanhas.

SEGMENTAÇÃO DE CLIENTES

Identificação de grupos com comportamentos e necessidades similares.

RECOMENDAÇÃO DE PRODUTOS

Sugestão personalizada de itens, aumentando o valor do carrinho.

DETECÇÃO DE CHURN

Identificação de clientes com risco de deixar de comprar.

> **Exemplo Prático:** Mulheres de 25-34 anos compram mais produtos da categoria X no inverno. Isso nos permite criar ações de marketing e gerenciar o estoque de forma mais eficiente.

POR QUE A PREPARAÇÃO DE DADOS É CRUCIAL?

1 **ENTENDIMENTO DO NEGÓCIO**

2 **ENTENDIMENTO DOS DADOS**

3 **PREPARAÇÃO DOS DADOS (NOSSO FOCO!)**

Esta etapa consome até **80% do tempo** de um projeto de Ciência de Dados.

4 **MODELAGEM**

5 **AVALIAÇÃO**

6 **IMPLANTAÇÃO**

LIMPEZA E TRATAMENTO DE DADOS

VALORES NULOS

Identificar NaN, null ou vazios. Estratégias incluem remoção de registros ou imputação (média, mediana, moda, interpolação).

DUPLICATAS

Remover registros duplicados utilizando IDs únicos para garantir a unicidade dos dados.

OUTLIERS

Identificar via Boxplot ou desvio padrão. Decidir se devem ser mantidos, corrigidos ou removidos, conforme o contexto.

BOAS PRÁTICAS:

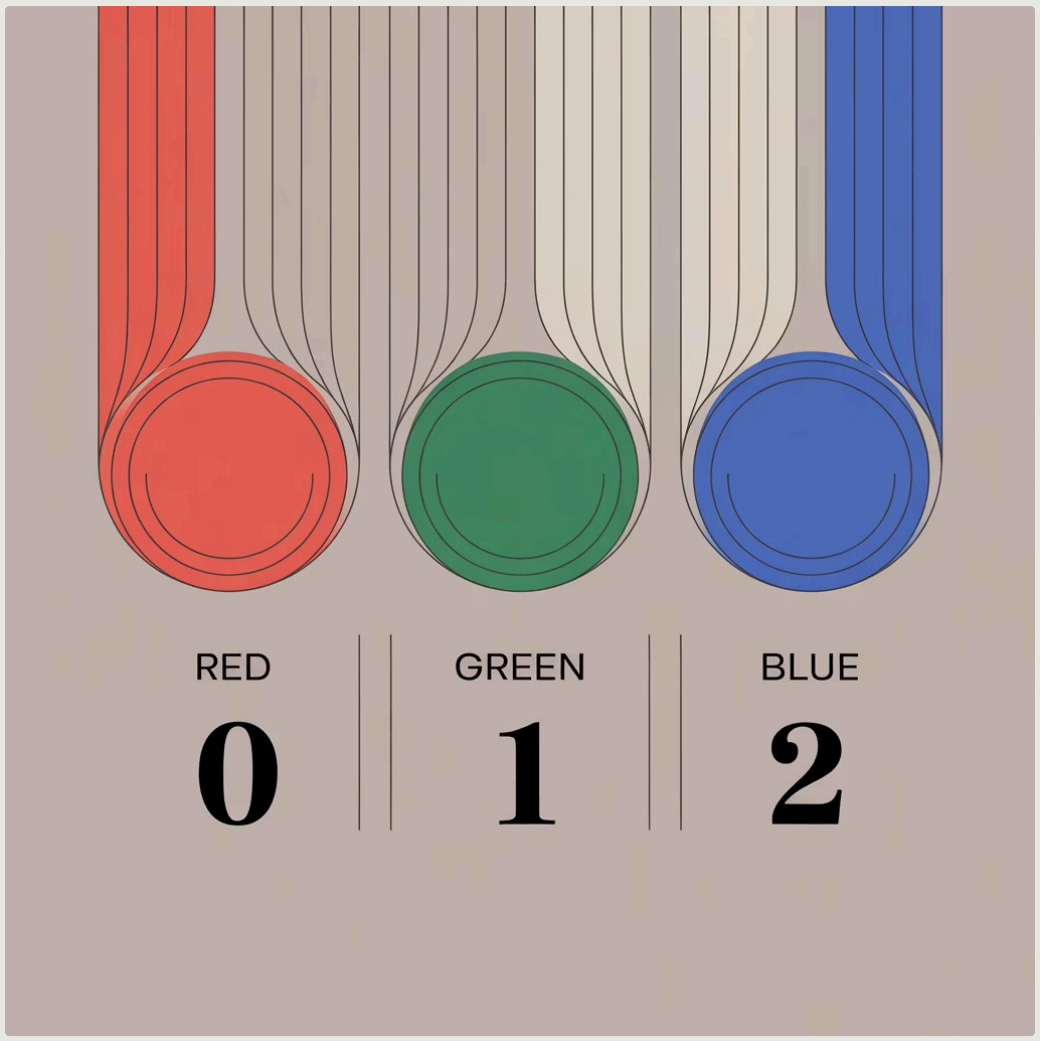
- **Backup:** Sempre faça backup dos dados originais.
- **Documentação:** Registre todas as transformações.
- **Validação:** Verifique a qualidade dos dados após o tratamento.
- **Automação:** Automatize processos para consistência.

CODIFICAÇÃO DE VARIÁVEIS CATEGÓRICAS

Algoritmos de Machine Learning não processam texto. Precisamos transformar categorias em números!

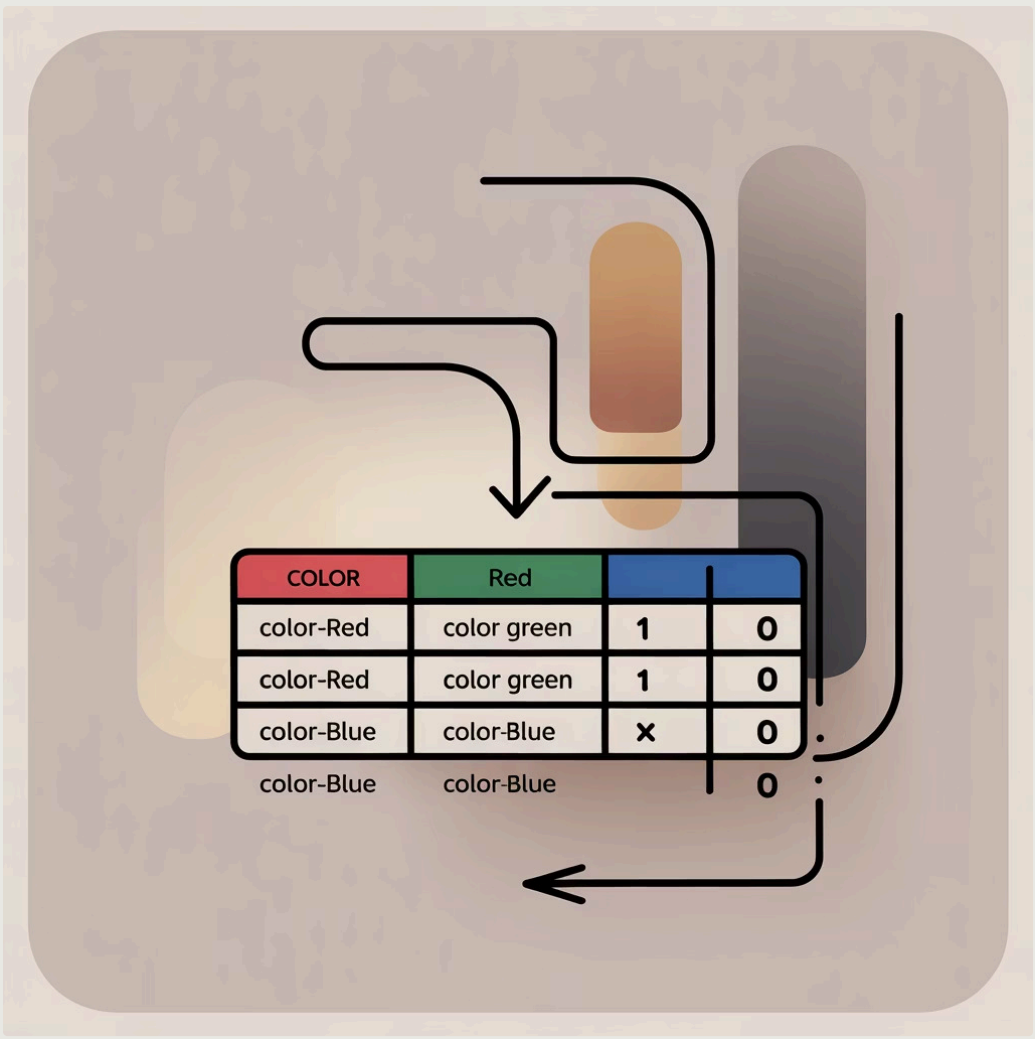
LABEL ENCODING:

- Atribui um número inteiro a cada categoria.
- Simples de usar, mas pode criar uma ordem artificial que não existe nos dados (ex: P = 0, M = 1, G = 2).



ONE-HOT ENCODING:

- Cria uma nova coluna binária (0 ou 1) para cada categoria.
- Evita a criação de ordem artificial, mas pode aumentar a dimensionalidade do dataset (muitas colunas).



DICAS PRÁTICAS:

Padronize nomes, agrupe categorias raras e documente todas as transformações realizadas.

NORMALIZAÇÃO E PADRONIZAÇÃO

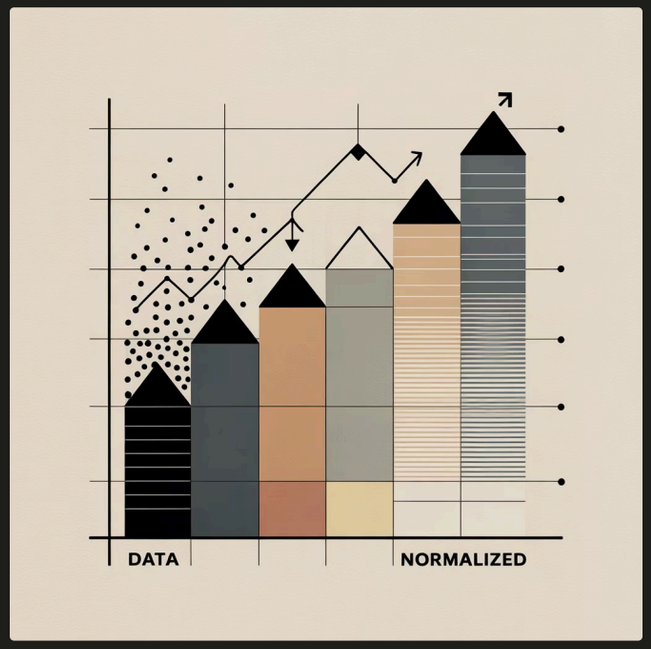
Para evitar que variáveis com diferentes escalas dominem o modelo, ajustamos seus valores.

NORMALIZAÇÃO (MIN-MAX SCALING)

Escala os valores para um intervalo fixo, geralmente entre 0 e 1. Útil quando a distribuição dos dados não é Gaussiana.

PADRONIZAÇÃO (Z-SCORE)

Transforma os dados para que tenham média 0 e desvio padrão 1. Ideal para dados com distribuição normal ou quando o algoritmo assume normalidade.



DICA IMPORTANTE:

Sempre aplique essas técnicas apenas no conjunto de treino e transforme o conjunto de teste com base nos parâmetros aprendidos do treino. A escolha da técnica depende do algoritmo e da distribuição dos seus dados.