



**DESENVOLVIMENTO DE SOFTWARE MULTIPLATAFORMA**

**Disciplina: ISW-039 - Mineração de Dados**

**Aula 02: KDD – Processo de Descoberta de Conhecimento.**

Data 17/08/2024

Prof. Me. Anderson Silva Vanin

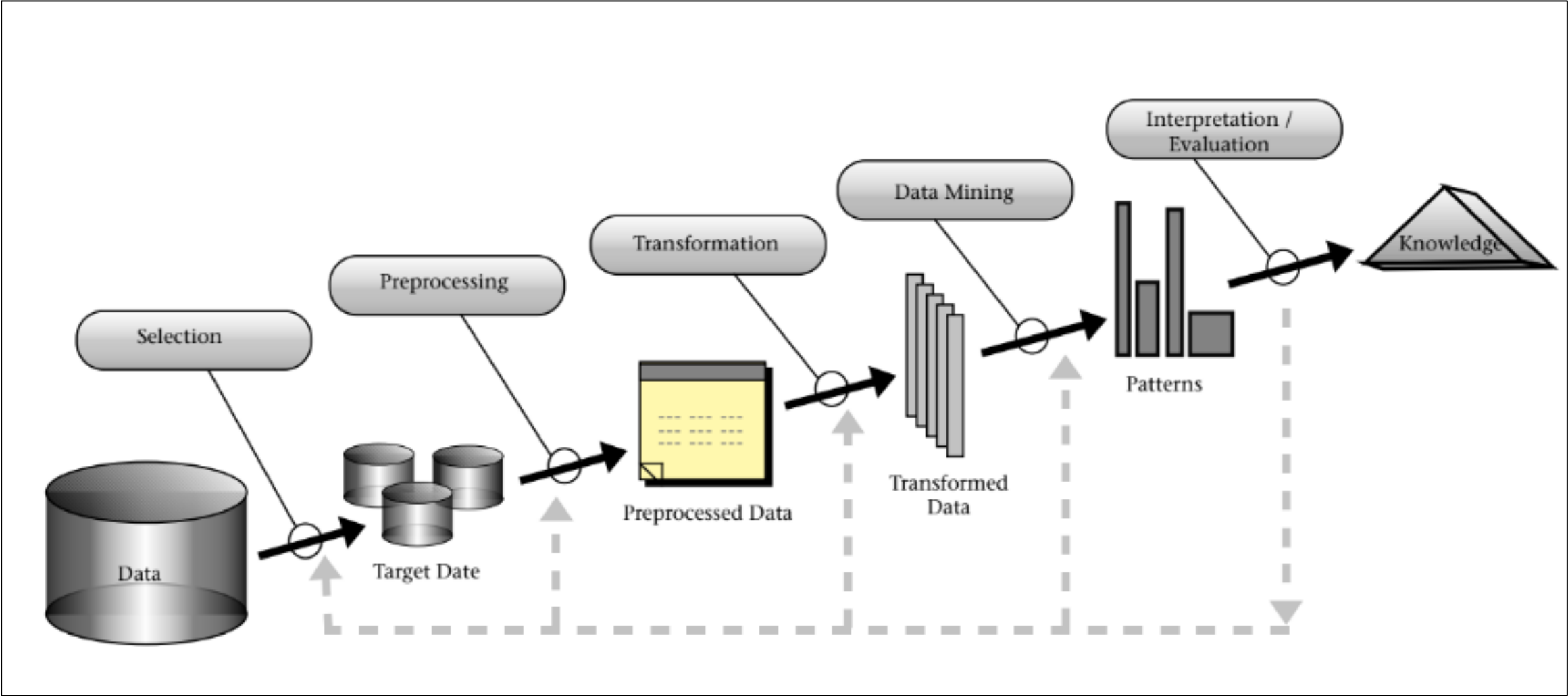
# Knowledge Discovery in Databases

## KDD

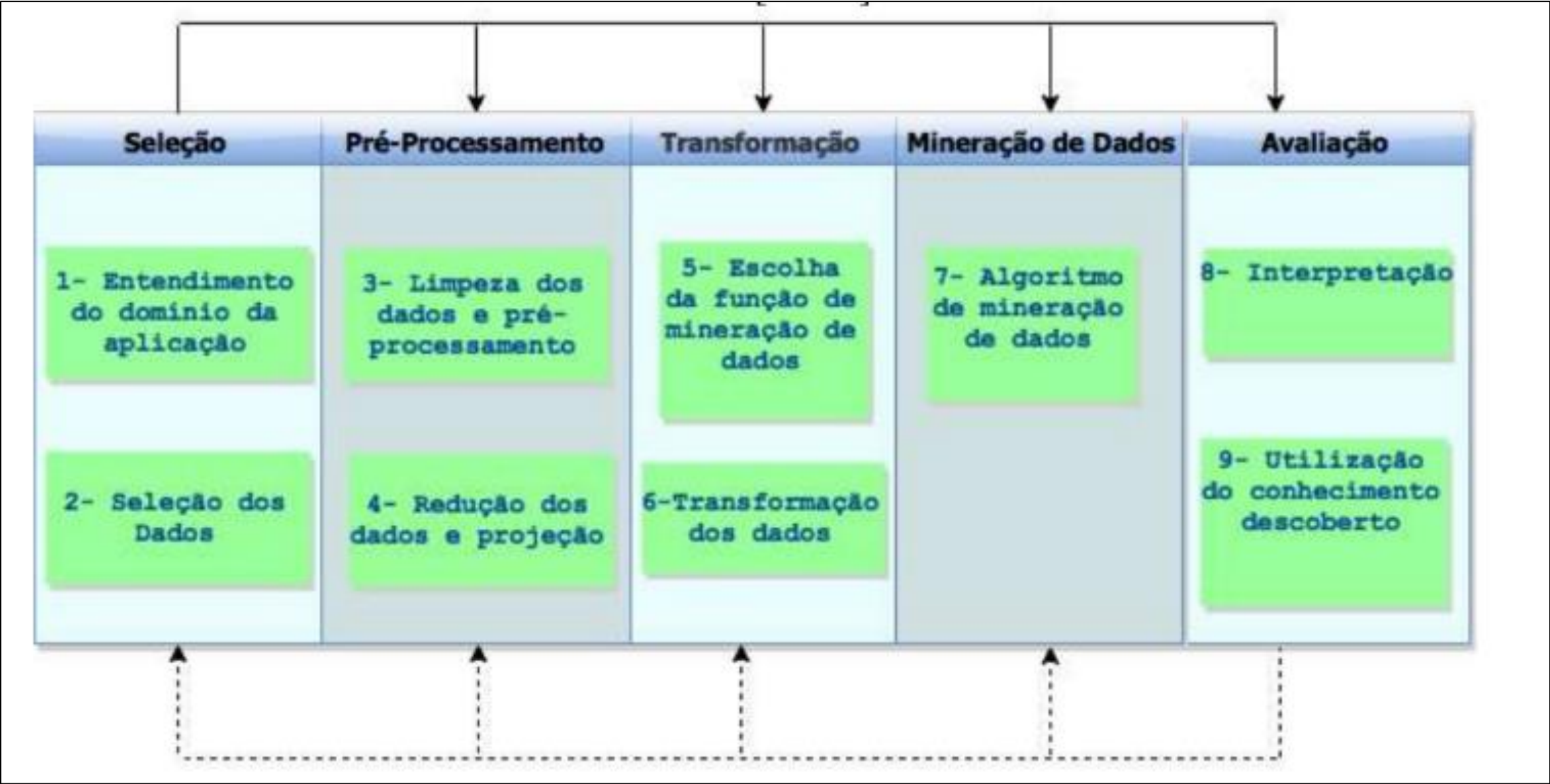
# KDD

O KDD é um processo mais amplo do que um algoritmo de Aprendizado de Máquina.  
**É um processo não trivial, iterativo, interativo e com múltiplos estágios que manipula e transforma os dados no intuito de descobrir padrões relevantes.**

# Knowledge Discovery in Databases - KDD



# Knowledge Discovery in Databases - KDD



# KDD

As cinco macro etapas do KDD:

## SELEÇÃO:

- 1- Entendimento do domínio da aplicação
- 2- Seleção de dados

## PRÉ-PROCESSAMENTO:

- 3- Limpeza dos dados e pré-processamento
- 4- Redução dos dados e projeção

## TRANSFORMAÇÃO:

- 5- Escolha da função de mineração de dados
- 6- Transformação dos dados

# KDD

## MINERAÇÃO DE DADOS:

7- Algoritmo de mineração de dados

## AVALIAÇÃO:

8- Interpretação

9- Utilização do conhecimento descoberto

# SELEÇÃO DE DADOS

A primeira etapa do processo de KDD é a seleção de dados. Nessa fase, são selecionados os dados relevantes para a análise. Isso pode incluir a identificação das fontes de dados apropriadas, a definição dos critérios de inclusão/exclusão e a coleta dos dados necessários. Por exemplo, suponhamos que desejamos realizar uma análise de vendas de uma empresa de varejo. Selecionamos os dados de vendas, incluindo informações como data, valor da venda, produtos vendidos e informações dos clientes.



# PRÉ-PROCESSAMENTO

Após a seleção dos dados, é necessário realizar o pré-processamento dos mesmos. Nessa etapa, os dados brutos são preparados para análise. Isso envolve a limpeza dos dados, tratamento de valores ausentes, lidar com ruídos e anomalias, e realizar a normalização dos dados, se necessário. Por exemplo, em nosso conjunto de dados de vendas, podemos remover entradas duplicadas, preencher valores ausentes e ajustar formatos de data.

# TRANSFORMAÇÃO

Uma vez que os dados tenham passado pelo pré-processamento, é hora de transformá-los em uma forma adequada para a análise. Isso pode envolver técnicas como agregação, redução de dimensionalidade e codificação de dados. Por exemplo, podemos agrupar as vendas por mês para obter uma visão mais ampla dos padrões sazonais de vendas.

# MINERAÇÃO DE DADOS

A etapa de mineração de dados é o coração do processo de KDD. Aqui, aplicamos algoritmos de mineração para descobrir padrões, relações e insights nos dados. Existem diferentes técnicas de mineração de dados disponíveis, como associação, classificação, clusterização e detecção de anomalias. Voltando ao nosso exemplo de vendas, poderíamos usar técnicas de mineração para identificar quais produtos são frequentemente comprados juntos por clientes.

# AVALIAÇÃO

Uma vez que os padrões e insights tenham sido descobertos, é importante avaliar e interpretar os resultados. Isso envolve a análise dos padrões encontrados, a validação de sua significância estatística e a interpretação do seu impacto no contexto do problema em estudo. Voltando ao exemplo de vendas, podemos avaliar a força das associações encontradas entre os produtos e identificar oportunidades de cross-selling.

# Utilização do Conhecimento Descoberto

A última etapa do processo de KDD é a utilização do conhecimento descoberto para tomar decisões informadas. Os insights obtidos podem ser usados para melhorar processos, otimizar estratégias, fazer previsões e tomar medidas corretivas, dependendo do contexto do problema. No caso das vendas, o conhecimento descoberto pode direcionar campanhas de marketing segmentadas com base nas associações encontradas entre os produtos.

## Exemplo Prático 1: Previsão de Churn em Telecomunicações

Agora que entendemos todas as etapas do processo do KDD, vamos considerar um exemplo prático do seu uso em uma empresa de telecomunicações. O objetivo é prever o churn (cancelamento de serviço) dos clientes.

## Exemplo Prático 1: Previsão de Churn em Telecomunicações

1. **Seleção de Dados:** Coletamos dados sobre os clientes, incluindo tempo de serviço, tipos de serviço utilizados, reclamações anteriores e histórico de pagamentos.
2. **Pré-processamento de Dados:** Lidamos com valores ausentes, removemos duplicatas e corrigimos erros. Por exemplo, se alguns clientes tiverem registros de tempo de serviço inconsistentes, esses registros seriam corrigidos ou removidos.
3. **Transformação de Dados:** Criamos variáveis como a média de reclamações nos últimos seis meses e a proporção de pagamentos atrasados. Essas variáveis podem ajudar a capturar padrões mais complexos.

## Exemplo Prático 1: Previsão de Churn em Telecomunicações

4. **Mineração de Dados:** Aplicamos algoritmos de machine learning para identificar padrões que indicam a probabilidade de um cliente cancelar o serviço nos próximos meses.
5. **Avaliação:** Avaliamos a precisão do modelo usando dados de clientes anteriores, comparando as previsões do modelo com os casos reais de churn.
6. **Interpretação:** Identificamos os principais fatores que contribuem para o churn, como reclamações frequentes ou histórico de pagamentos inconsistentes. Essas descobertas podem orientar estratégias para reter clientes.



## Exemplo Prático 2: Análise de Comportamento de Clientes em um E-commerce

**Objetivo:** Melhorar a estratégia de marketing e retenção de clientes de uma loja online, identificando padrões de compra.

## Exemplo Prático 2: Análise de Comportamento de Clientes em um E-commerce

### 1. Seleção dos Dados: Escolha do conjunto de dados relevante.

- Dados disponíveis: Transações de clientes, incluindo informações como ID do cliente, data da compra, itens comprados, quantidades, preços e dados demográficos dos clientes.
- Fonte: Banco de dados de transações do e-commerce.
- Exemplo de dados:

Cliente_ID	Data	Item	Quantidade	Preço	Idade	Sexo	Localização
123	2024-01-15	Camiseta	2	30.00	25	F	São Paulo
456	2024-01-16	Tênis	1	120.00	32	M	Rio de Janeiro
789	2024-01-17	Jaqueta	1	80.00	28	F	Salvador

## Exemplo Prático 2: Análise de Comportamento de Clientes em um E-commerce

### 2. Pré-processamento: Limpeza e preparação dos dados.

- Ações: Remoção de registros duplicados, tratamento de valores faltantes, e normalização de dados.
- Exemplo: Remoção de transações duplicadas. Preenchimento de valores ausentes para a idade média do cliente ou local. Conversão de datas para um formato padrão.

## Exemplo Prático 2: Análise de Comportamento de Clientes em um E-commerce

Antes:

Cliente_ID	Data	Item	Quantidade	Preço	Idade	Sexo	Localização
123	15-01-2024	Camiseta	2	30.00	25	F	São Paulo
123	15-01-2024	Camiseta	2	30.00	25	F	São Paulo
456	16-01-2024	Tênis	1	120.00	32	M	Rio de Janeiro
789	17-01-2024	Jaqueta	1	80.00	NaN	F	Salvador

Depois:

Cliente_ID	Data	Item	Quantidade	Preço	Idade	Sexo	Localização
123	2024-01-15	Camiseta	2	30.00	25	F	São Paulo
456	2024-01-16	Tênis	1	120.00	32	M	Rio de Janeiro
789	2024-01-17	Jaqueta	1	80.00	28	F	Salvador

## Exemplo Prático 2: Análise de Comportamento de Clientes em um E-commerce

**3. Transformação:** Conversão dos dados para formatos adequados para a mineração de dados.

- Ações: Criação de novas variáveis ou agregação de dados.
- Exemplo: Nova variável → **Total gasto por transação**. Nova variável → **Mês da compra**.

Cliente_ID	Data	Item	Quantidade	Preço	Idade	Sexo	Localização	Total_Gasto	Mês
123	2024-01-15	Camiseta	2	30.00	25	F	São Paulo	60.00	Janeiro
456	2024-01-16	Tênis	1	120.00	32	M	Rio de Janeiro	120.00	Janeiro
789	2024-01-17	Jaqueta	1	80.00	28	F	Salvador	80.00	Janeiro

## Exemplo Prático 2: Análise de Comportamento de Clientes em um E-commerce

**4. Mineração de Dados:** Aplicação de algoritmos para identificar padrões. Técnicas:

Agrupamento, classificação, regressão, etc. Exemplo:

Agrupamento → Segmentação de clientes com base no total gasto e frequência de compra.

Regressão → Previsão do total gasto futuro com base em comportamentos passados.

Resultado (agrupamento):

Segmento A: Clientes com alto valor de vida (Gasto médio  $> 100$ ).

Segmento B: Clientes com valor moderado de vida (Gasto médio entre 50 e 100).

Segmento C: Clientes com baixo valor de vida (Gasto médio  $< 50$ ).

## Exemplo Prático 2: Análise de Comportamento de Clientes em um E-commerce

**5. Interpretação e Avaliação:** Interpretação dos padrões encontrados e avaliação de sua validade e utilidade.

Ações: Avaliação de métricas de desempenho, validação cruzada, e análise de significância.

Exemplo:

Interpretação → Clientes do Segmento A são os que mais compram produtos de alta margem de lucro e poderiam ser alvo de programas de fidelidade.

Avaliação → Medição da precisão e recall para prever a compra futura com base no modelo.

## Exemplo Prático 2: Análise de Comportamento de Clientes em um E-commerce

**6. Apresentação e Ação:** Apresentação dos resultados e implementação de ações baseadas nos insights.

Ações: Relatórios, visualizações, e estratégias de marketing baseadas nos segmentos identificados.

Exemplo:

Relatório → Gráficos e tabelas que mostram os diferentes segmentos de clientes.

Ação → Implementação de campanhas de marketing direcionadas para Segmento A com ofertas especiais.



## Exemplo Prático 2: Análise de Comportamento de Clientes em um E-commerce

Segmentos de Clientes		
-----		
Segmento	Quantidade	Gasto Médio
-----	-----	-----
A	50	150.00
B	150	75.00
C	300	30.00

**Conclusão:** Baseado nos padrões descobertos, a loja pode personalizar suas campanhas para diferentes segmentos de clientes, otimizar promoções e melhorar a retenção de clientes de alto valor.