Data Mining e Graph Mining

Inteligência Artificial

Professor André Luiz Esperidião

Release 2.0

INTRODUÇÃO

Transformar dados em insights estratégicos através de uma metodologia comprovada e aplicada globalmente em projetos de mineração de dados.

INTRODUÇÃO AO DATA MINING

FUNDAMENTOS

O que é Data Mining?

Data Mining é o processo sistemático de descobrir padrões, tendências e insights valiosos em grandes volumes de dados. Esta técnica poderosa combina estatística, aprendizado de máquina e análise de dados para extrair conhecimento acionável que permaneceria oculto em bases de dados complexas.

Empresas líderes utilizam Data Mining para apoiar decisões estratégicas críticas em diversas áreas: negócios, marketing, finanças e operações. A capacidade de antecipar comportamentos e identificar oportunidades representa uma vantagem competitiva significativa no mercado atual.

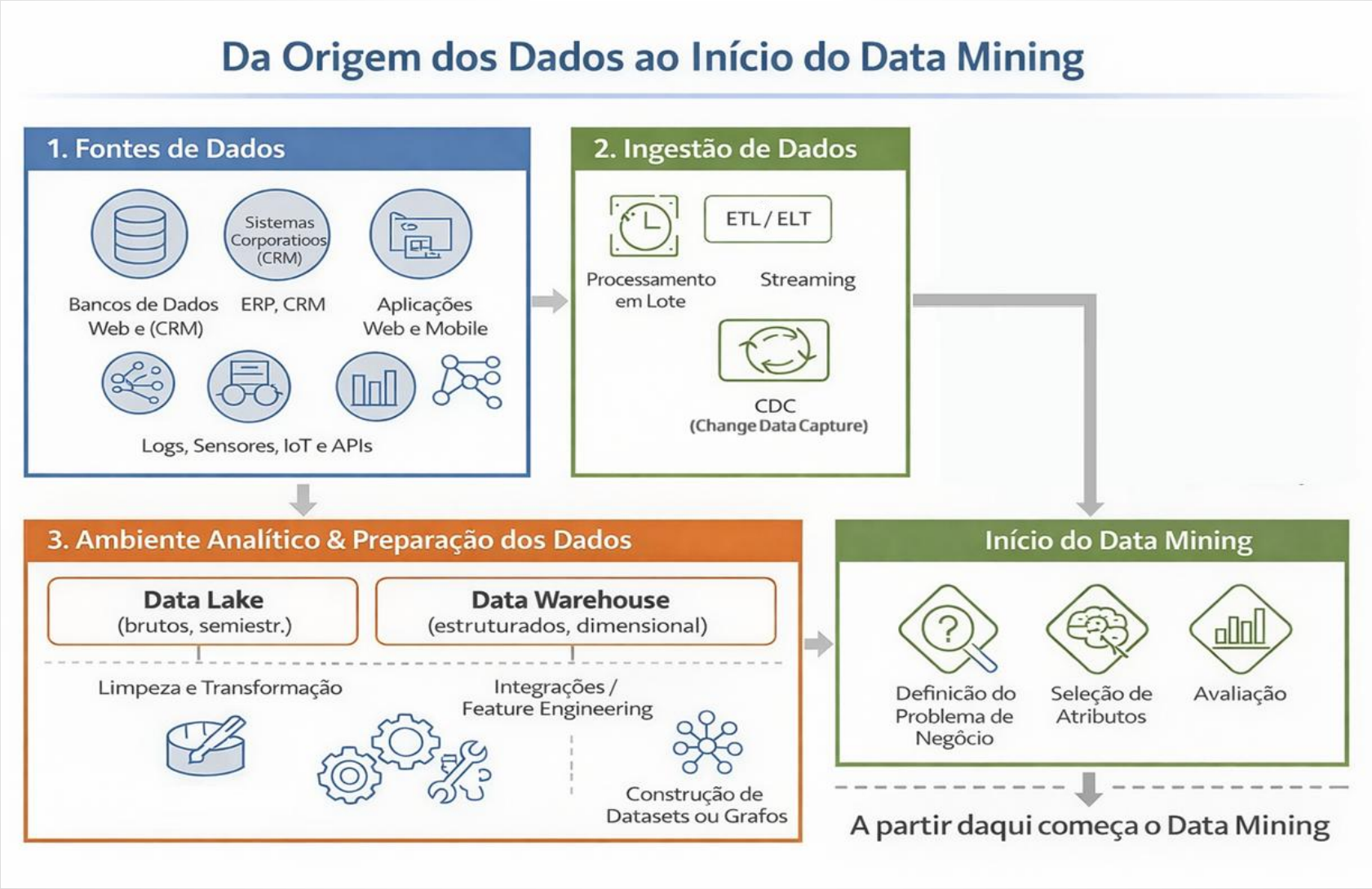
Aplicações práticas incluem: prever churn de clientes antes que aconteça, detectar fraudes em tempo real, otimizar campanhas de marketing com precisão, personalizar experiências do consumidor e identificar novos segmentos de mercado.

INTRODUÇÃO AO DATA MINING

O QUE É DATA MINING (MINERAÇÃO DE DADOS)?

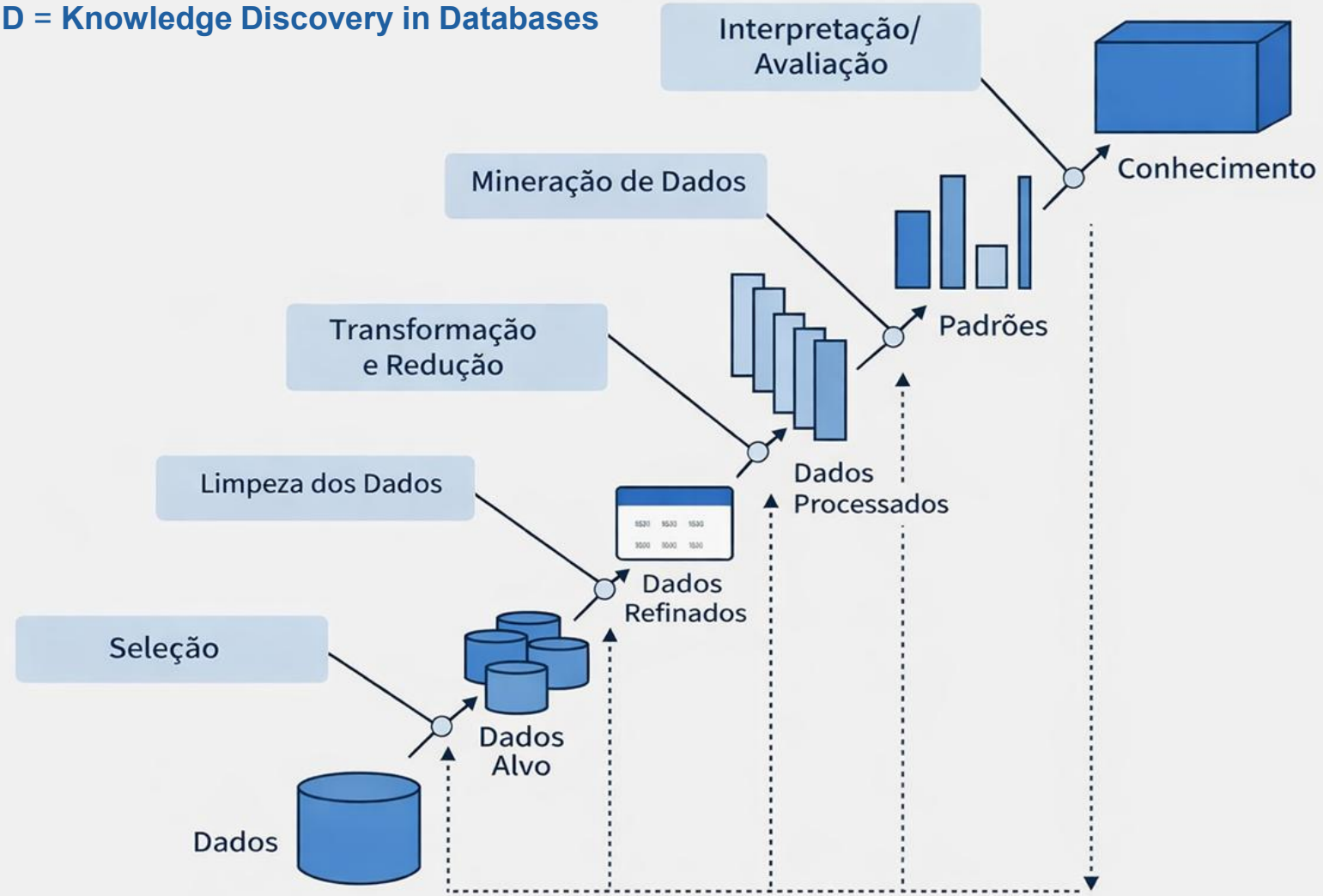
- Data Mining ≠ Banco de Dados
- Data Mining ≠ BI
- Data Mining ≠ Machine Learning
- ➡ Data Mining = **descoberta de padrões úteis**

INTRODUÇÃO AO DATA MINING



INTRODUÇÃO AO DATA MINING

KDD = Knowledge Discovery in Databases

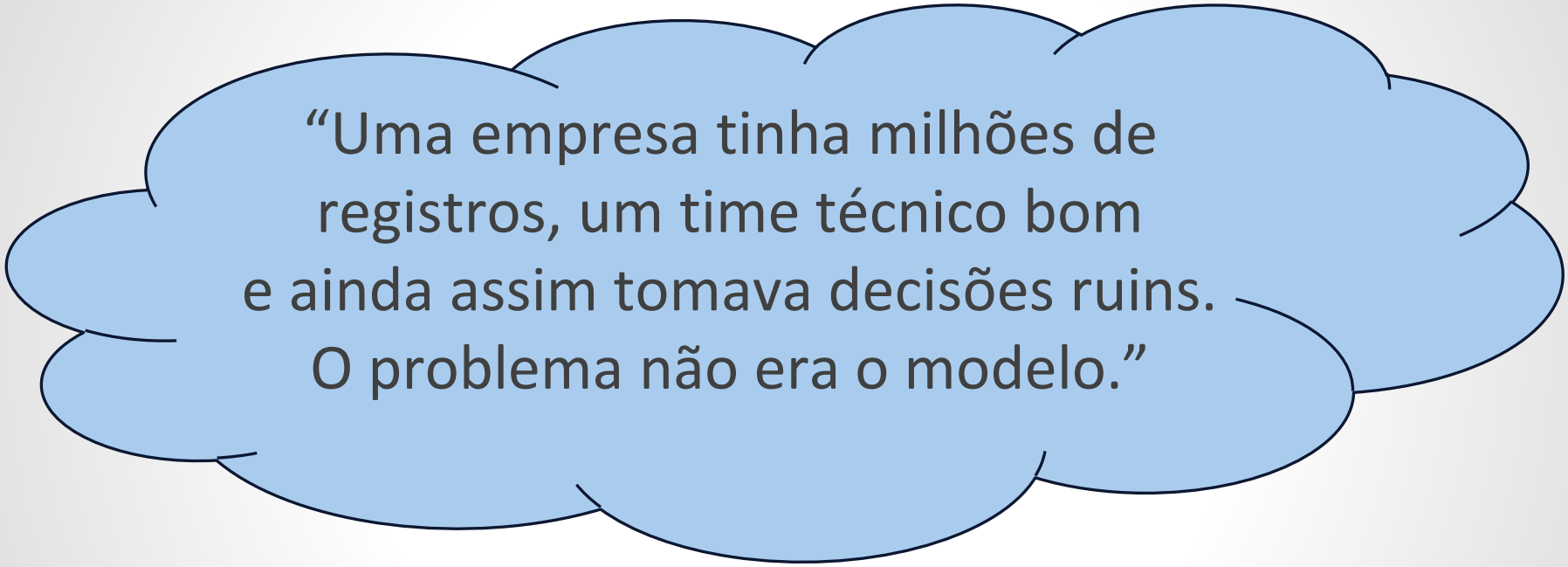


INTRODUÇÃO AO DATA MINING

🧠 Comparação KDD vs CRISP-DM

Aspecto	KDD	CRISP-DM
Nome completo	Knowledge Discovery in Databases	Cross-Industry Standard Process for Data Mining
Origem	Acadêmica / pesquisa	Industrial / mercado
Foco principal	**Descoberta de conhecimento**	**Entrega de valor ao negócio**
Natureza	Conceitual	Metodológica e operacional
Adoção	Artigos científicos, livros	Empresas, projetos reais
Ciclo	Iterativo	Iterativo e incremental

PENSANDO DATA MINING



Data Mining começa antes do código.

METODOLOGIA

CRISP-DM: O Padrão para Projetos de Data Mining

O CRISP-DM (Cross-Industry Standard Process for Data Mining) foi desenvolvido em 1999 por um consórcio de empresas líderes para padronizar e estruturar processos de mineração de dados em diferentes indústrias. Desde então, tornou-se a metodologia mais adotada mundialmente para projetos de Data Mining.

Esta abordagem se destaca por sua **flexibilidade e independência de ferramentas**, permitindo que organizações de qualquer porte e setor apliquem suas práticas. O framework é iterativo, reconhecendo que descobertas em fases posteriores frequentemente levam a revisões de etapas anteriores.

Composto por 6 fases interconectadas, o CRISP-DM guia todo o projeto desde a concepção inicial até a implantação em produção, garantindo alinhamento estratégico e qualidade dos resultados.

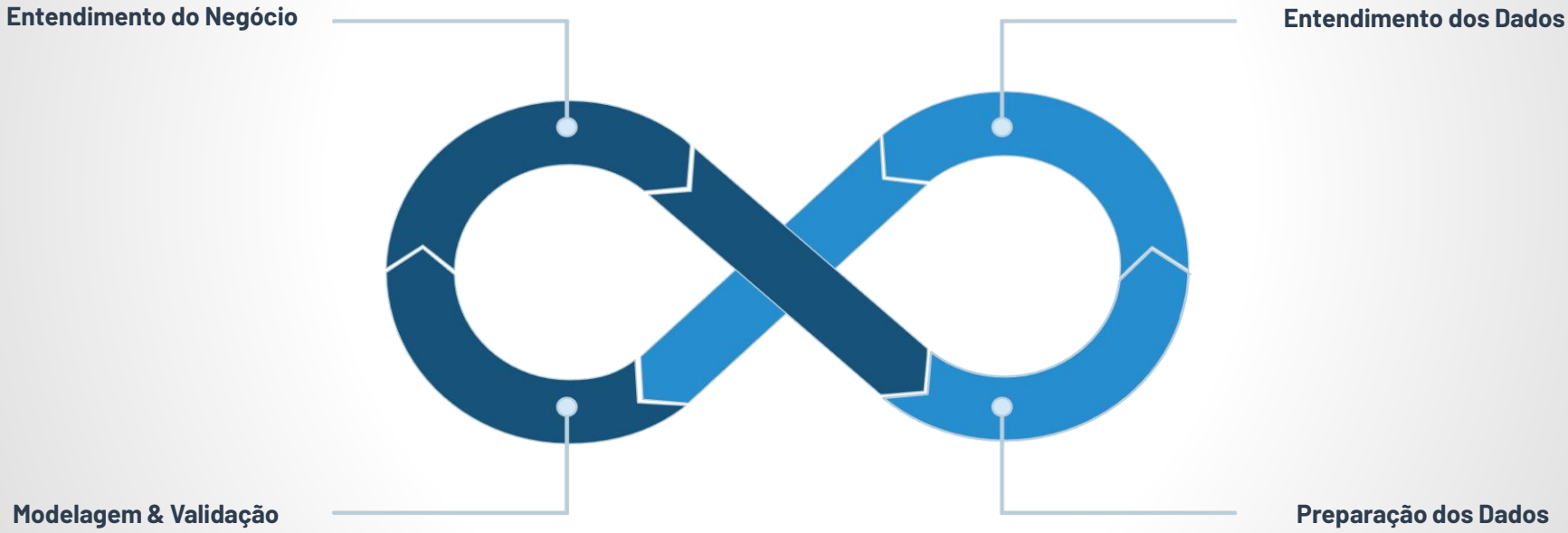
Benefícios Principais

- Redução de riscos e custos em projetos
- Maior taxa de sucesso e ROI
- Comunicação clara entre equipes
- Repetibilidade e escalabilidade
- Melhoria contínua dos processos

INTRODUÇÃO AO DATA MINING - CRISP-DM

As 6 Fases do CRISP-DM

O framework CRISP-DM estrutura projetos de Data Mining em seis fases distintas e complementares. Cada fase possui objetivos específicos, entregáveis claros e critérios de validação que garantem a qualidade e o alinhamento do projeto com as necessidades do negócio.



Este ciclo não é linear - projetos frequentemente retornam a fases anteriores conforme novos insights são descobertos. A natureza iterativa permite refinamento contínuo e adaptação às descobertas emergentes durante o processo.

INTRODUÇÃO AO DATA MINING - CRISP-DM

FASE1

Business Understanding

Definição de Objetivos

Estabelecer objetivos de negócio mensuráveis e critérios de sucesso claros e quantificáveis. Esta etapa fundamental garante que o projeto técnico esteja alinhado com as prioridades estratégicas da organização.

Avaliação de Recursos

Analisar disponibilidade de dados, ferramentas tecnológicas, orçamento, equipe e tempo. Identificar riscos potenciais, restrições técnicas e dependências que podem impactar o projeto.

Planejamento Estratégico

Desenvolver roadmap detalhado com etapas, marcos, responsabilidades e ferramentas. Definir metodologia de trabalho e processos de governança para garantir execução eficiente.

FASES 2 E 3

Entendimento e Preparação dos Dados

Data Understanding

A coleta e análise inicial exploram estrutura, formatos, qualidade e padrões dos dados disponíveis. Estatísticas descritivas revelam distribuições, outliers e relacionamentos iniciais entre variáveis.

Data Preparation

Esta fase consome tipicamente 70-80% do tempo do projeto. Inclui limpeza rigorosa para corrigir erros e inconsistências, tratamento inteligente de valores faltantes, normalização de formatos, integração de múltiplas fontes e criação de variáveis derivadas (feature engineering).

FASES 4

Modelagem

A fase de modelagem é onde a ciência de dados ganha vida. Aqui, técnicas estatísticas e de aprendizado de máquina são aplicadas aos dados preparados para criar modelos preditivos ou descritivos que respondem às questões de negócio definidas.

Seleção de Técnicas

Escolha de algoritmos apropriados como regressão logística, árvores de decisão, random forests, redes neurais, clustering K-means ou algoritmos de associação, dependendo do tipo de problema.

Treinamento do Modelo

Divisão dos dados em conjuntos de treino e teste. Aplicação dos algoritmos selecionados e ajuste fino de hiperparâmetros para otimizar métricas de desempenho como acurácia, precisão e recall.

Validação Cruzada

Uso de técnicas como k-fold cross-validation para garantir que o modelo generaliza bem e não está apenas memorizando os dados de treino (overfitting).

Exemplo aplicado: Utilizar modelo de classificação Random Forest para prever quais clientes têm maior probabilidade de cancelar o serviço nos próximos 30 dias, baseado em padrões de uso, histórico de reclamações e perfil demográfico.

INTRODUÇÃO AO DATA MINING - CRISP-DM

FASES 5 E 6

Evaluation (Avaliação)

Verificar se o modelo atende aos objetivos de negócio estabelecidos na fase 1 e se seus resultados são confiáveis e interpretáveis. Validação rigorosa com métricas apropriadas (accuracy, F1-score, AUC-ROC) e testes em dados não vistos anteriormente.

Revisão crítica com stakeholders para garantir que insights gerados são acionáveis e fazem sentido no contexto do negócio. Esta etapa pode revelar necessidade de retornar a fases anteriores para refinamento.

Deployment (Implantação)

Integrar o modelo validado em ambiente de produção para uso contínuo e automatizado. Estabelecer processos de monitoramento para acompanhar performance ao longo do tempo e detectar degradação.

Sistema em Produção

Criar documentação completa, treinar usuários finais e estabelecer processos de manutenção e atualização periódica dos modelos.

Exemplo real:
Sistema automatizado que analisa dados de clientes diariamente e alerta a equipe de retenção sobre os 100 clientes com maior risco de churn, priorizando ações proativas de relacionamento.

INTRODUÇÃO AO DATA MINING - CRISP-DM

FASES 5 E 6

Evaluation (Avaliação)

Verificar se o modelo atende aos objetivos de negócio estabelecidos na fase 1 e se seus resultados são confiáveis e interpretáveis. Validação rigorosa com métricas apropriadas (accuracy, F1-score, AUC-ROC) e testes em dados não vistos anteriormente.

Revisão crítica com stakeholders para garantir que insights gerados são acionáveis e fazem sentido no contexto do negócio. Esta etapa pode revelar necessidade de retornar a fases anteriores para refinamento.

Deployment (Implantação)

Integrar o modelo validado em ambiente de produção para uso contínuo e automatizado. Estabelecer processos de monitoramento para acompanhar performance ao longo do tempo e detectar degradação.

Sistema em Produção

Criar documentação completa, treinar usuários finais e estabelecer processos de manutenção e atualização periódica dos modelos.

Exemplo real:
Sistema automatizado que analisa dados de clientes diariamente e alerta a equipe de retenção sobre os 100 clientes com maior risco de churn, priorizando ações proativas de relacionamento.

Por que Usar CRISP-DM?

Metodologia Comprovada

Framework estruturado, testado e validado por milhares de projetos bem-sucedidos em dezenas de indústrias ao redor do mundo desde 1999.

Alinhamento Estratégico

Garante que esforços técnicos em dados e modelagem estejam sempre conectados aos objetivos reais de negócio e gerem valor mensurável.

Comunicação Eficaz

Facilita diálogo produtivo entre equipes técnicas e de negócio através de linguagem comum e etapas claramente definidas.

Melhoria Contínua

Permite repetir, refinar e escalar projetos de mineração de dados com segurança, aprendendo com cada iteração e construindo conhecimento organizacional.

Adotar o CRISP-DM não é apenas seguir uma metodologia - é investir em uma abordagem sistemática que maximiza chances de sucesso, reduz riscos e transforma dados em vantagem competitiva sustentável para sua organização.

CASO REAL

Exemplo Prático: Prevendo Churn de Clientes

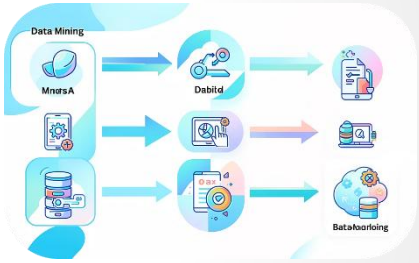
Objetivo do Projeto

Reduzir significativamente a perda de clientes na empresa de telecom, por meio de identificação de sinais precoces de insatisfação e risco de cancelamento.

Dados

Histórico completo de uso de serviços, registros de reclamações e atendimentos, perfil demográfico, dados de faturamento e interações com suporte técnico.

telecom_churn_synthetic.csv



Metodologia

Seguir rigorosamente todas as fases do CRISP-DM, desde o entendimento do negócio até a implantação, garantindo alinhamento estratégico e qualidade técnica.

CASO REAL

Exemplo Prático: Prevendo Churn de Clientes

Dados

telecom_churn_synthetic.csv

Visão Geral do Dataset

- **6000 linhas**
- **27 colunas**

CASO REAL

Exemplo Prático: Prevendo Churn de Clientes

Dados

telecom_churn_synthetic.csv

Visão Geral do Dataset

- **6000 linhas**
- **27 colunas**

https://github.com/andrelbe/Data-Mining-e-Graph-Mining/blob/main/telecom_churn_synthetic.csv

CASO REAL

Exemplo Prático: Prevendo Churn de Clientes

Dados

telecom_churn_synthetic.csv

Visão Geral do Dataset

- **6000 linhas**
- **27 colunas**

- **Quais dados?**
- **Quais desses dados são relevantes?**
- **Qual a taxa de Churn (churn rate)?**
- **Valores faltantes (missing)?**
  - total_charges?
  - avg_download_mbps?

CASO REAL

Exemplo Prático: Prevendo Churn de Clientes

Dados

telecom_churn_synthetic.csv

Visão Geral do Dataset

- **6000 linhas**
- **27 colunas**
- **Taxa de Churn**: **14.05%** (minoria)
- **Missing**:
  - total_charges: **~1.23%**
  - avg_download_mbps: **~0.77%**

O que é "missing" (em português claro)

- **Missing** = campo **sem valor** para alguns registros.
- Pode aparecer como:
  - NULL
  - vazio
  - NaN
  - ?
  - -1 (pior prática, mas acontece)

CASO REAL

Exemplo Prático: Prevendo Churn de Clientes

Dados

telecom_churn_synthetic.csv

Visão Geral do Dataset

- **6000 linhas**
- **27 colunas**
- **Taxa de Churn**: **14.05%** (minoria)
- **Missing**:
  - total_charges: **~1.23%**
  - avg_download_mbps: **~0.77%**

Interpretando os números

- **total_charges: ~1,23%**

≈ 1,23% dos clientes não têm valor registrado para o **total gasto**.

Exemplo:

Dataset com 10.000 clientes

~123 clientes sem total_charges

CASO REAL

Exemplo Prático: Prevendo Churn de Clientes

Dados

telecom_churn_synthetic.csv

Visão Geral do Dataset

- **6000 linhas**
- **27 colunas**
- **Taxa de Churn**: **14.05%** (minoria)
- **Missing**:
  - total_charges: **~1.23%**
  - avg_download_mbps: **~0.77%**

Interpretando os números

- **avg_download_mbps: ~0,77%**

≈ 0,77% dos clientes não têm média de download registrada.

Exemplo:

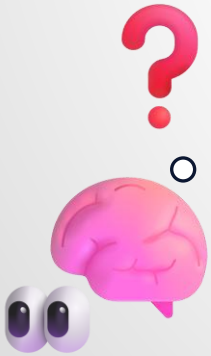
10.000 clientes

~77 clientes sem esse valor

CASO REAL

Exemplo Prático: Prevendo Churn de Clientes

Por que isso aparece no Entendimento dos Dados (CRISP-DM – Fase 2)?



Porque nessa fase você responde:
Os dados estão completos?
Onde estão os buracos?
Isso pode afetar o modelo?
👉 Isso **não é correção ainda**, é diagnóstico.

CASO REAL

Exemplo Prático: Prevendo Churn de Clientes

Dados

Visão Geral do Dataset

- **6000 linhas**
- **27 colunas**
- **Taxa de Churn**: **14.05%** (minoria)
- **Missing**:
  - total_charges: **~1.23%**
  - avg_download_mbps: **~0.77%**

Sobre o "Missing":

- Isso é muito? É pouco? Depende:
  - **Regra prática (bem usada em projetos reais):**
    - < 1% → geralmente aceitável
    - 1% a 5% → atenção
    - > 5% → problema sério
    - > 20% → variável candidata a descarte
  - **Aqui:**
    - 1,23% → ok, mas observar
    - 0,77% → tranquilo

STACK ESSENCIAL DE FERRAMENTAS

STACK ESSENCIAL DE FERRAMENTAS

🧠 STACK FINAL (RESUMO EXECUTIVO)

Finalidade	Ferramenta
Data Mining	Python + pandas + sklearn
Exploração	Jupyter
Graph Mining	NetworkX
Visualização	Gephi ou pyvis
Grafos reais	Neo4j
Escala	Spark + GraphFrames
Ambiente	Colab / Local
Profissional	Git

Data Mining e Graph Mining

Inteligência Artificial

Professor André Luiz Esperidião

Release 2.0