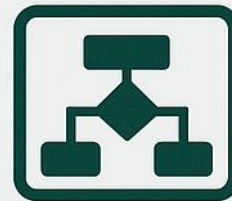


FRAMEWORKS E METODOLOGIAS PARA PROJETOS DE DATA SCIENCE



Objetivos

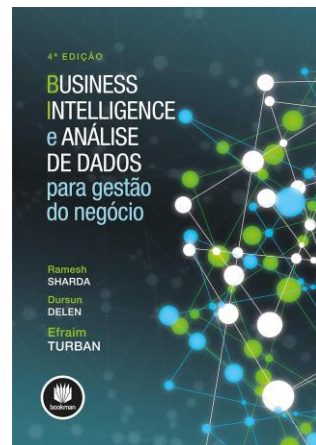
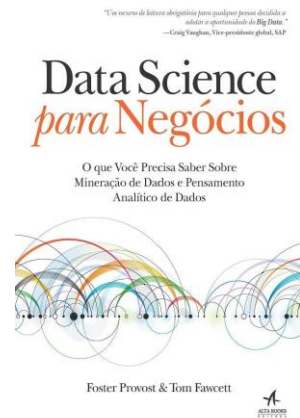
- Desenvolver uma compreensão abrangente sobre frameworks e metodologias aplicáveis à Ciência de Dados
- Desenvolver conceitos de preparação de dados desde a coleta de dados até a implementação de soluções para a geração de competitividade organizacional
- Aplicação prática de frameworks e metodologias em ambientes corporativos

O que iremos aprender ?

1. Conceitos e Definições dos principais frameworks e metodologias aplicáveis à Ciência de Dados
2. frameworks
 - CRISP-DM
 - SEMMA
 - KDD

Referências Bibliográficas

- Foster Provost e Tom Fawcett, **Data Science para Negócios**, Alta Books, 2016, ISBN: 9788576089728
- SHARDA, R.; DELEN, D.; TURBAN, E. **Business Intelligence e Análise de Dados para Gestão do Negócio**. 4. ed. Porto Alegre: Bookman, 2019.



Framework

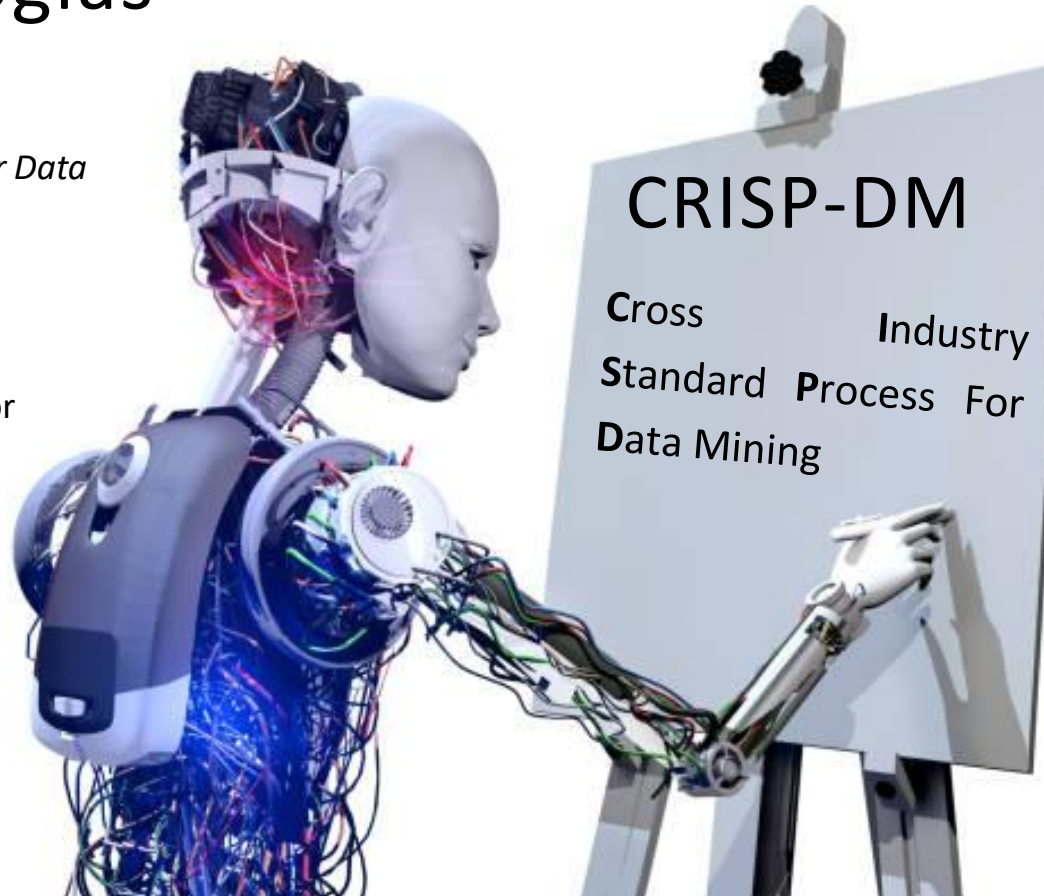
Metodologias

- Os frameworks **CRISP-DM** (*Cross Industry Standart Process for Data Mining*), **KDD** (*Knowledge Discovery in Databases*) e **SEMMA** (*Sample, Explore, Modify, Model e Assess*) são metodologias amplamente utilizadas em projetos de Ciência de Dados e Mineração de Dados
- Cada um tem um foco específico, mas todos compartilham etapas semelhantes para estruturar o processo analítico

Framework Metodologias

O CRISP-DM (*Cross Industry Standard Process for Data Mining*) que é dividida em seis fases principais

Criada há mais de 20 anos atrás e passou por vários ajustes e atualizações

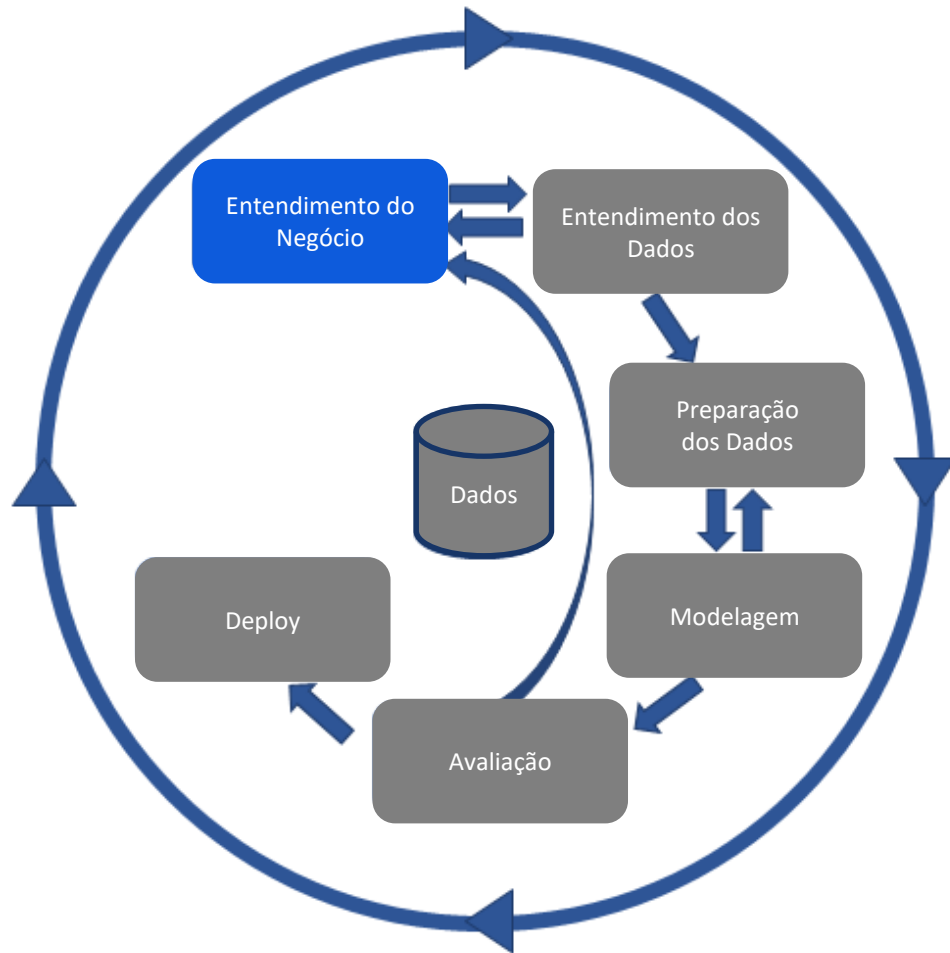


CRISP-DM

Entendimento do Negócio

- Definição dos objetivos do projeto
- Compreensão do domínio do problema
- Identificação das restrições e requisitos

Dica: Importância da Comunicação com o Negócio

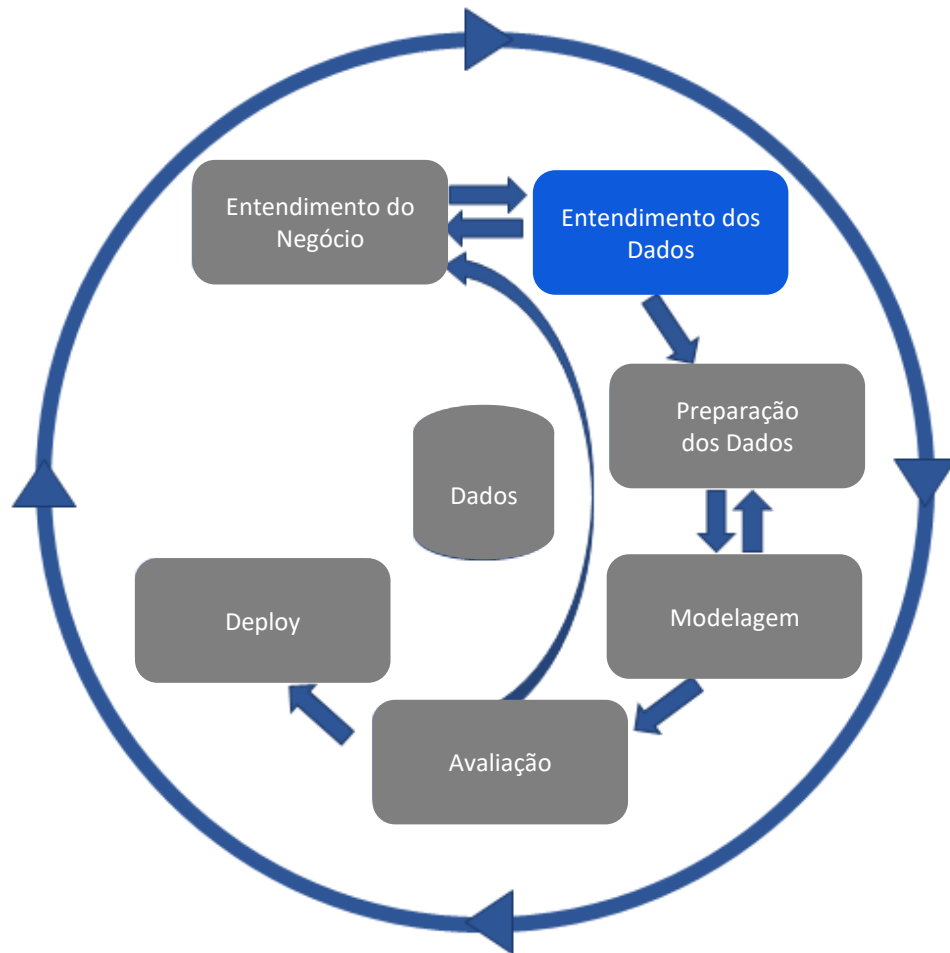


CRISP-DM

Entendimento dos Dados

- Coleta
- Análise exploratória
- Avaliação da qualidade dos dados

Dica: Importância dos principais conceitos de Estatística

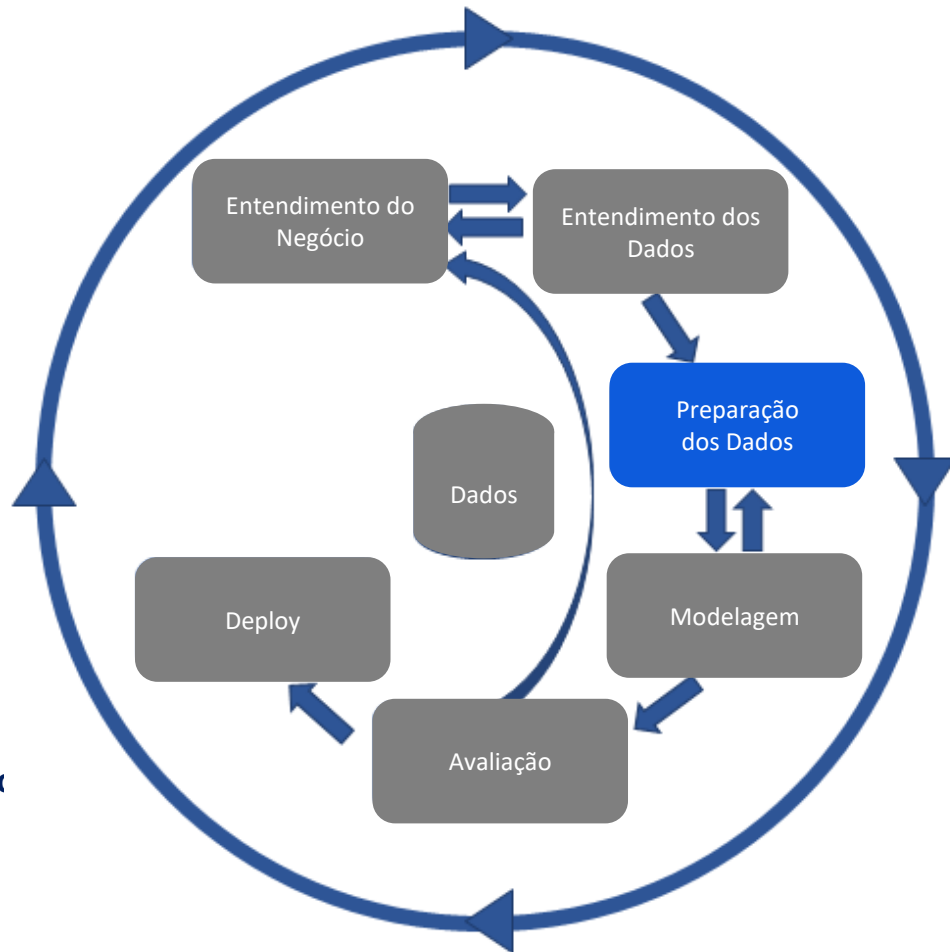


CRISP-DM

Preparação dos Dados

- Limpeza e tratamento
- Seleção de variáveis relevantes
- Transformação e formatação

Dica: É a parte mais demorada e trabalhosa de todas, porém um bom trabalho aqui significa menos retrabalho futuro

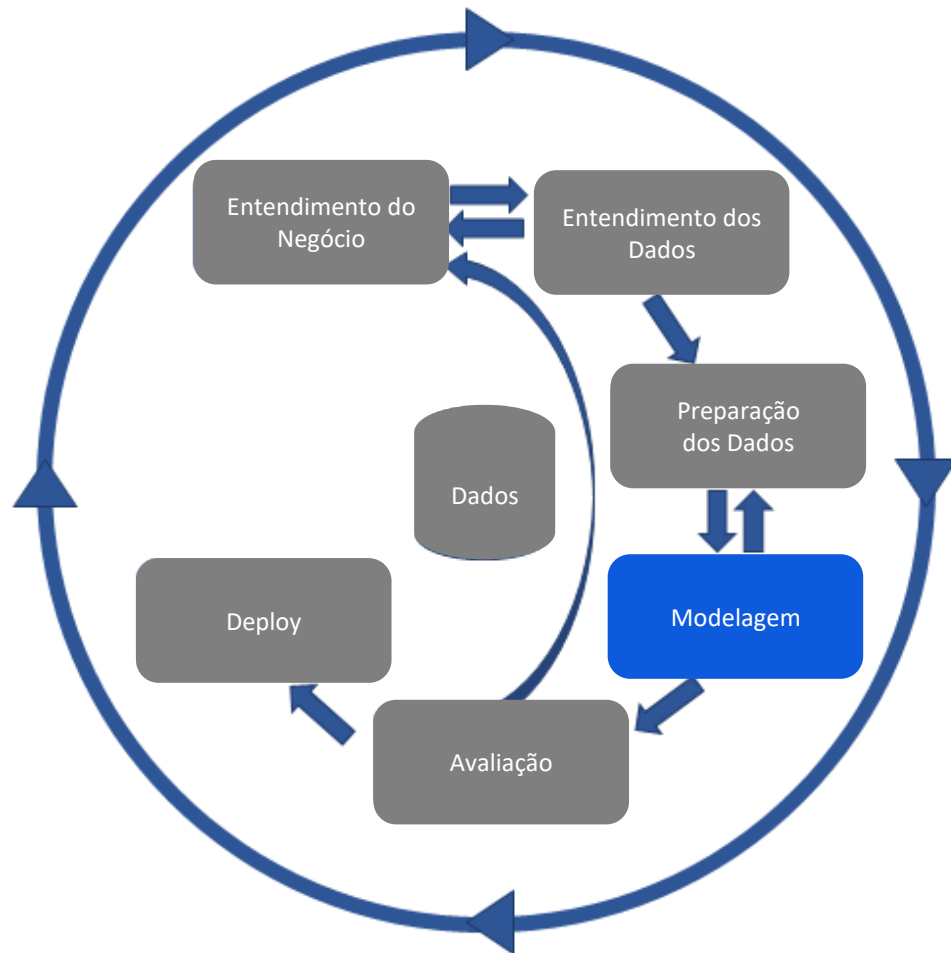


CRISP-DM

Modelagem

- Escolha da técnicas de modelagem
- Construção do Modelo
- Avaliação do desempenho do modelo

Dica: Menos é Mais!

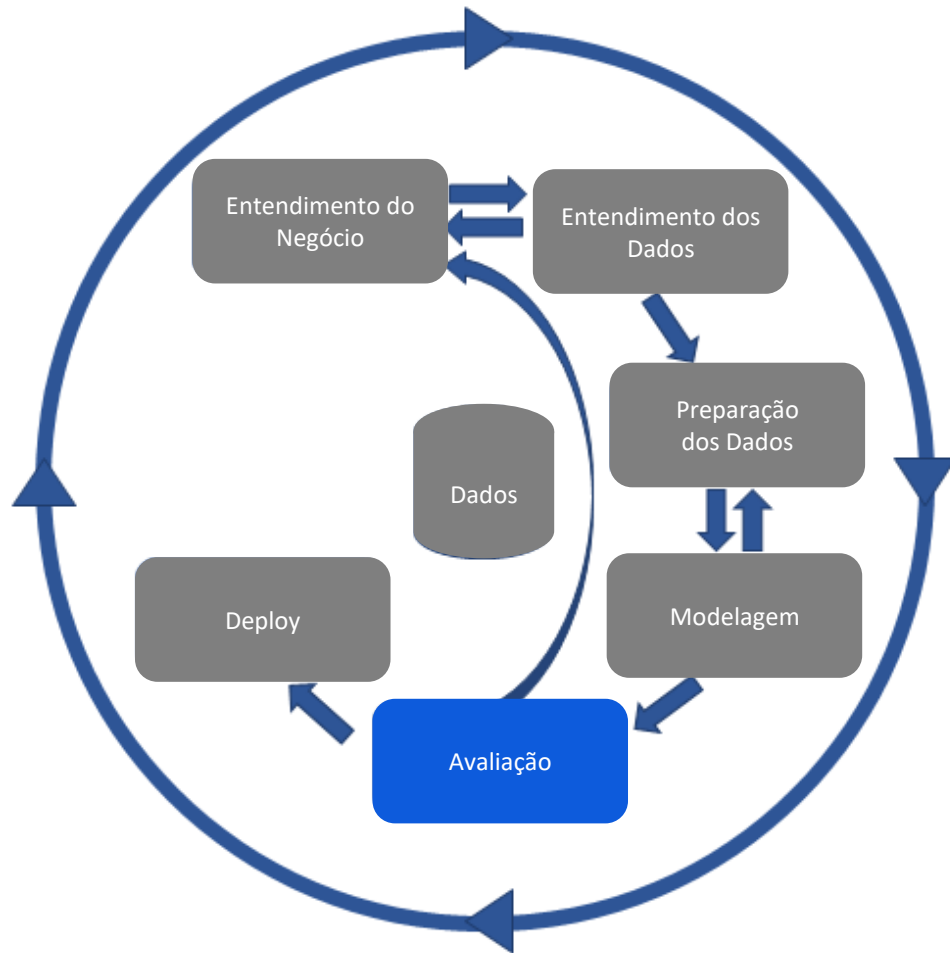


CRISP-DM

Avaliação

- Verificação se os objetivos do negócio foram atendidos
- Análise de resultados e insights
- Decisão sobre a implementação do modelo

Dica: Importância de análises de impactos apropriadas para o negócio

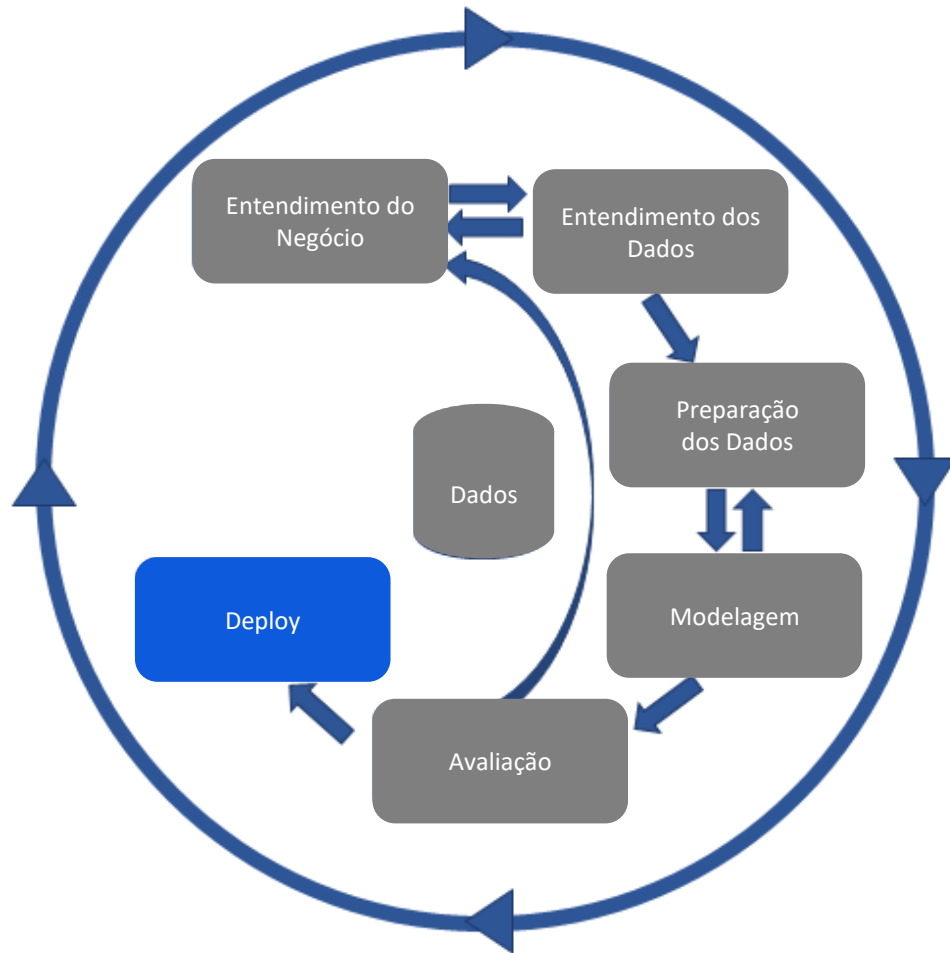


CRISP-DM

Deploy

- Planejamento da Implementação do modelo em produção
- Monitoramento do desempenho do modelo
- Documentação

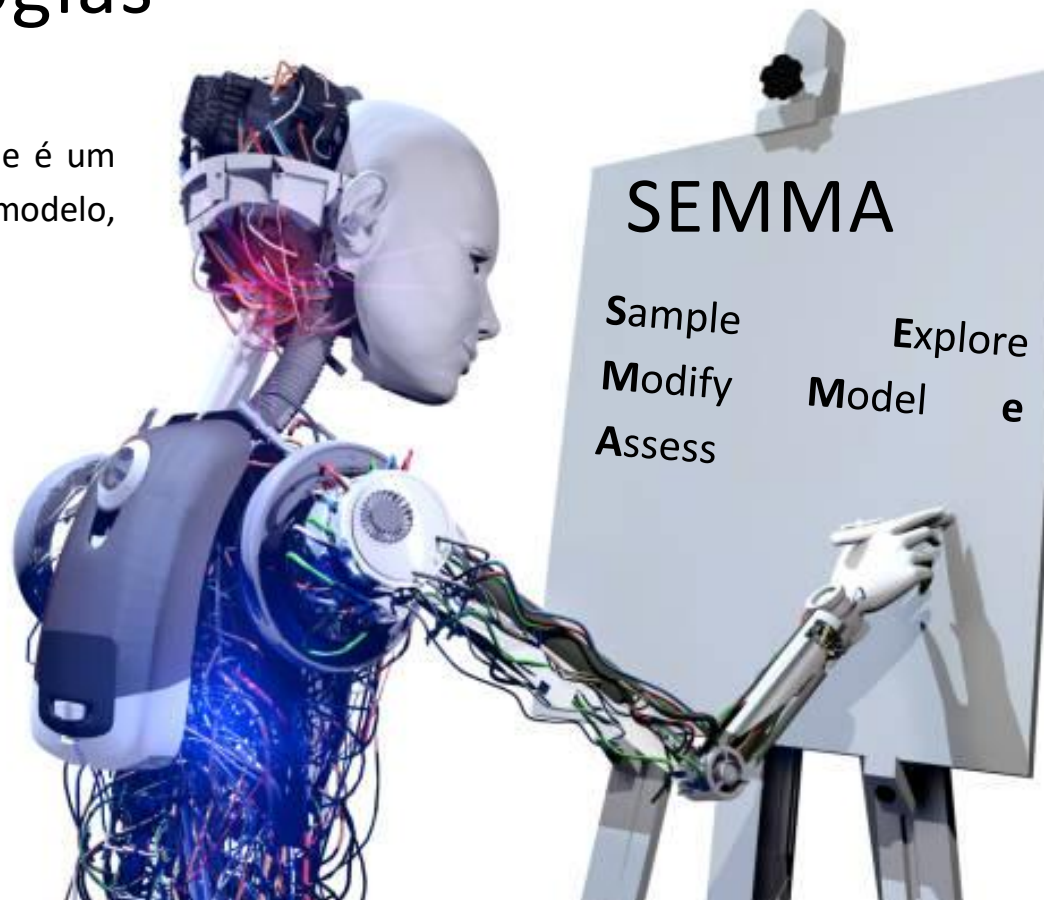
Dica: Importância do Acompanhamento do piloto



Framework Metodologias

O **SEMMA** foi desenvolvido pela *SAS Institute* e é um framework com foco nas tarefas de criação do modelo, deixando as questões de negócio de fora

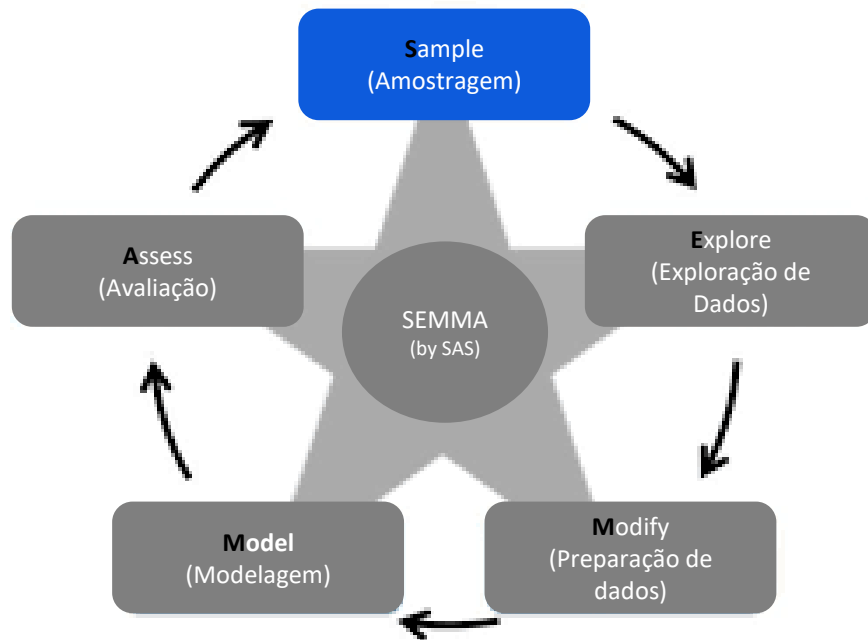
As tarefas são semelhantes com as do CRISP-DM



SEMMA

Sample (Amostragem)

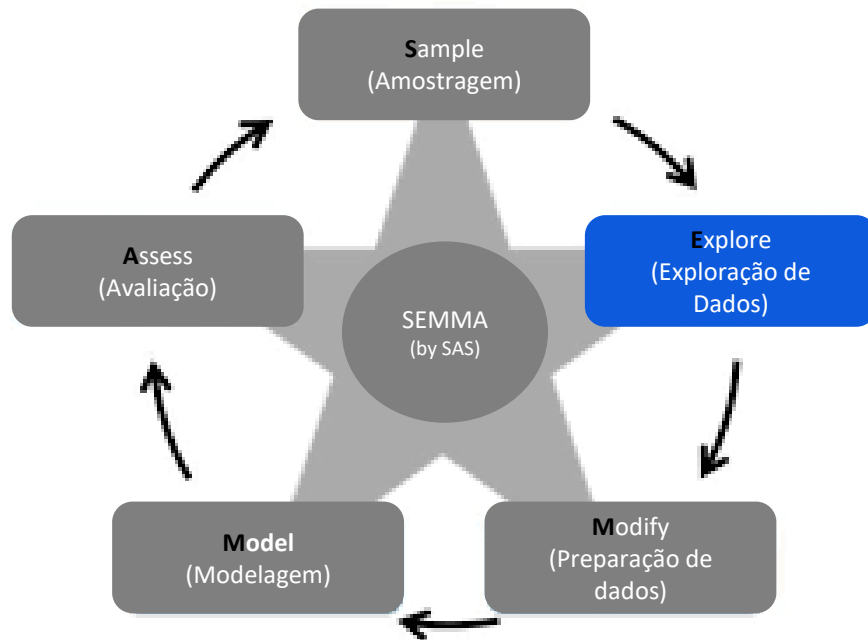
- Seleção de uma amostra representativa dos dados



SEMMA

Explore (Exploração de Dados)

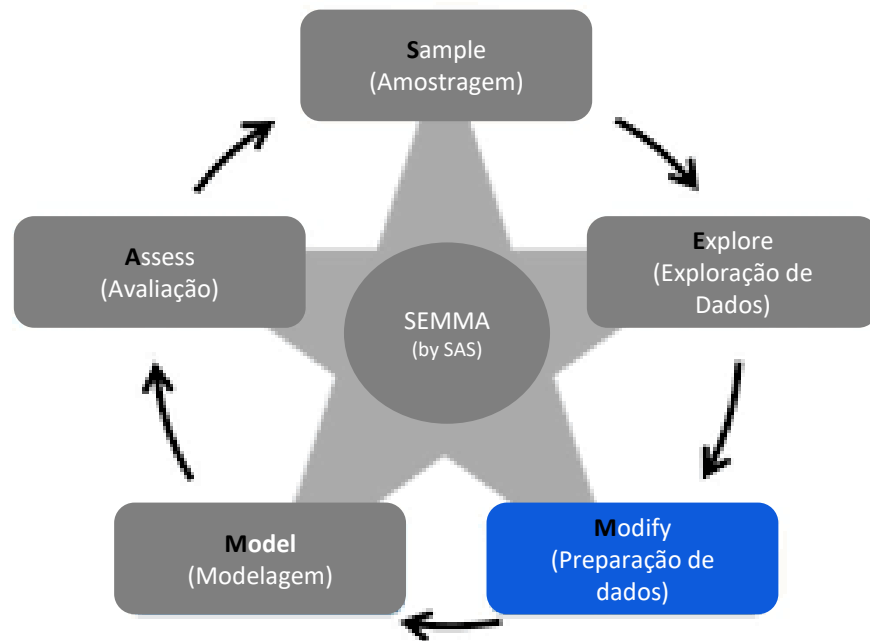
- Análise exploratória para entender padrões e relações nos dados



SEMMA

Modify (Preparação de Dados)

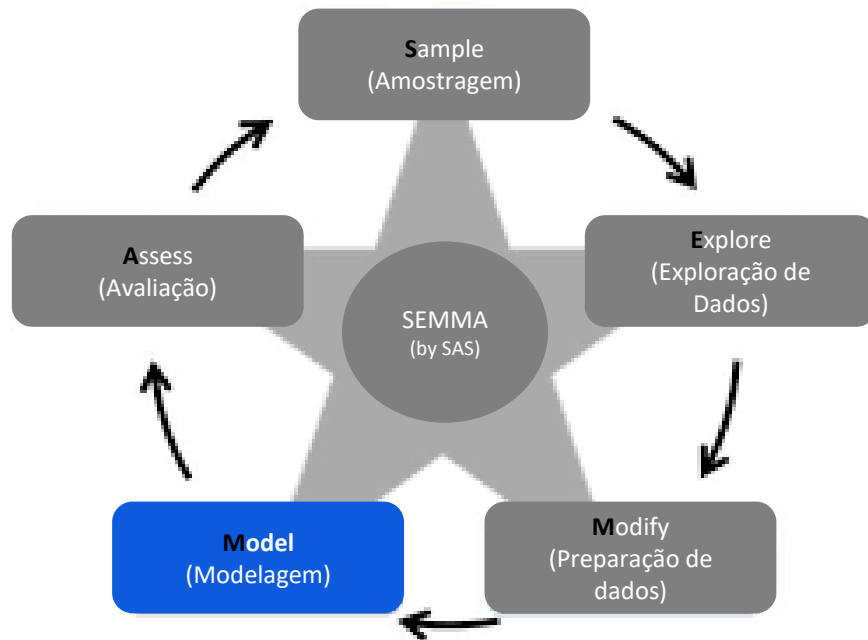
- Transformação e criação de novas variáveis



SEMMA

Model (Modelagem)

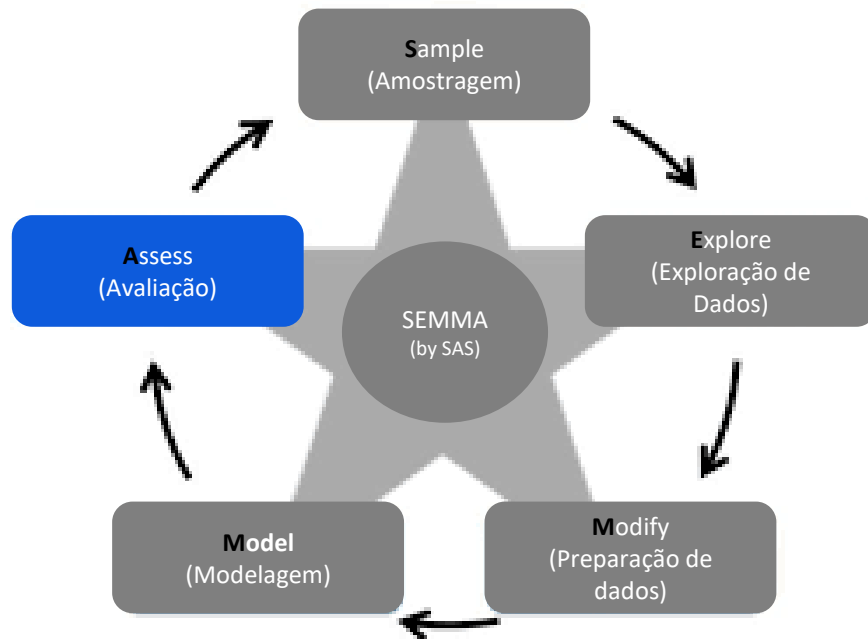
- Aplicação de algoritmos estatísticos e de Aprendizado de Máquina



SEMMA

Assess (Avaliação)

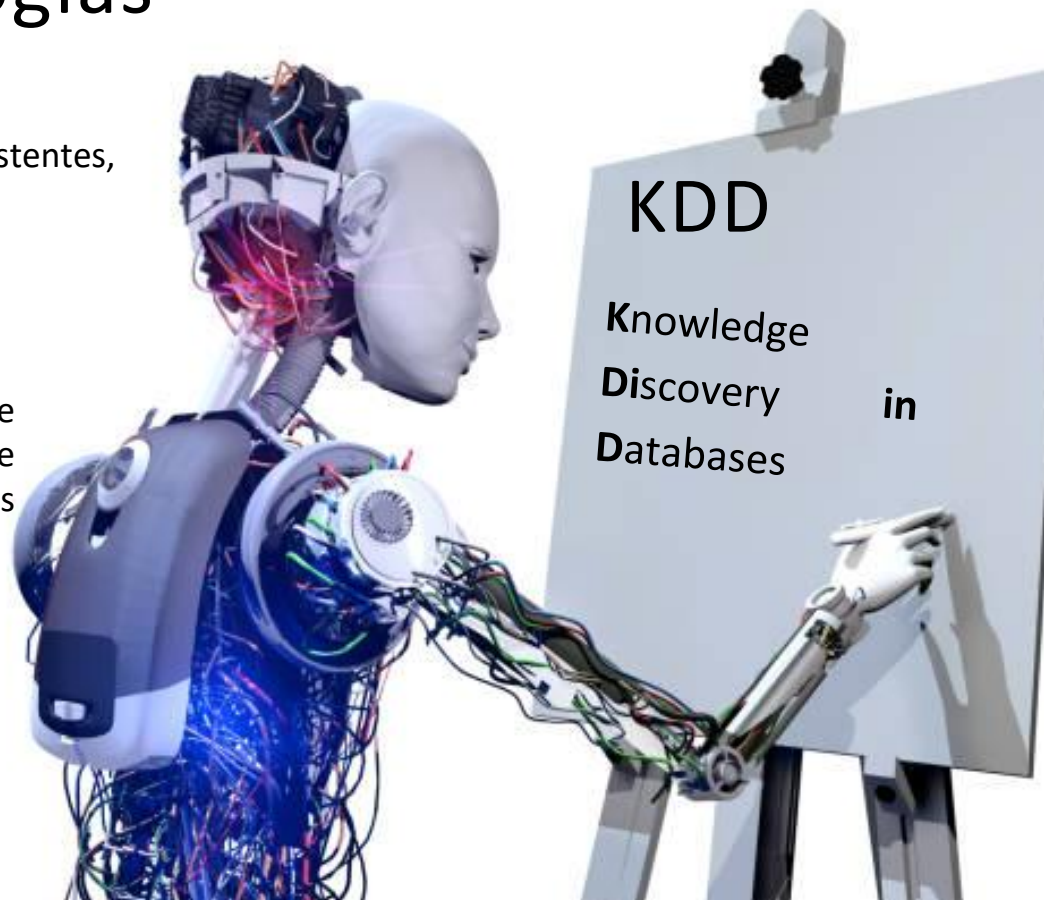
- Medição do desempenho do modelo e validação dos resultados



Framework Metodologias

O **KDD** é um dos métodos mais antigos existentes, tendo sua criação feita em 1980

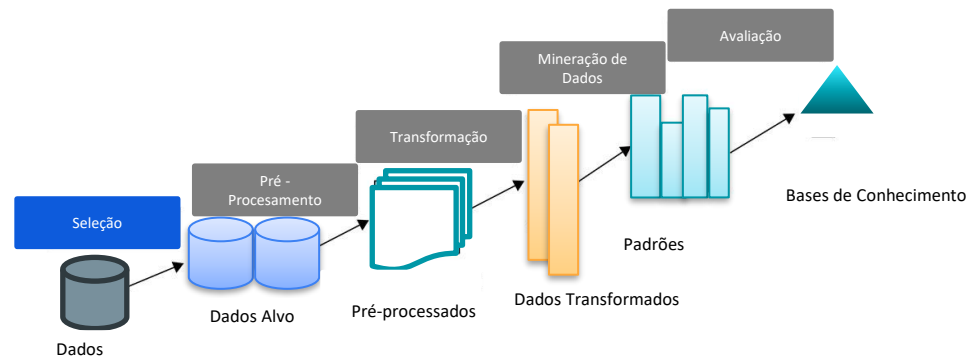
Refere-se ao processo de descoberta de padrões, conhecimentos e informações úteis e potencialmente valiosos a partir de grandes conjuntos de dados



KDD

Seleção dos Dados

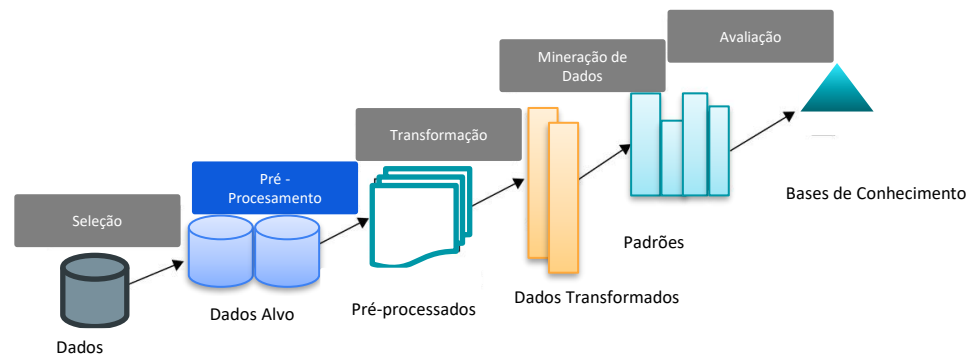
- Escolha dos dados relevantes para análise



KDD

Pré-processamento

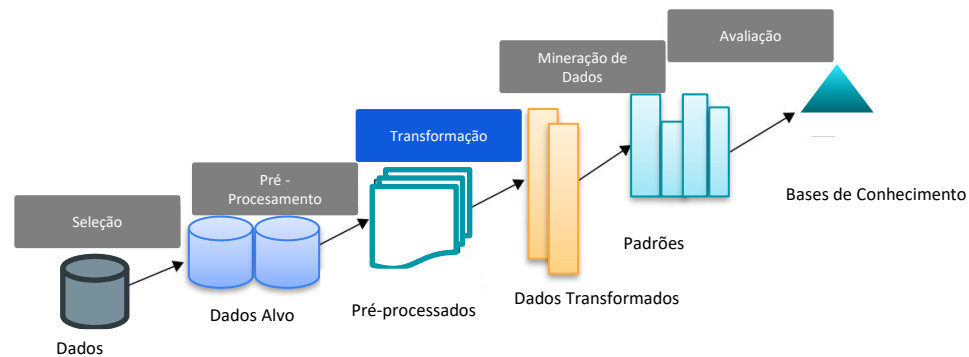
- Limpeza e tratamento de dados inconsistentes ou faltantes



KDD

Transformação

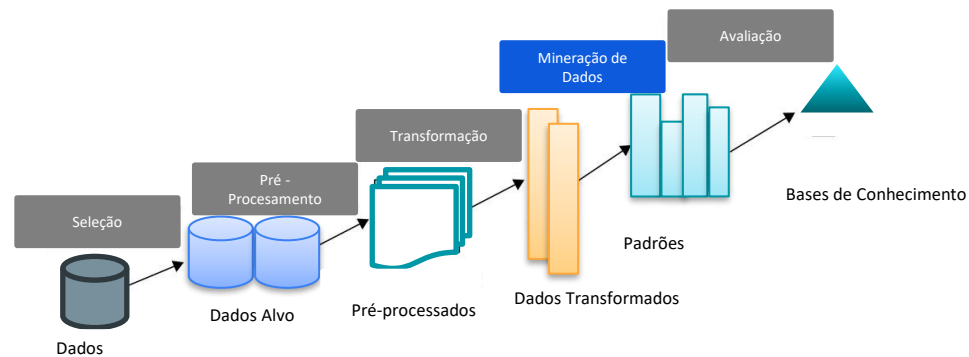
- Criação de novas variáveis e redução de dimensionalidade



KDD

Mineração de Dados

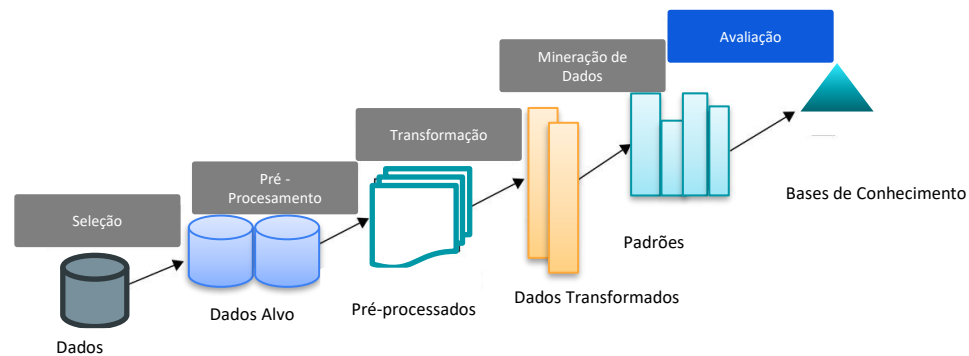
- Aplicação de algoritmos para encontrar padrões



KDD

Avaliação

- Análise dos padrões descobertos e sua utilidade para tomada de decisão



Comparação entre os frameworks

KDD	SEMMA	CRISP-DM
Pre KDD	-	Business Understanding
Selection	Sample	Data Understanding
Pro Processing	Explore	
Transformation	Modify	Data Preparation
Data Mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	-	Deployment

Aspecto	CRISP-DM	SEMMA	KDD
Foco	Processo e Negócio	Dados e Modelagem	Descoberta de Conhecimento
Iterativo ?	Sim	Não	Sim
Contexto de Negócio	Forte ênfase	Limitado	Ênfase Moderada
Flexibilidade	Alta	Moderado	Alta
Facilidade de Uso	Iniciante	Necessário Expertise	Complexidade Moderada