Canal Educação
PROGRAMA DE MEDIAÇÃO TECNOLÓGICA

2ª série

Inteligência Artificial

Pré-processamento de dados em algoritmos inteligentes

Rildo Oliveira

04/04/2024

OBJETO DO CONHECIMENTO: Pré-processamento de dados em algoritmos inteligentes

HABILIDADE:

EF05HI06 - Comparar o uso de diferentes linguagens e tecnologias no processo de

PCRP03 - Identificar, entender e explicar em que situações o computador pode ou não ser utilizado para solucionar um problema.

OBJETIVOS:

- Compreender os conceitos fundamentais do pré-processamento de dados;
- Identificar técnicas comuns de pré-processamento de dados, como normalização, tratamento de valores ausentes e codificação de variáveis categóricas.

DA TEORIA À PRÁTICA: Uso de imagens, texto e conceitos para um melhor entendimento do tema abordado.

Pasta Compartilhada

Ensino Médio 2° ano

https://github.com/rildexter/2ST-IA-2024

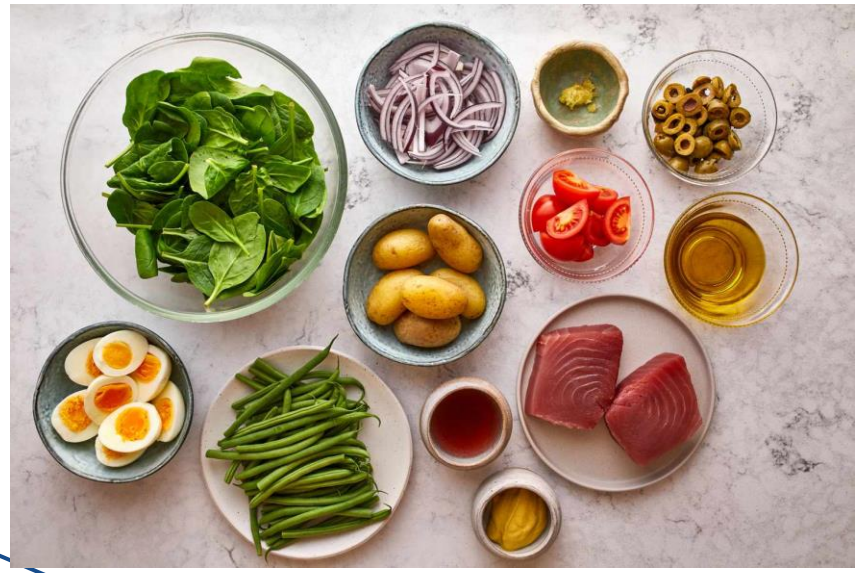
A Importância do Pré-processamento de Dados

A função de um profissional que trabalha com dados, seja focado em ferramentas de análise ou de machine learning, exige uma intensa atividade de pré-processamento de dados. Embora seja uma etapa "menos glamourosa", a maior parte do tempo gasto é ocupado nessa atividade. Estima-se que ela consuma em torno de 70-80% do tempo e esforço total de um projeto de análise de dados.

Objetivo do Pré-processamento de Dados

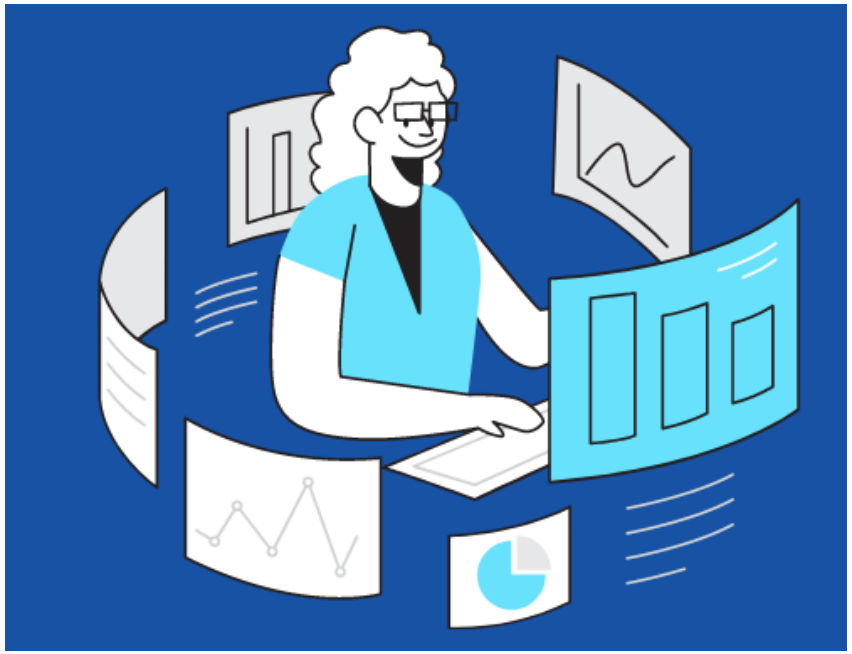
Pré-processamento de dados envolve limpar, organizar e estruturar dados brutos antes da análise.

É como preparar ingredientes antes de cozinhar.



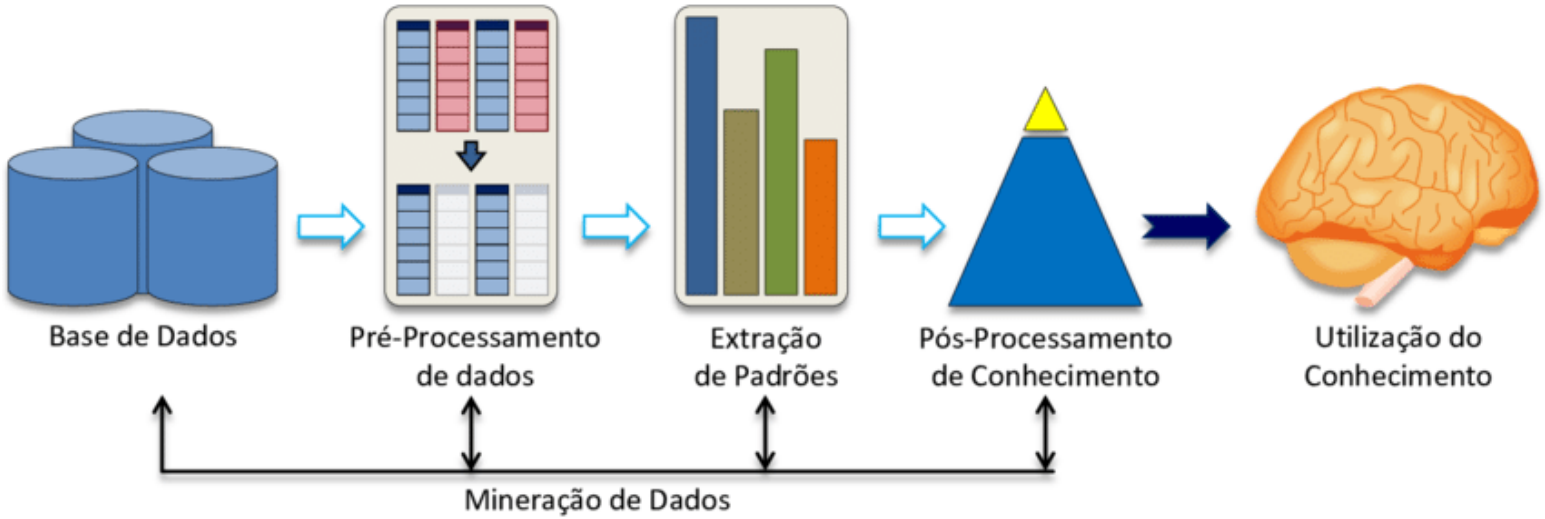
Problemas comuns em Datasets

Entre os principais problemas encontrados dentro de um conjunto de dados, também conhecido como dataset, podemos elencar os atributos com valores faltantes, os outliers e as escalas diferentes para valores iguais.



Processo de Descoberta de Conhecimento

O objetivo de qualquer atividade relacionada a dados, como mineração de dados, é a aquisição de conhecimento. É um conhecimento indispensável à tomada de decisões mais sólidas. A mineração de dados é um processo de negócios para explorar grandes quantidades de dados com foco no reconhecimento de regras e padrões.

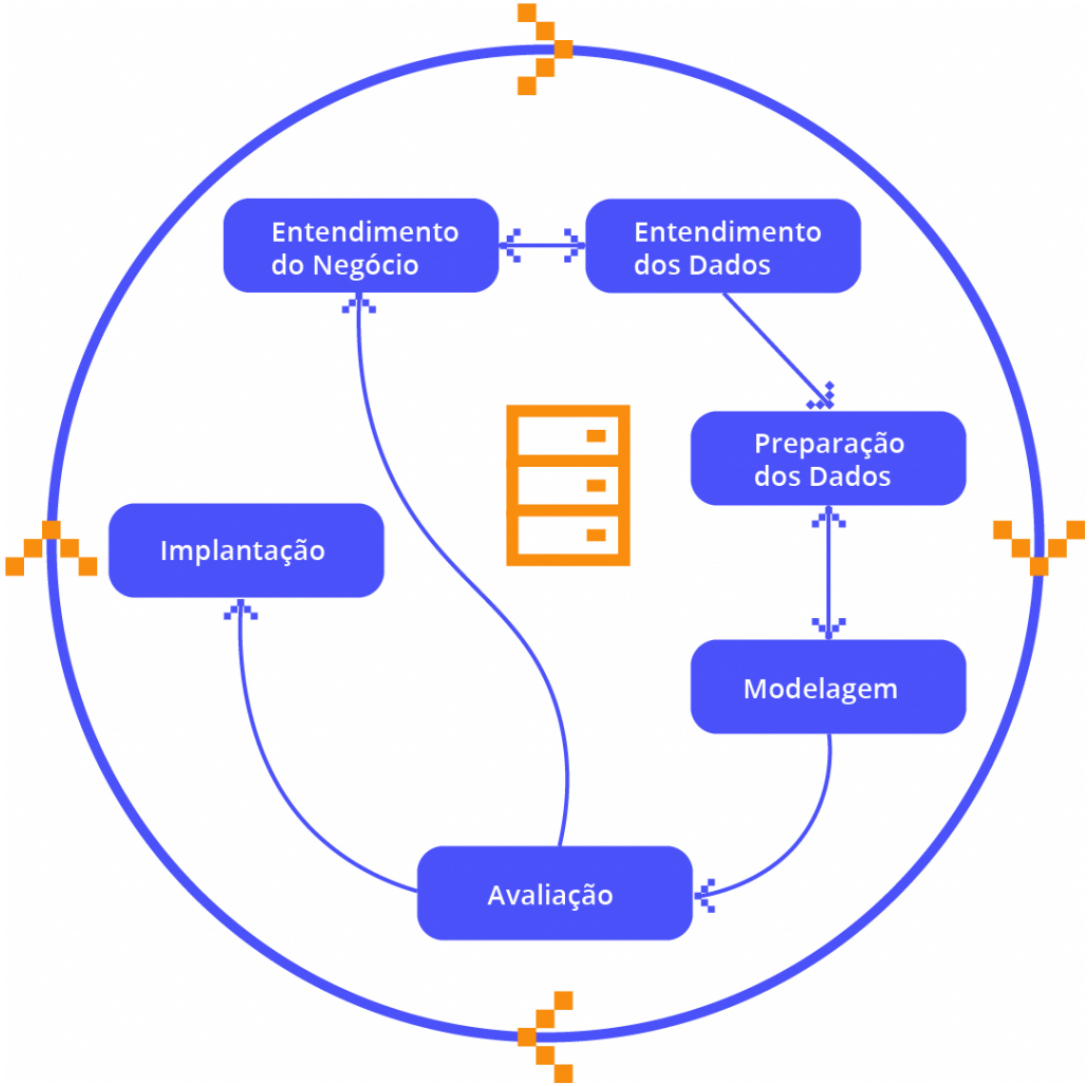


Modelos de Descoberta de Conhecimento

Dentro da mineração de dados, os três principais modelos utilizados para descoberta/aquisição de conhecimento são: KDD, SEMMA e CRISP-DM, sendo o último o mais popular na indústria.

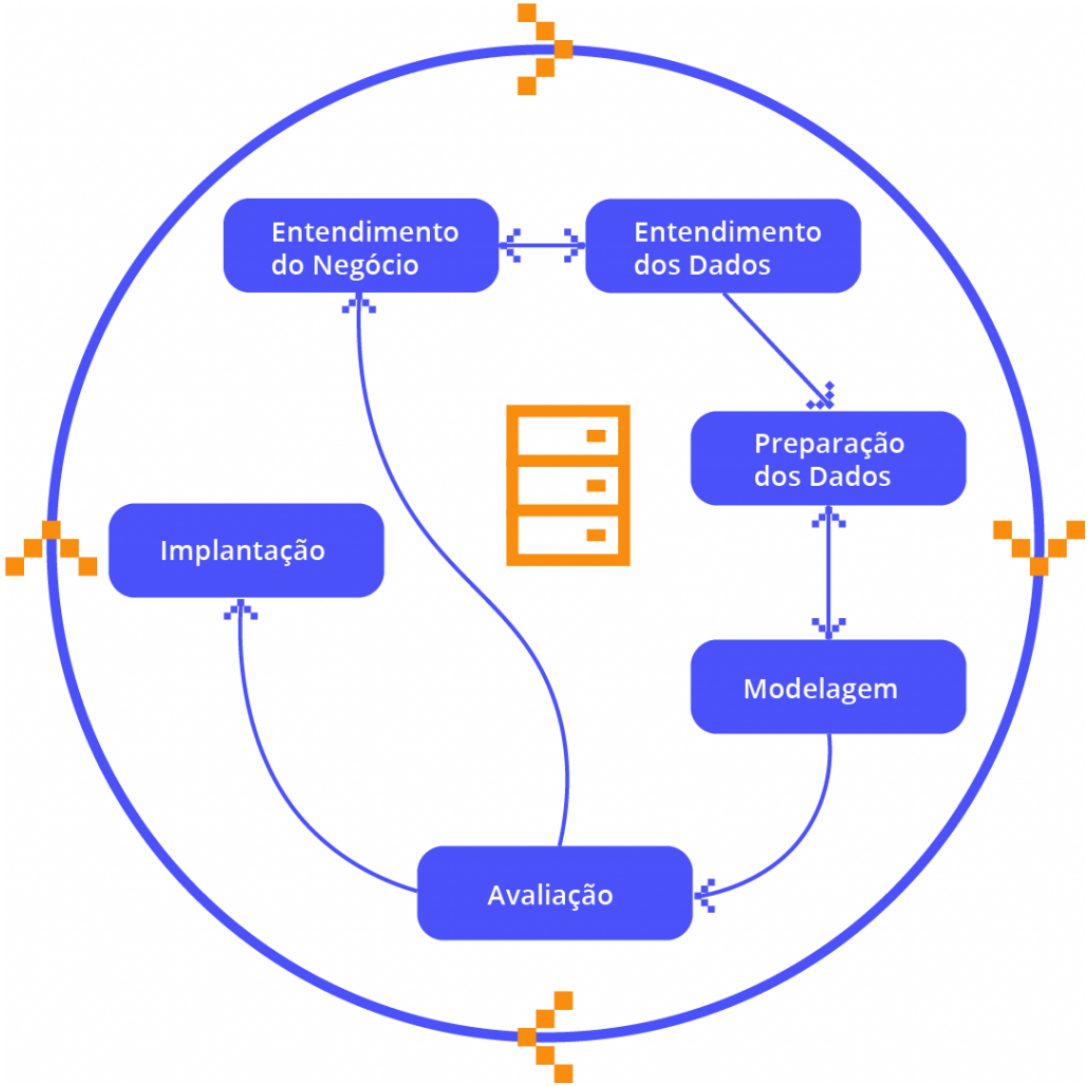
Processo CRISP-DM: Entendimento de Negócio

O processo de CRISP-DM envolve várias etapas, começando com o entendimento do negócio. Esta fase visa entender os objetivos do projeto e os requisitos de negócios.



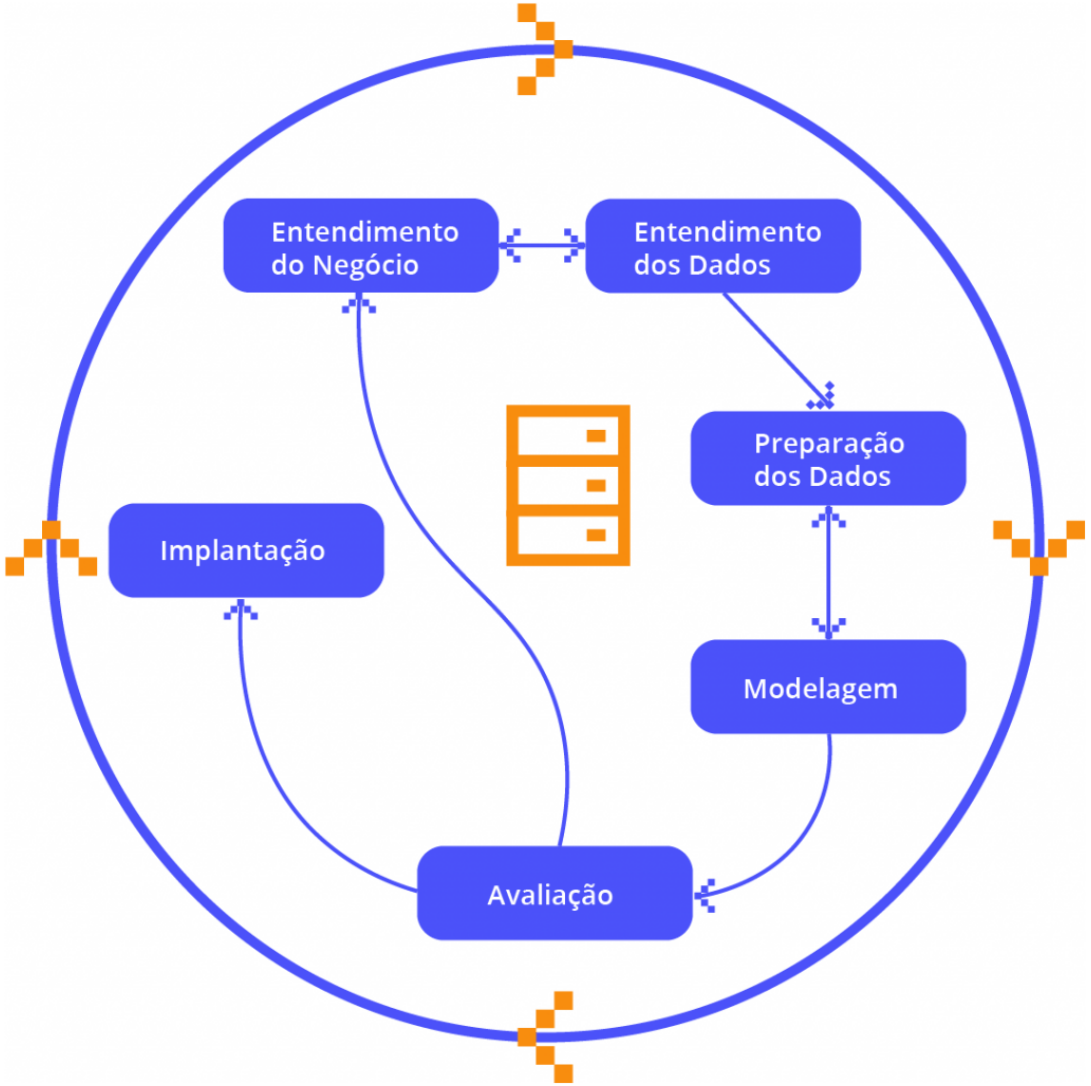
Processo CRISP-DM: Entendimento dos Dados

Após entender o negócio, é importante compreender os dados disponíveis. Isso inclui coletar os dados relevantes, explorá-los para entender sua estrutura e qualidade, e identificar quais dados são necessários para alcançar os objetivos do projeto.



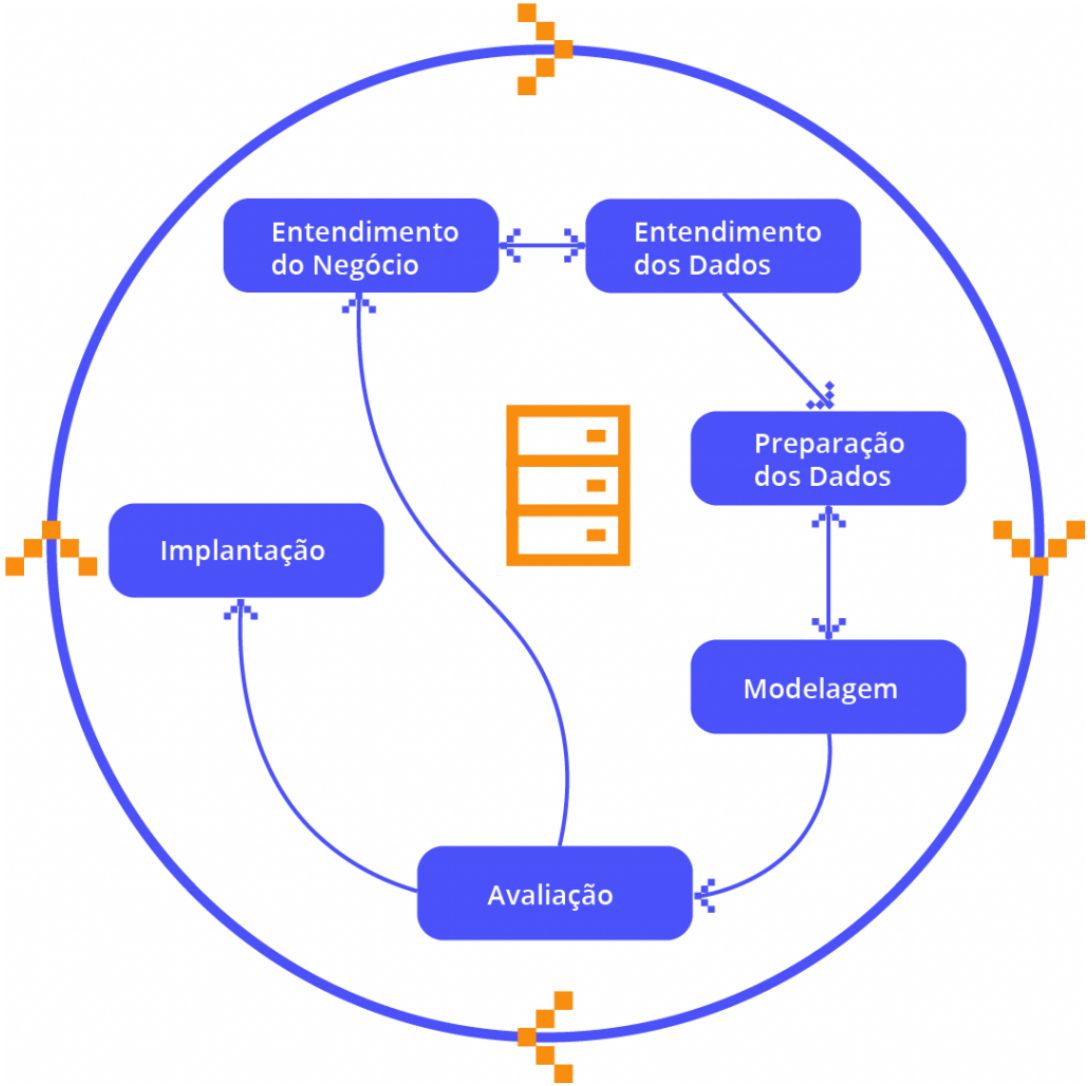
Processo CRISP-DM: Preparação dos Dados

Compreendidos os dados, o próximo passo é prepará-los para análise. Isso pode envolver limpeza de dados, transformação de formatos, seleção de atributos relevantes e integração de diferentes fontes de dados.



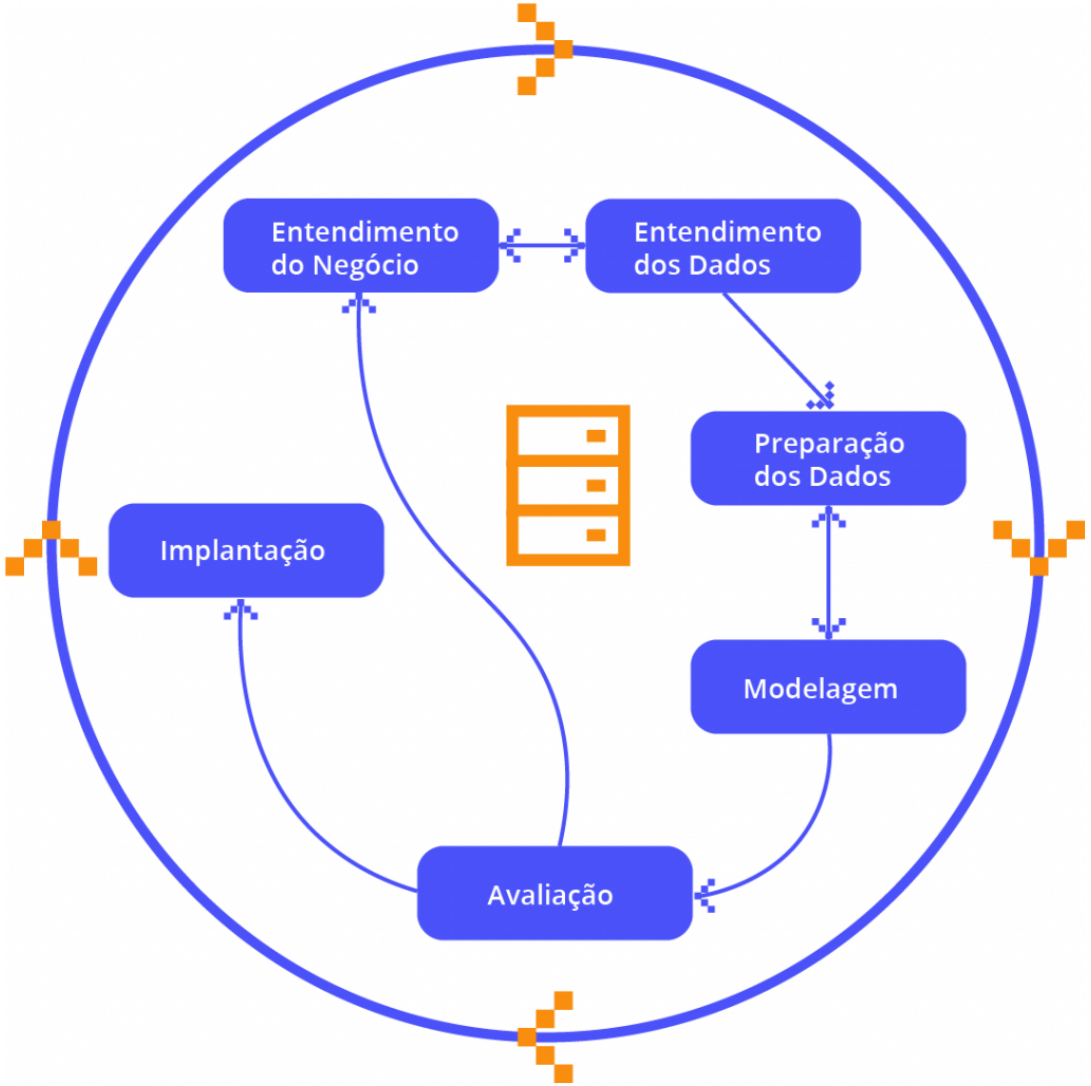
Processo CRISP-DM: Modelagem dos Dados

A etapa de modelagem envolve a seleção e aplicação de técnicas de modelagem para construir um modelo preditivo ou descritivo com base nos dados preparados.



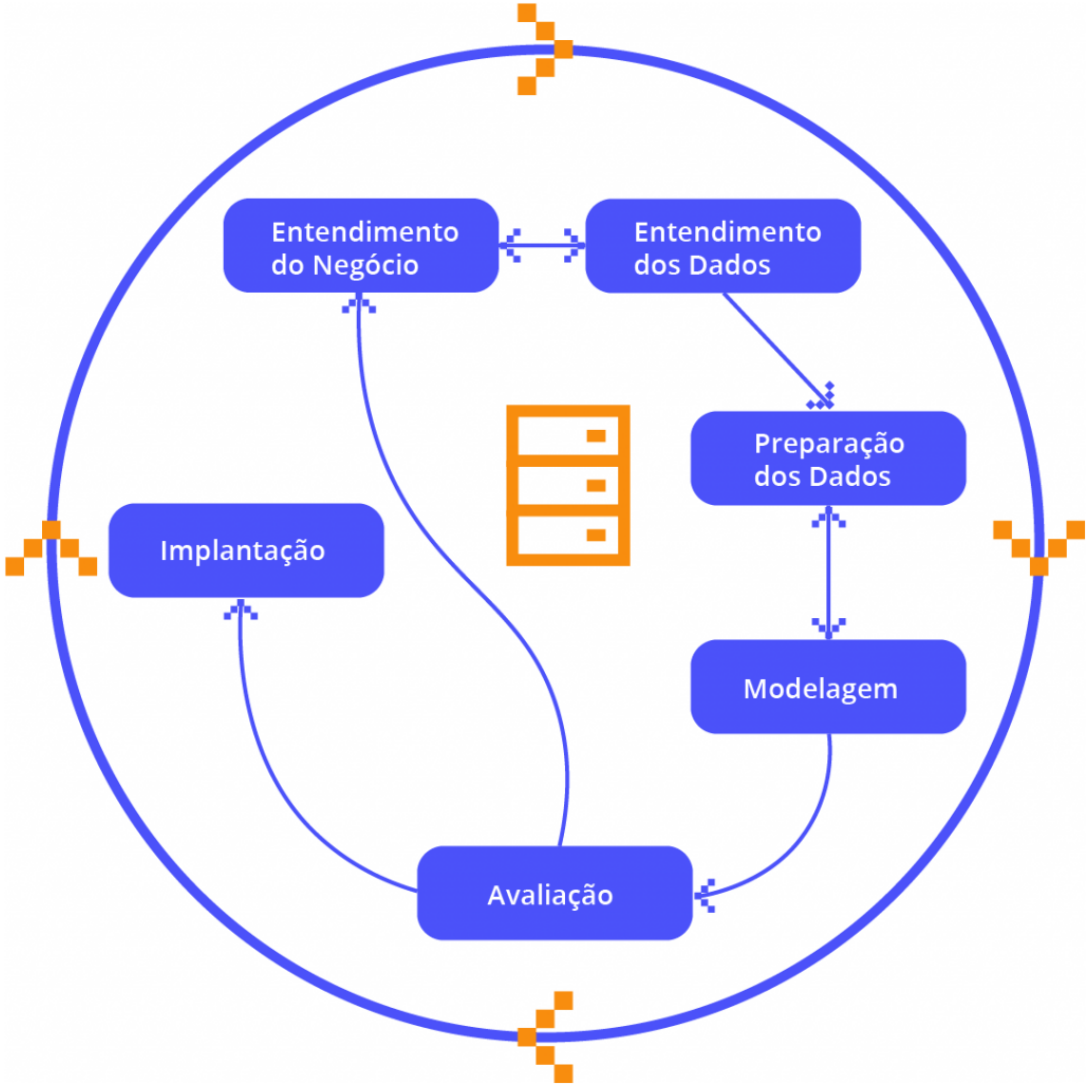
Processo CRISP-DM: Avaliação dos Resultados

Depois de construir o modelo, é essencial avaliar sua eficácia e precisão. Isso pode ser feito usando métricas de desempenho e testes estatísticos para garantir que o modelo atenda aos requisitos do projeto.



Processo CRISP-DM: Inplantação

Finalmente, o modelo desenvolvido é implantado em um ambiente de produção, onde ele é usado para tomar decisões ou gerar insights de negócios. Este processo pode envolver integração com sistemas existentes e monitoramento contínuo do desempenho do modelo.



Estrutura de Dados

Antes de iniciar o pré-processamento em si, é importante conhecer as principais estruturas dos dados. De modo geral, elas podem ser classificadas em três categorias:

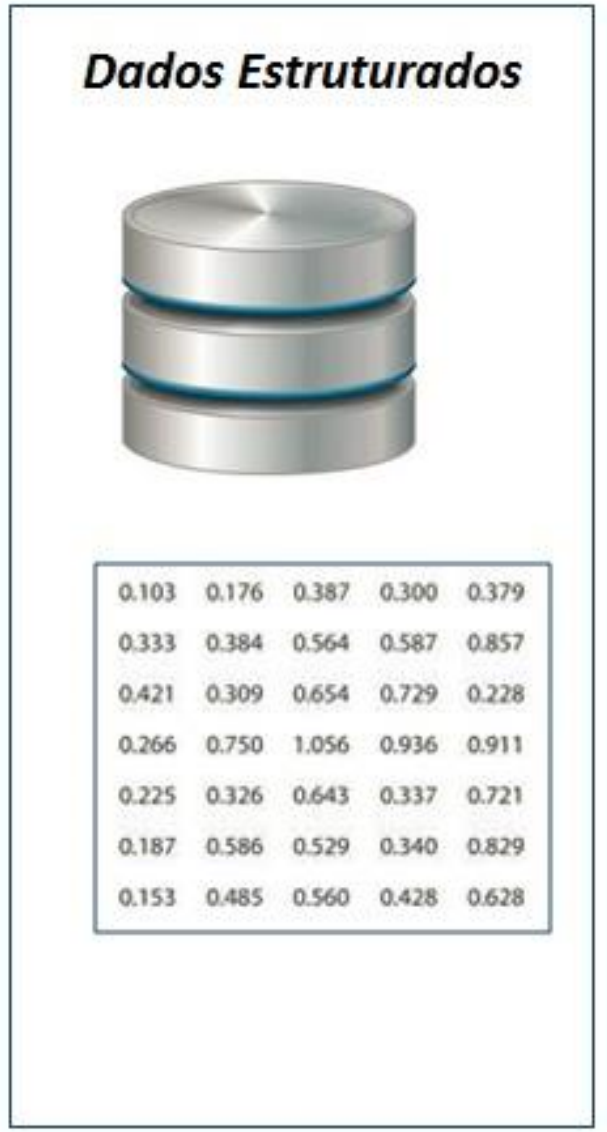
- Estruturados;
- Semiestruturados;
- Não estruturados.

Dados Estruturados

São os dados que contêm uma organização rígida e previamente planejada. Normalmente são "etiquetados" em linhas e colunas que identificam suas características a respeito de determinados assuntos. Podem ser organizados em blocos semânticos (relações) e definição de descrições para dados de um mesmo grupo (atributo).

Exemplos de Dados Estruturados

Exemplos de dados estruturados incluem bancos de dados relacionais, planilhas Excel, arquivos CSV, entre outros. Este tipo de estrutura é comum em ambientes corporativos e é altamente organizado para facilitar a manipulação e análise dos dados.



Dados Semiestruturados

Os dados semiestruturados possuem uma estrutura, porém, não seguem os formatos formais dos modelos associados a bancos de dados relacionais. Geralmente, eles contêm marcadores, como tags, para separar elementos semânticos e criar hierarquias para os registros e campos.

Exemplos de Dados Semiestruturados

Exemplos comuns de dados semiestruturados incluem arquivos XML, JSON, HTML, entre outros. Esses formatos são amplamente utilizados na web e em ambientes digitais para armazenar e compartilhar informações de maneira flexível e adaptável.

Dados Não Estruturados

Os dados não estruturados são caracterizados pela ausência de uma organização clara. Para extrair insights desses dados, é necessário realizar um intenso pré-processamento para recuperar a informação útil.

Exemplos de Dados Não Estruturados

Exemplos comuns de dados não estruturados incluem documentos de texto, áudio, imagens e outros formatos onde a informação não está organizada em tabelas ou esquemas definidos. O processamento desses dados requer técnicas especializadas para extrair significado e insights.


DADOS
Estruturados
Semi Estruturados
Não Estruturados

Técnicas de Pré-Processamento

O pré-processamento de dados é importante para transformar dados brutos em formatos úteis e eficientes. Esse processo envolve três principais passos: limpeza de dados, transformação de dados e redução de dados, cada um com suas próprias atividades.

Limpeza de Dados

A limpeza de dados é essencial para lidar com partes irrelevantes ou ausentes nos dados originais. Isso inclui o tratamento de dados faltantes, redução de ruídos, identificação e remoção de valores aberrantes, e resolução de inconsistências.

Dados Faltantes

Os dados faltantes são comuns em conjuntos de dados e exigem estratégias específicas para lidar com eles. Opções incluem remover registros com atributos nulos, preencher com a média ou mediana dos valores do mesmo atributo, ou preencher com os valores mais frequentes no dataset.

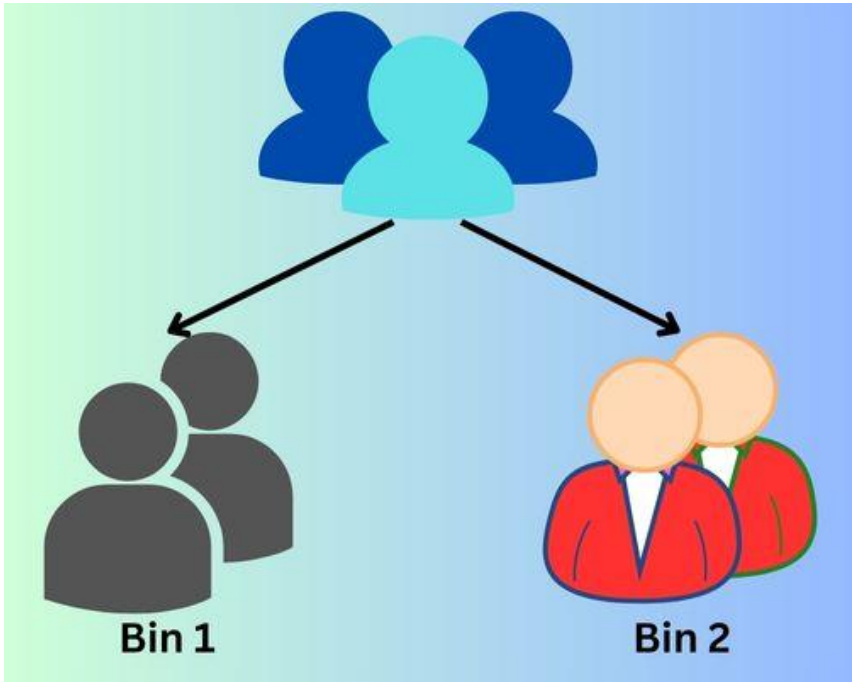
Área (m^2)	Andares	Preço (R$)
50	2	280.000
120	3	400.000
100	NaN	390.000
35	NaN	150.000

Como Tratar Dados Ausentes

Dados incompletos podem distorcer análises. Métodos como remoção de registros ou preenchimento com valores médios são comuns para lidar com essa questão.

Dados Ruidosos

Dados ruidosos são sem sentido e podem ser tratados com técnicas como o método de Binning, que suaviza os dados dividindo-os em intervalos conhecidos como compartimentos e substituindo-os por um valor geral.



Transformação de Dados

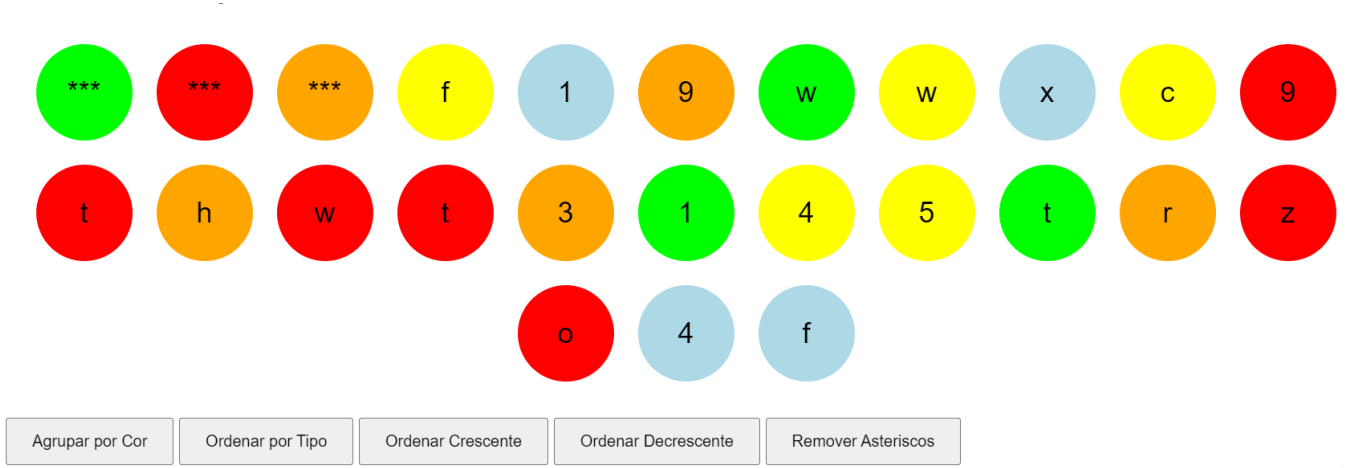
De forma geral podemos entender o pré-processamento de dados na figura a seguir:

Conclusão

O pré-processamento de dados pode ser uma etapa trabalhosa, mas é fundamental para garantir a qualidade e eficácia da análise. Sem ele, os modelos podem ser prejudicados pela entrada de dados inadequados.

Exemplo na prática

h lio/pre-



Referencias

1. Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques. Elsevier.

2. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data mining: practical machine learning tools and techniques. Morgan Kaufmann.

3. Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media, Inc.

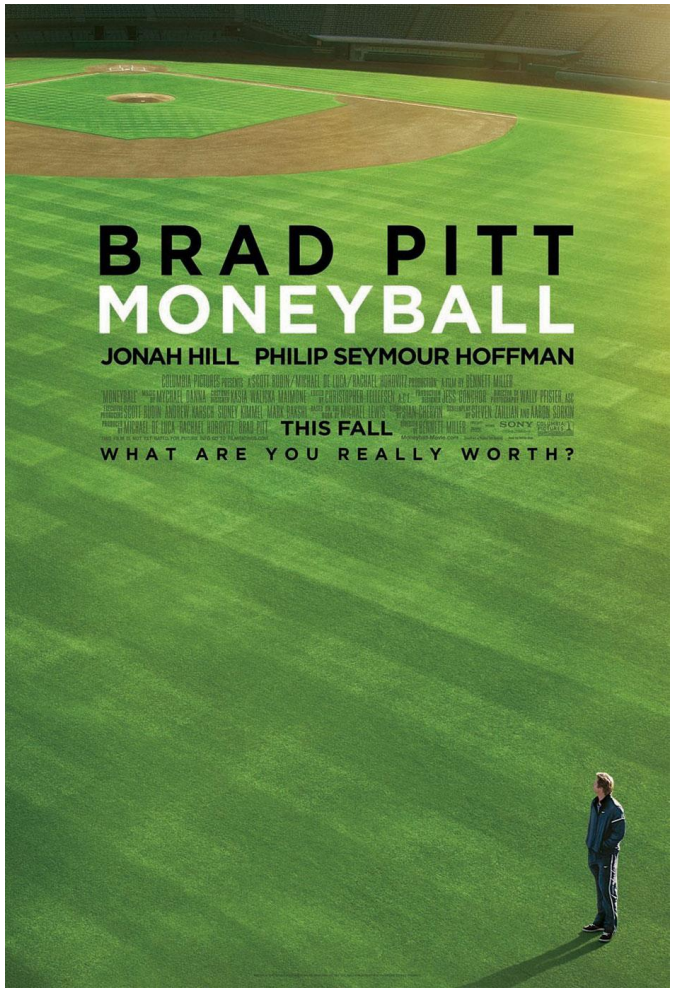
4. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

5. García, S., Luengo, J., & Herrera, F. (2015). Data preprocessing in data mining (pp. 35-64). Springer, Cham.

Dica de cinema!

O Homem que Mudou o Jogo (2011)

O filme é baseado em uma história real e segue a trajetória de Billy Beane, gerente geral do time de beisebol Oakland Athletics, que utiliza análise estatística avançada para montar uma equipe competitiva com um orçamento limitado. Esta história fascinante mostra como o pré-processamento de dados e a análise inteligente podem revolucionar uma indústria tradicional, oferecendo insights valiosos para tomadas de decisão estratégicas.



ATÉ A PRÓXIMA AULA!