Inteligência Artificial

Projeto da Disciplina: Previsão Preço imóveis de São Caetano do Sul
Etapa: *Data Preparation*

Prof. Dr. Ivan Carlos Alcântara de Oliveira

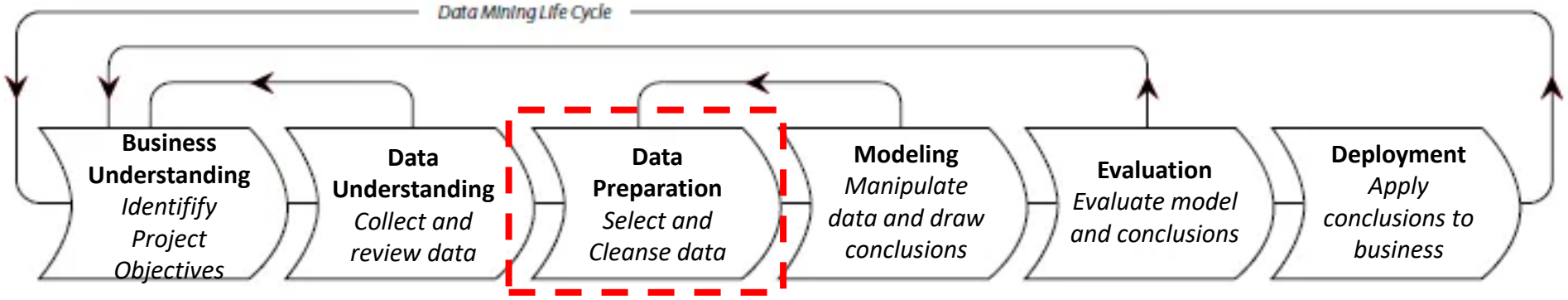
https://orcid.org/0000-0002-6020-7535

Preparação dos dados e Pré-Processamento

Ciclo de Vida de Projeto de Data Science

CRISP-DM

- **CRISP-DM (*Cross Industry Standard Process for Data Mining* -** Processo Padrão Interindústrias para Mineração de Dados) é um processo de fases bastante aceito na indústria para representar um ciclo completo de Ciência e Análise de Dados, incluído a aplicação de modelos de Aprendizado de Máquina.



Auxiliar na Identificação	Obtenção dos dados	Pré-processamento	Seleção do modelo	Validação do modelo	Aplica solução obtida ao negócio
Compreensão	Integração	Transformação de Variáveis	Cross-Validation	Otimização	Desenvolvimento de Aplicação Analítica
Pesquisas e Estudos	Análise Exploratória	Seleção de Atributos	Métricas de performance		Geração de relatório
Reunião com Especialistas		Redução de Dimensionalidade	Otimização		Construção de Dashborads etc.
Definição dos Atributos		Amostragem			
Definição medida de erro					

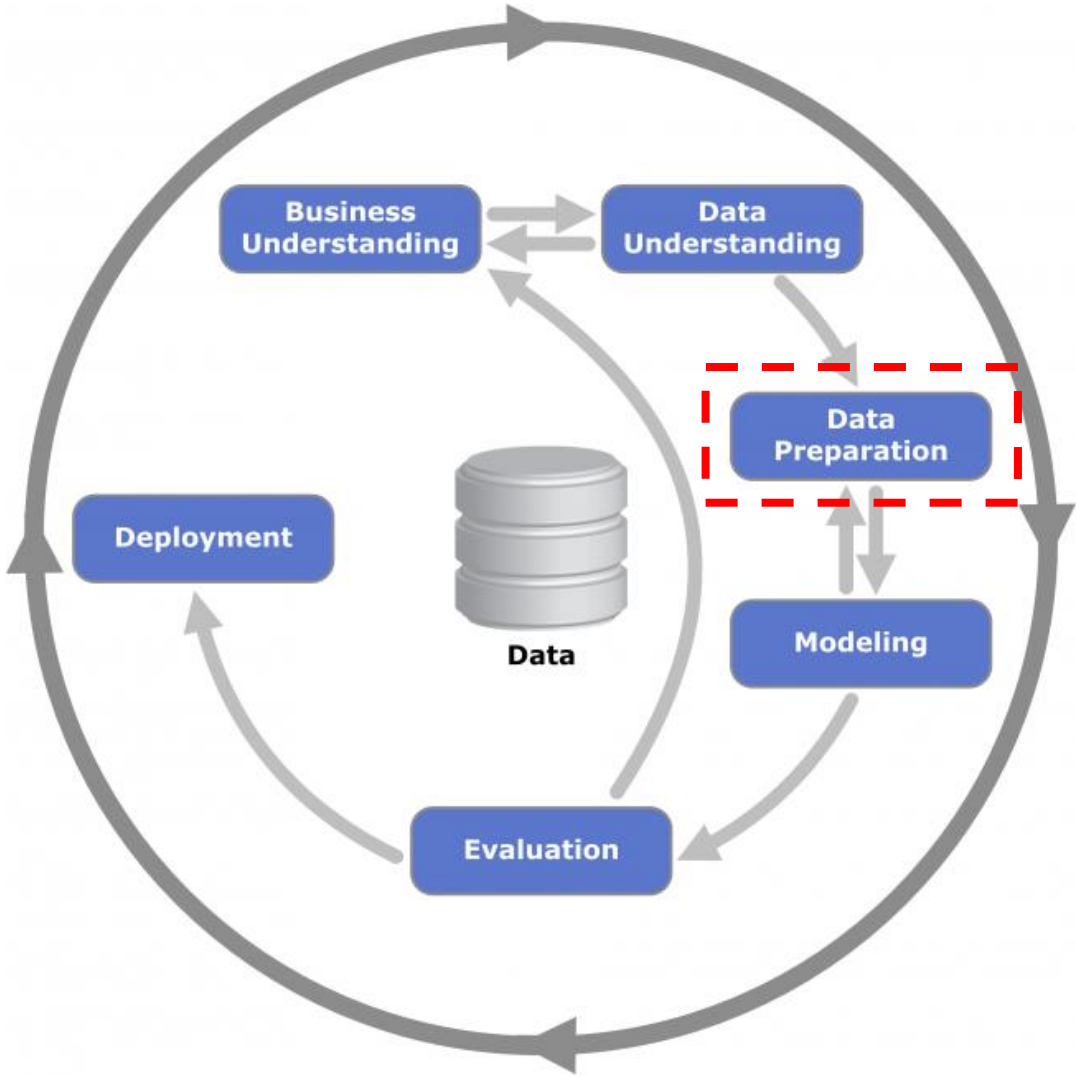
Ciência de Dados

Ciclo de Vida de Projeto de *Data Science*

3. **Preparação dos Dados**

Data Wrangling: pré-processamento, transformação, limpeza dos dados, etc.

Até aqui, aproximadamente 70% do tempo de um Cientista de Dados.



Data Preparation

Previsão de Valor (Compra e Venda) de Imóvel – São Caetano do Sul

Considere que fazemos parte de uma imobiliária que tem imóveis localizados UNICAMENTE em São Caetano do sul. Ela quer disponibilizar um preditor aos seus clientes para sugerir o valor de venda ou compra de um imóvel.

- Algumas perguntas de negócio poderiam ser:
- Quero comprar um imóvel seguindo algumas características, é possível prever qual o preço ou valor do imóvel?
- Quais dados devem ser capturados para avaliar o valor a pagar?
- Posso montar um conjunto de dados para gerar um modelo preditivo que permite um vendedor avaliar o preço que deve ser ofertado do imóvel? Com esse mesmo conjunto de dados, caso eu queira comprar, é possível utilizá-lo para prever o valor de compra?

Data Preparation

Previsão de Valor (Compra e Venda) de Imóvel – São Caetano do Sul

- A coleta de dados foi realizada por alunos de um curso de Ciência da Computação e Sistemas de Informação em março de 2021 de uma instituição acadêmica fazendo uso de coleta de dados manuais e *webscraping* nos sites:
  - www.vivareal.com.br/
  - www.creditas.com
  - www.imovelweb.com.br
- Nesse cenário, foram coletados e montados 23 datasets por diferentes elementos, dos quais 20 deles constam em uma pasta de nome bd_SCS.

Data Preparation

Previsão de Valor (Compra e Venda) de Imóvel – São Caetano do Sul

Os atributos considerados durante essa coleta de dados foram:

- Tipo: Casa (1), Apto (2).
- Tempo do imóvel em anos: (0 = novo)
- Localização: Rua Luís Cavana, 70 - Centro, São Caetano do Sul – SP.
- Área total (m2):
- Área útil (m2):
- Número de quartos:
- Suíte: Sim (1) ou Não (0)

Data Preparation

Previsão de Valor (Compra e Venda) de Imóvel – São Caetano do Sul

Os atributos considerados durante essa coleta de dados foram:

- Número de Banheiros
- Número de Vagas para Carros
- Academia: Sim (1) ou Não (0)
- Churrasqueira: Sim (1) ou Não (0)
- Elevador: Sim (1) ou Não (0)
- Salão de Festas: Sim (1) ou Não (0)
- Piscina: Sim (1) ou Não (0)
- Valor do IPTU: em R$
- Valor do Condomínio: em R$.
- Valor do Imóvel: em R$.

Data Preparation

Previsão de Valor (Compra e Venda) de Imóvel – São Caetano do Sul

- Em um primeiro momento, o professor da disciplina fez uma integração de três *datasets*, sendo dois arquivos texto com extensão “csv”: bdSCS_1_csv.csv e bdSCS_2_csv.csv, e um do Excel com extensão xlsx: bdSCS_3_excel.xlsx, gerando ao final único dataset com nome bdSCS_final_csv.csv.
- No processo de integração, foram realizadas manipulações nos conteúdos internos de cada um dos arquivos fazendo uso de algumas bibliotecas do Python, dentre elas: pandas, numpy, matplotlib, seaborn e re (expressões regulares), chegando-se a um resultado final de 13 atributos para bdSCS_final_csv.csv, sendo 12 deles candidatos a atributos de entrada e 1 o atributo alvo (Valor do Imóvel).
- Após esse processo, os atributos resultantes foram:

Data Preparation

Previsão de Valor (Compra e Venda) de Imóvel – São Caetano do Sul

- Tipo: Casa (1), Apto (2).
- Bairro.
- Área total (m2)
- Número de quartos
- Número de Banheiros
- Número de Vagas para Carros
- Academia: Sim (1) ou Não (0)
- Churrasqueira: Sim (1) ou Não (0)
- Elevador: Sim (1) ou Não (0)
- Salão de Festas: Sim (1) ou Não (0)
- Piscina: Sim (1) ou Não (0)
- Valor do Condomínio: em R$.
- Valor do Imóvel: em R$.

Data Preparation

Previsão de Valor (Compra e Venda) de Imóvel – São Caetano do Sul

- A partir do apresentado, considere os arquivos:

1. IA_EAD_Atividade_DataPreparation.ipynb: template da Preparação dos Dados em arquivo do Jupyter Notebook/Google Colab.
2. bdSCS_final_profIvan.csv: arquivo texto que integra os arquivos: bdSCS_1_csv.csv, bdSCS_2_csv.csv e bdSCS_3_excel.xlsx, contendo 6381 registros e os atributos (colunas): tipo, bairro, area_total, quartos, banheiros, vagas, academia, churrasqueira, elevador, salao_festa, piscina, condominio e preco .
3. Template do Relatório do Projeto (Continuação) (IA_EAD_relatorio_projetoImoveisSCS_dataPreparation_20241.docx).

Data Preparation

Previsão de Valor (Compra e Venda) de Imóvel – São Caetano do Sul

- Utilize o arquivo do Jupyter Notebook “IA_EAD_Atividade_DataPreparation.ipynb” e faça a preparação dos dados utilizando o *dataset* RESULTANTE que você obteve após a integração do arquivo “bdSCS_final_csv.csv” com os outros dois arquivos, dos 20 disponibilizados, realizada na etapa anterior da disciplina.
- **Para fazer a preparação dos dados**, basta: **ler o *dataset* gerado por você (obtido na etapa anterior da disciplina)**, reexecute a análise exploratória que consta no “IA_EAD_Atividade_DataPreparation.ipynb.
- **Fique atento e veja se o seu dataset NECESSITA DE ALGUMA preparação adicional além daquelas indicadas e realizadas por mim no meu *dataset*. Percebendo essa necessidade, realize essa preparação e informe no relatório, ok!**

Data Preparation

Previsão de Valor (Compra e Venda) de Imóvel – São Caetano do Sul

- Atualizar o Relatório no Template “IA_EAD_relatorio_projetoImoveisSCS_dataPreparation_20241.docx” com os dados internos da etapa *Data Preparation*, conforme consta no seu conteúdo interno.
- Entregar um único arquivo compactado com os itens:
  - A sua base de dados: bdSCS_final_csv.csv (ou com o nome que você atribuiu).
  - A base de dados final com os dados preparados e colunas ok: imoveis_SCS_preparadosColOk.csv.
  - Relatório do projeto preenchido: IA_EAD_relatorio_projetoImoveisSCS_dataPreparation_20241.docx
  - Arquivo do jupyter notebook com a execução da preparação dos dados.

Bibliografia

BÁSICA:

- AGGARWAL, Charu C. **Artificial Intelligence: A Textbook**. New York: Springer: 2021.
- CHOLLET, François. **Deep Learning with Python, 2ed.** Shelter Island: Manning, 2021.
- GÉRON, Aurélien. **Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**, 2 ed**.** Sebastopol: O'Reilly, 2019.

COMPLEMENTAR:

- GOODFELLOW, Ian; BENGIO, Yoshua, COURVILLE, Aaron. **Deep Learning**. Cambridge: MIT Press, 2016.
- RASCHKA, Sebastian; MIRJALILI, Vahid. **Python Machine Learning**. 3 ed. Birmingham: Packt, 2017.
- RUSSEL, Stuart; NORVIG, Peter. **Artificial Intelligence: A Modern Approach**. 3 ed. Upper Saddle River: Pearson, 2010.
- TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introduction to Data Mining**. 2 ed. Upper Saddle River: Pearson, 2018.
- VANDERPLAS, Jake. **Python Data Science Handbook**. Sebastopol: O'Reilly, 2017.

ADICIONAIS:

- FACELI, Katti et al. **Inteligência artificial: uma abordagem de aprendizado de máquina**. 2ª Ed. Rio de Janeiro: LTC-Livros Técnicos e Científicos, 2021.
- LUGER, George F. **Inteligência Artificial** - 6ª ed. São Paulo: Pearson Education do Brasil, 2015.
- RUSSELL, Stuart; NORVIG, Peter. **Inteligência artificial: Uma Abordagem Moderna** - 4ª. Ed. GEN LTC, 2022.