Professor: Rafael Stubs Parpinelli

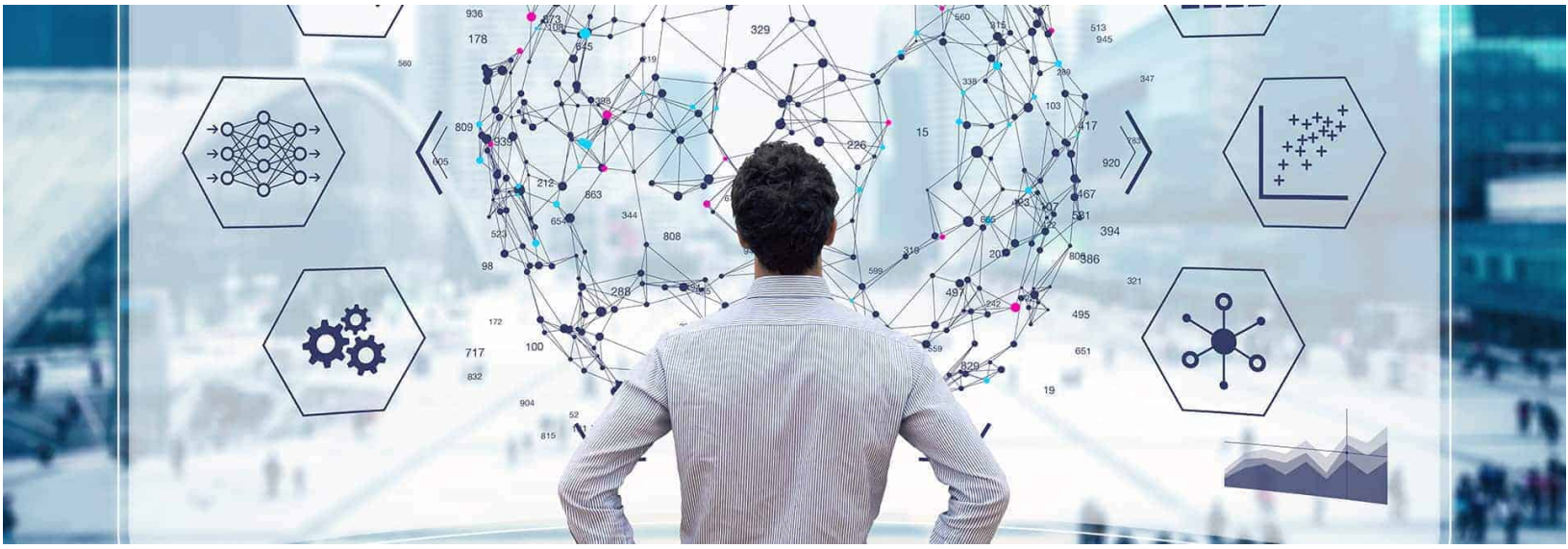
Estagiário Docente: Douglas Macedo Sgrott

Data de lançamento: 02/06/2021

Data prevista de término: 07/06/2021

Disciplina: Inteligencia Artifial

Trabalho 1 - Primeiros passos em Python + Análise Exploratória de Dados (EDA)



Objetivos:

- Iniciar a programação em Python, utilizando conforme necessário bibliotecas para manipulação numéricas (ex: Numpy e Pandas)
- Iniciara a explorar a área de Data Science utilizando conforme necessário bibliotecas para aprendizagem de máquina (ex: Scikit Learn)
- Iniciar a explorar a área de visualização de dados (exemplo de bibliotecas úteis: Matplotlib, Seaborn, Plotly, entre outros)
- Explorar o máximo possível sobre Análise Exploratória de Dados, aproveitando da vasta quantidade de exemplos disponíveis na plataforma Kaggle

Sobre o dataset:

- **Variáveis de entrada / independentes:** 'city', 'area', 'rooms', 'bathroom', 'parking spaces', 'floor', 'animal', 'furniture', 'hoa (R$)', 'rent amount (R$)', 'property tax (R$)', 'fire insurance (R$)'
- **Variável de saída / dependente:** 'total (R$)'

O que poderá ser avaliado no trabalho:

Análise de dados:

- Identificar os tipos das variáveis de entrada e saída do dataset
- Perguntas interessantes de serem respondidas: Qual a porcentagem de dados ausentes? Quanto é a assimetria (skewness) da distribuição dos dados? Qual a proporção das categorias dos dados qualitativos?
- Realizar Análise Univariada das variáveis (podendo incluir: cálculo da média, mediana, moda, cálculo da assimetria da distruibuição, limites dos intervalos, variância, gráfico de histogramas, gráfico de densidade da distribuição, entre outros…)
- Realizar Análise Bivariada das variáveis (gráfico de dispersão 2D ou 3D, gráfico de correlação / heatmap, contagem de frequências usando gráfico de barras ou pizza, entre outros...)

Data cleaning / Limpeza de dados:

- Quantificar a quantidade de dados ausentes, caracteres especiais e outliers.
- Analisar se irá deletar ou imputar variáveis ausentes
    - Se deletar, contar quantidade de dados ANTES e DEPOIS da remoção e qual o critério de deleção (1 dado ausente por linha, 2 dados ausentes por linha…)
    - Se não deletar, relatar o método usado para imputação (Substituição por constante, por média ou por regressão).
- Quantificar quantidade de dados inconsistentes, se houver, e corrigí-los
- Analisar a necessidade de retirar outliers para continuar análise
    - Se retirar, relatar o método utilizado para isso
    - Se não retirar, cuidado em deixar os gráficos legíveis

Data Scaling / Reescalonamento:

- Reescalonar os dados (recomenda-se utilizar as classes MinMaxScaler, StandardScaler ou RobustScaler da biblioteca Scikit Learn por simplicidade caso não queira criar código do zero para essa finalidade)

Feature Engineering / Engenharia de Atributos:

- Realizar algum método de Feature Engineering, como Decomposição, Binning, Cruzamento de variáveis, etc (ou seja, gerar uma nova variável de entrada a partir das variáveis de entrada já existentes. Importante: esta variável deve ser integrada no dataset junto com as outras variáveis de entrada já existentes).

Feature Selection / Seleção de Atributos:

- Aplicar algum método para realizar Feature Selection (Importante: caso o algoritmo retire alguma variável de entrada, elimine esta variável do dataset)

[EXTRA] Trazer novidades:

- Descobriu alguma análise/algoritmo não mencionado durante a aula e aplicou no dataset? Ótimo! Será muito bem vindo, contanto que esteja minimamente explicado o resultado da aplicação.

Algumas documentações técnicas dos algoritmos mencionados durante a aula:

Visualização de dados:

- Tutorial sobre Seaborn: https://seaborn.pydata.org/tutorial.html
- Diagrama de Caixas no Matplotlib e Seaborn
  https://matplotlib.org/3.1.1/api/_as_gen/matplotlib.pyplot.boxplot.html
  https://seaborn.pydata.org/generated/seaborn.boxplot.html
- Histograma no Matplotlib:
  https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.hist.html
- Gráfico de dispersão 2D no Matplotlib:
  https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.scatter.html
- Gráfico de barras no Matplotlib:
  https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.bar.html
- Gráfico de histogramas com 1 e 2 variáveis e gráfico de densidade no Seaborn
  https://seaborn.pydata.org/generated/seaborn.histplot.html#seaborn.histplot

- Gráfico de Correlação / Heatmap em Seaborn: https://seaborn.pydata.org/generated/seaborn.heatmap.html
- Tabela “resumo” estatístico em Pandas: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.describe.html
- Gráficos interativos (e que é possível executar no Kaggle notebook): https://plotly.com/python/

Data Cleaning:

- Preenchendo valores ausentes em Pandas (tem como preencher com constante e interpolação): https://pandas.pydata.org/pandas-docs/stable/user_guide/missing_data.html#na-values-in-groupby
- Preenchendo valores ausentes em Scikit Learn (preenchimento com constante ou regressor): https://scikit-learn.org/stable/modules/impute.html
- Filtragem de outliers (é StackOverflow, mas pode ajudar. Indicado para distribuições normais): https://stackoverflow.com/questions/23199796/detect-and-exclude-outliers-in-pandas-data-frame

Reescalonamento:

- Reescalonamento em Scikit Learn https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html#sphx-glr-auto-examples-preprocessing-plot-all-scaling-py

Feature Engineering:

- Aparentemente não existem bibliotecas de uso intuitivo para feature engineering. É mais recomendável neste trabalho aplicar manualmente, como por exemplo, criando uma nova feature através da multiplicação ou divisão de outras features, ou através de binning/discretização).

Feature Selection:

- Feature Selection em Scikit Learn (tem de diferentes tipos: abordagem de filtro e wrapper, incluindo Chi Test Squared, VarianceThreshold, Recursive Feature Elimination, Sequential Feature Selection, etc): https://scikit-learn.org/stable/modules/feature_selection.html