Bootcamp: Cientista de Dados

Trabalho Prático

Módulo 3: Técnicas para o processamento de Big Data

Objetivos de Ensino

Exercitar os seguintes conceitos trabalhados no Módulo:

1. Conhecimento do dataset
2. Limpeza dos dados
3. Identificação de outliers
4. Aplicação e análise de modelos de Modelo de Aprendizado de Máquina

Enunciado

Para o desenvolvimento de qualquer aplicação que envolva o desenvolvimento de algoritmos de aprendizado de máquina na ciência de dados, são necessárias 7 etapas básicas:

1. Coleta dos dados.
2. Preparação dos dados.
3. Seleção do modelo.
4. Treinamento do modelo.
5. Avaliação do modelo.
6. Sintonia dos parâmetros.
7. Previsão.

Dentre todas essas etapas, a que provavelmente demanda um maior esforço por parte do analista/cientista de dados é a etapa de preparação dos dados.

Isso ocorre porque é a partir dessa etapa que o analista/cientista de dados realiza a “limpeza” dos dados, identifica possíveis dados faltosos, possíveis outliers e prepara os dados para a construção dos modelos de previsão. Desse modo, realizar uma preparação correta dos dados ajuda a compreender o problema e obter resultados mais precisos com as previsões.

Atividades

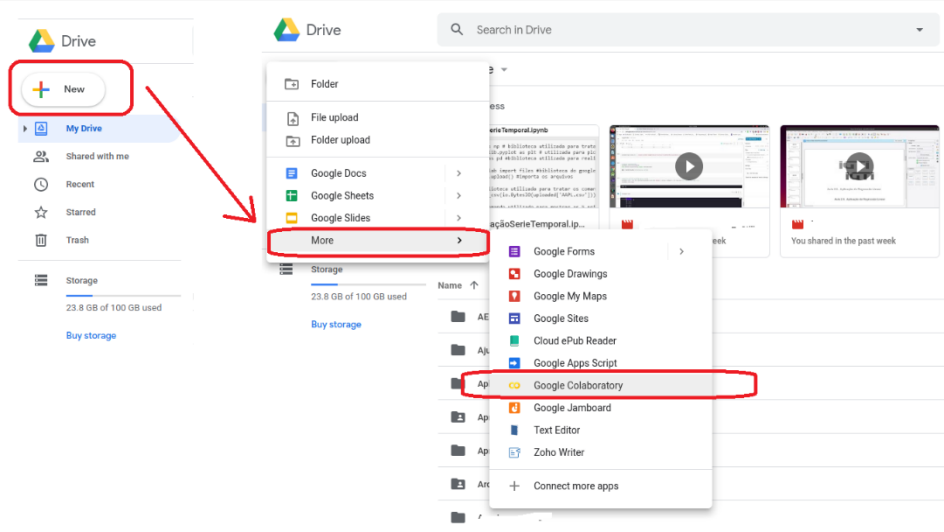
Os alunos deverão desempenhar as seguintes atividades:

Criar uma conta no Google.

Acessar o “Google Colaboratory”.

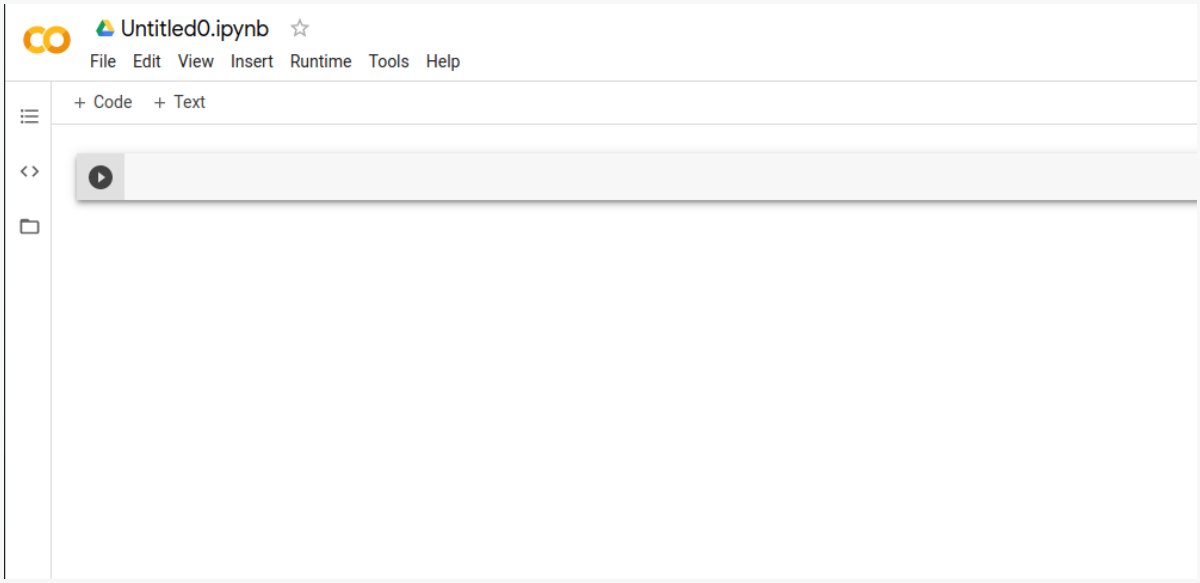
*Para este trabalho prático, será utilizado o ambiente de desenvolvimento do Google Colab. Para acessar esse ambiente, basta ter uma conta do Google ativa e acessar o Google Drive. Dentro do Google Drive, clique em “New”, depois em “More” e em seguida selecione “Google Colaboratory”. A Figura 1 mostra as etapas necessárias:

Figura 1 – Criando um arquivo no Google Colab



Após acessar o “Google Colaboratory”, você será direcionado(a) para o ambiente de desenvolvimento do Google. A Figura 2 apresenta a página que deve aparecer ao acessar o ambiente:

Figura 2 – Ambiente do Google Colab.



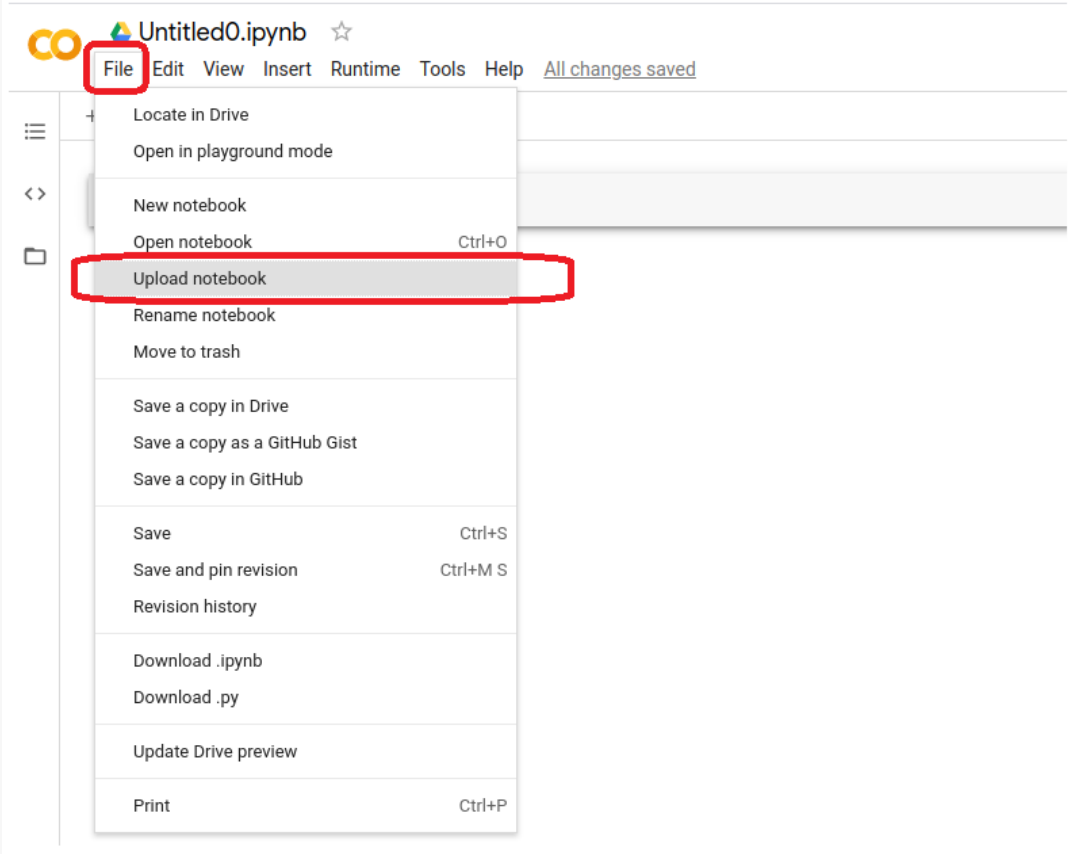
Para essa prática, será utilizado o dataset "wholesale_customers_data.csv". Para baixar esse dataset, acesse o link abaixo e realize o download dos arquivos "wholesale_customers_data.csv" e "trabalho_pratico_TPD_bootcamp.ipynb".

https://drive.google.com/drive/folders/10LIlGiWyYJltXVW36iCucoHG5rtcekkd?usp=sharing

Com todo o ambiente preparado, é necessário realizar o "upload" do arquivo "trabalho_pratico_TPD_bootcamp.ipynb" para o "Google Colab".

Para isso, acesse no canto superior esquerdo o menu "File" e clique em "Upload Notebook". No local onde realizou o download dos arquivos anteriores, selecione o arquivo "trabalho_pratico_TPD_bootcamp.ipynb". A Figura 3 demonstra como realizar esse procedimento. Após essa etapa, já é possível iniciar o seu Trabalho Prático.

Figura 3 – Upload do arquivo "trabalho_pratico_TPD_bootcamp.ipynb".



Para a realização do trabalho, é necessário executar, em sequência, cada uma das células presentes no “Google Colab”. Para executar uma célula, selecione a célula desejada e clique o ícone “play” (▶) ou pressione “Ctrl+Enter”.

Após executar a célula 2, será necessário realizar o upload do dataset utilizado para essa prática. Clique no botão “Escolher Arquivo” e selecione o arquivo “wholesale_customers_data.csv”. Prossiga executando cada uma das células.