04

Obtendo a base de dados

Transcrição

Nesta aula, geraremos nossos clusters. Primeiramente, deveremos fazer o download da nossa base de dados que utilizaremos ao longo do curso.

Utilizaremos a base de dados do **Kaggle**, uma plataforma de machine learning. A base que utilizaremos possui informações de aproximadamente nove mil clientes e como utilizam os cartões de crédito. Para realizar o download da base, precisaremos fazer um cadastro rápido.

Ao analisaremos o arquivo csv, teremos uma coluna que representa atributos diferentes: id do cliente, balanço (limite disponível na conta), frequência que o balanço é alterado, valor em compras a vista e parcelado, entre outros. São no total 18 atributos.

No Jupyter notebook, utilizaremos a ferramenta Anaconda (na descrição do curso há o link do curso de Pandas, em que existem instruções precisas de como instalar o Jupyter e o Anaconda).

Inseriremos o arquivo de baixamos dentro da ferramenta Anaconda, e então criaremos um novo notebook, que chamaremos de "Notebook Cartão de Crédito".

Primeiramente, geraremos a visualização dos dados por meio do Pandas, então realizaremos sua importação e então leremos o arquivo `pd.read.csv()` ,a função responsável pela leitura, e então passaremos o nome do arquivo. E então atribuiremos uma variável `dataframe` que representará nosso arquivo.

Escreveremos `dataframe.head()` , e então serão exibidas as cinco primeiras linhas do dataframe.

```
import pandas as  pd

dataframe = pd.read_csv("CC GENERAL.csv")
dataframe.head()
```

Teremos os atributos exibidos na tela, como o esperado. Vermos os diferentes comportamentos de compra entre os cinco primeiros clientes da lista, o primeiro não fez uma única compra parcelada nos últimos seis meses outros parcelaram compras com um determinado valor e assim por diante.

São muitos dados, por isso utilizaremos o algorítimo de clusterização para organizarmos as informações. Mas antes disso, precisamos eliminar os atributos que não são necessários para a clusterização, o que aprenderemos a seguir.