Aula 02: Coleta de Dados para Responder à Pergunta

Professor: Paulo Rogério Pires Manseira

Alunos: Maruan Biasi El Achkar e Ricardo Falcão Schilieper

GitHub: github.com/manseiracredit

Parte 1 Caça aos dados –Encontrando o dataset certo?

Critérios:

- **Relevância**

O dataset encontrado se encaixa muito bem no nosso tema e pergunta. Pois inclui diversos dados sobre compras com cartão de crédito, incluindo se aquela compra foi fraudulenta ou não.

- **Qualidade**

O dataset tem grande quantidade e diversidade de dados, além disso, não parece apresentar dados corrompidos nem faltando.

- **Tamanho**

8.000.008 dados individuais. Sendo 1.000.001 linhas e 8 colunas.

- **Formato**

CSV (72,7 MB.)

Fonte dos dados:

Fonte original: https://www.kaggle.com/datasets/dhanushnarayananr/credit-card-fraud
Nosso repositório: https://github.com/manseiracredit/dataset

Parte 2: Descrevendo sua Descoberta: O Dicionário de Dados

Nome do dataset: Credit Card Fraud

Fonte: https://www.kaggle.com/datasets/dhanushnarayananr/credit-card-fraud

Justificativa: O dataset escolhido possui mais de oito milhões de células de dados sobre compras feitas com cartão de crédito, incluindo um indicador se elas foram fraudulentas ou não. Todas as informações do nosso dataset serão utilizadas em nossas pesquisas para descobrir quais fatores indicam uma compra de cartão de crédito fraudulenta.

Dicionário de Dados Preliminar:

- **distance_from_home** ***(Float)***

Distância da compra atual em relação à residência do titular do cartão.

- **distance_from_last_transaction** ***(Float)***

Distância da compra atual em relação à compra anterior do cartão.

- **ratio_to_median_purchase_price** ***(Float)***

Razão do valor da compra atual em relação à média de compras do cartão.

- **repeat_retailer** ***(Boolean)***

Se o cartão já foi usado naquele estabelecimento anteriormente.

- **used_chip** ***(Boolean)***

Se o chip do cartão foi usado, ou seja, se o cartão foi inserido na máquina.

- **used_pin_number** ***(Boolean)***

Se a senha do cartão foi inserida durante a transação.

- **online_order** ***(Boolean)***

Se a transação foi feita online.

- **fraud** ***(Boolean)***

Se a transação foi considerada fraudulenta.

Desafios previstos:

Uma possível falta de relação entre os dados das compras fraudulentas.