Bootcamp IGTI

Desafio

Módulo 5	Arquiteto de Big Data

Objetivos

Para este desafio vamos exercitar todas as etapas do processo de Big Data. Para isso, vamos utilizar datasets distintos. O primeiro enunciado visa coletar, processar, analisar e aplicar um algoritmo de *Machine Learning* no *dataset* de clientes de uma rede de supermercados, cujo o principal objetivo é conhecer o perfil do cliente. No segundo enunciado vamos coletar, tratar, analisar e aplicar o algoritmo de regras de associação em um *dataset* de vendas de produtos. O objetivo dessa atividade é tentar encontrar padrões de compras de produtos realizadas pelos clientes.

Objetivos geral do desafio

Exercitar os seguintes conceitos trabalhados no curso:

- ✓ Coleta de dados estruturados.
- ✓ Coleta de dados na Web.
- ✓ Criação de estrutura de armazenamento em banco de dados.
- ✓ Tratamento, limpeza e processamento de dados.
- ✓ Analise de dados.
- ✓ Visualização de dados.
- ✓ Desenvolvimento de algoritmos de Machine Learning:
    a. K-means.
    b. Regras de associação.
- ✓ Práticas de manipulação de dados.
- ✓ Exercitar comandos Python, SQL.

Enunciado I

Uma rede de supermercados viu a necessidade de criar uma maneira de entender e conhecer mais o seu público-alvo. Diante desse desafio, a rede precisa criar um processo de Big Data para auxiliar essa análise. A rede quer conhecer seu cliente como um todo, das compras que foram realizadas aos produtos mais vendidos e, dessa forma, criar uma estrutura que permita tomar decisões mais assertivas.

Os analistas de Big Data da rede identificaram que é necessário desenvolver um processo bem elaborado para transformar dados variados em informações úteis. Para isso é necessário:

1. Coletar dados em diversas fontes;
2. Armazenar os dados em um repositório;
3. Realizar análises de dados coletados;
4. Criar modelo analítico de *Machine Learning*;
5. Criar visualizações para os dados processados.

Para esse primeiro momento, vamos analisar os dados e realizar um agrupamento dos clientes baseados em algumas características que eles possuem.

As compras foram separadas por usuário. Deste modo, cada compra necessita possuir cliente, produto, quantidade de produtos, valor unitário e valor total da compra.

ATENÇÃO

Informação importante: não existe compra sem produto ou sem cliente.

Os dados de clientes e compras é um dado fictício utilizado para o desenvolvimento das atividades a serem realizadas nesse trabalho. Deste modo, os dados foram criados de forma aleatória e não possuem nenhuma relação com dados no mundo real.

Atividades do enunciado I

Os alunos deverão desempenhar as seguintes atividades:

1. Coletar dados das seguintes fontes de dados.
    - compras.xls
        - Contém dados das compras realizadas por cliente;
    - clientes.json
        - Contém dados de clientes (análise de perfil);
    - estados.txt
        - Contém dados de estados dos clientes;
    - O link: https://profleandrolessa.wordpress.com/exercicio-de-coleta-de-dados/
        - Contém dados de produtos.
2. Criar estrutura de armazenamento;
3. Avaliar dados ausentes das colunas e corrigi-los;
4. Criar algoritmo de clusterização k-means;
5. Responder as questões 1 a 10 práticas do desafio.

Informações de identificação de códigos base de clientes:

código	estado civil
0	solteiro(a)
1	casado(a)
2	viuvo(a)
3	divorciado(a)

código	indicador
0	não hipertenso
1	hipertenso
0	não diabetes
1	diabetes

código	sexo
0	feminino
1	masculino

Enunciado II

Após a implantação de todas as etapas do processo do Big Data, os analistas identificaram a necessidade de entender melhor a relação entre os produtos comprados pelos clientes. A ideia é encontrar padrões ocultos nos dados, que possam auxiliar na tomada de decisão e assim criar promoções, kits de vendas e melhorar disponibilização de produtos nas prateleiras dos supermercados, por exemplo. Para isso, antes de implementar o modelo em produção, os analistas vão realizar uma POC com dados de vendas de produtos de outra rede de supermercados e, em seguida, aplicar o modelo nos dados de produção do supermercado.

Atividades do enunciado II

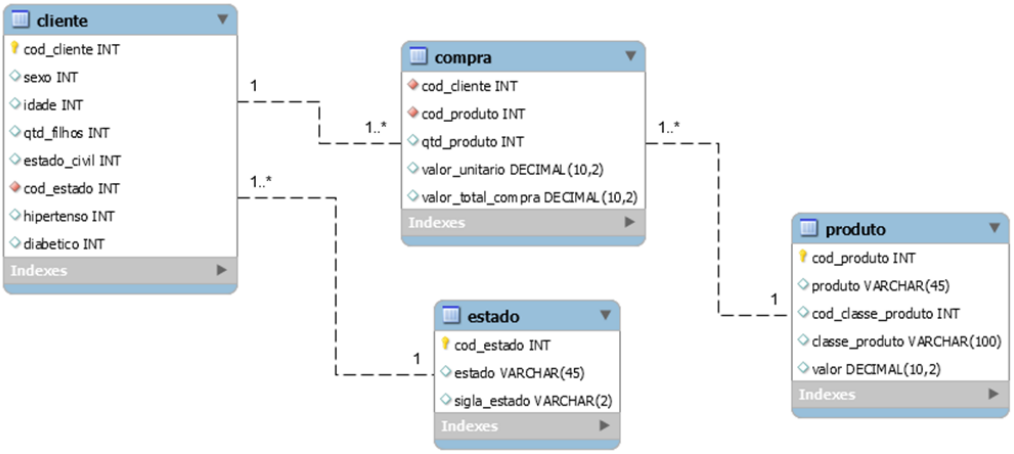
Os alunos deverão desempenhar as seguintes atividades:

1. Coletar dados do *dataset* mercado.csv;
2. Analisar os dados coletados;
3. Tratar os dados coletados;
4. Avaliar dados ausentes nas colunas;
5. Identificar os itens frequentes;
6. Criar regras de associação;
7. Responder as questões 11 a 15 práticas do desafio;

Dicas do professor:

- Para armazenar os dados, pode utilizar banco de dados relacional MYSQL ou banco de dados não relacional Mongodb.
- Para as atividades relacionados a clusterização realizada pelo k-means, crie um novo dataframe (compras_clientes) com os dados obtidos entre a relação das tabelas cliente, estado, produto e compras.
- Muita atenção para realizar inserts nas tabelas para o banco de dados relacional. Verifique a hierarquia dos dados nas tabelas.

- Segue uma sugestão para modelagem de dados.



- Ao coletar e armazenar os dados de compras no formato xls, pode ser que seja necessário instalar um complemento. Caso isso aconteça utilize: !pip install xlrd
- Muita atenção para as informações dos códigos 0 e 1 da tabela de clientes.
- Os datasets estão disponíveis no link:
    - https://github.com/ProfLeandroLessa/desafio-final-ABD

Respostas Finais

Os alunos deverão desenvolver a prática e, depois, responder às seguintes questões objetivas: