MVP ENGENHARIA DE DADOS – CAIO AUGUSTO RODRIGUES ALVES VIEIRA

Tema: Análise de Desempenho Estudantil

1. Objetivo do Projeto

O objetivo deste MVP é construir um pipeline de dados utilizando as camadas Bronze, Prata e Ouro no Databricks, a partir do dataset público 'Students Performance', obtido pelo Kagle.. Aqui buscaremos entender como as variáveis socioeducacionais influenciam as notas dos estudantes em matemática, leitura e escrita. A ideia é responder às seguintes perguntas:

a) Existe diferença de desempenho entre gêneros?;
b) A escolaridade dos pais influencia nas notas?
c) O curso preparatório melhora o desempenho?
d) Existe alguma diferença de performance entre etnias?
e) Há correlação entre as notas de matemática, leitura e escrita?

2. Coleta dos Dados

A coleta foi realizada por upload do arquivo no Databricks. Por ser um dataset público, não houve necessidade de web scraping ou coleta automatizada.

3. Modelagem dos Dados

A modelagem adotada segue o padrão de Data Lake com três camadas:

- Bronze: dados brutos, qualquer alteração da base extraída;
- Prata: dados tratados, padronizados e preparados para análise;
- Ouro: tabelas analíticas e métricas consolidadas.

Para a camada Ouro, foi construído um modelo dimensional simples (Esquema Estrela), contendo:

- Fato:
    - Fato_Performance (nota de matemática, leitura, escrita, ID do aluno)
- Dimensões:
    - Dim_Genero;
    - Dim_Etnia;
    - Dim_Educacao_Pais;
    - Dim_Curso_Preparatorio

Os dados utilizados no projeto representam informações acadêmicas de estudantes, incluindo notas e características socioeducacionais. Segue abaixo uma descrição de cada coluna, tipo, domínio e valores esperados (quando aplicáveis):

a) Genero:
   i) Tipo: string
   ii) Valores possíveis: "male", "female"
b) Etnia:
   i) Tipo: string
   ii) Valores possíveis: "group A", "group B", "group C", "group D", "group E"
c) Educacao_dos_pais:
   i) Tipo: string
   ii) Valores possíveis:
      1) "some high school";
      2) "high school";
      3) "some college";
      4) "associate's degree";
      5) "bachelor's degree";
      6) "master's degree".
d) Almoço:
   i) Tipo: string
   ii) Valores possíveis:
      1) "Standard";
      2) "free/reduced"
e) Curso_Preparatorio:
   i) Tipo: string
   ii) Valores possíveis:
      1) "None"
      2) "Completed"
f) Nota_Mat:
   i) Tipo: inteiro
   ii) Valores esperados: 0 a 100
g) Nota_Leitura:
   i) Tipo: inteiro
   ii) Valores esperados: 0 a 100
h) Nota_Escrita:
   i) Tipo: inteiro
   ii) Valores esperados: 0 a 100

Linhagem:

- Fonte: dataset público Students Performance

- Transformações: renomeação de colunas, criação de ID artificial, cast de tipos, agregações.

- Destino final: tabelas Delta nas camadas Prata e Ouro

4. Carga dos Dados (ETL no Databricks)

A carga foi realizada por meio de um pipeline simples utilizando notebooks em PySpark.

Etapas principais:

4.1) Bronze:

- Leitura do CSV;
- Inferência de schema;
- Armazenamento como tabela Delta

4.2) Prata:

- Padronização dos nomes de colunas;
- Criação da coluna ID_Estudante;
- Garantia de tipos numéricos corretos;
- Escrita em formato Delta

4.3) Ouro:

- Agregações por gênero, etnia e escolaridade;
- Cálculo de médias e estatísticas;
- Construção das tabelas para análise.

Todo o processo foi registrado nos notebooks correspondentes.

5. Análise

a) Qualidade dos Dados

Em relação a qualidade dos dados notamos que não possuímos valores nulos no dataset, todas as notas variam de 0 a 100 e não encontramos valores duplicados. Em resumo a qualidade dos dados eram boas, e não foi necessário um trabalho pesado neste sentido, e assim não tivemos problemas para realizar a análise

b) *Solução do Problema*

1) Existe diferença de desempenho entre gêneros?

   Notamos que em leitura e escrita, as mulheres possuem um desempenho melhor. Já em Matemática, os homens possuem melhor peformance.

2) A escolaridade dos pais influencia nas notas?

   Notamos que quanto maior a escolaridade dos pais, maior tendem a ser as notas dos filhos.

3) O curso preparatório melhora o desempenho?

   Os alunos que fizeram curso preparatório apresentam notas melhores.

4) Existe alguma diferença de performance entre etnias?

   Em relação à etnia, o grupo E apresentou o melhor desempenho geral, enquanto o grupo A foi o que obteve as menores médias.

5) Há correlação entre as notas de matemática, leitura e escrita?

   Sim. A maior correlação foi encontrada leitura e escrita (0,95) e a pior entre escrita e Matemática (0,80).

6. Conclusão

A construção do pipeline permitiu organizar o conjunto de dados em três diferentes camadas: bronze, prata e ouro, garantindo, assim, rastreabilidade, padronização e qualidade para realizarmos todas as análises propostas. Além disso, o processo reforçou a importância de boas práticas de engenharia de dados., já que todo o trabalho feito no dataset resultou em informações suficientes para responder às perguntas levantadas neste estudo. Do ponto de vista técnico, o MVP cumpriu sua proposta, já que criamos um pipeline funcional e documentado, capaz de transformar dados brutos em informações úteis.

7. Autoavaliação

Ao longo do desenvolvimento deste MVP, consegui atingir os principais objetivos propostos: construir um pipeline funcional, aplicar boas práticas de organização de dados e produzir análises que realmente respondem às perguntas de negócio. A estrutura Bronze → Prata → Ouro ajudou a visualizar com clareza cada etapa do processo e reforçou a importância de transformar um dataset bruto em algo analítico.

Durante a execução, percebi que algumas tarefas exigiram mais atenção do que eu imaginava, principalmente no uso do Databricks e na padronização das tabelas.

Para trabalhos futuros, vejo oportunidades de melhorar o pipeline com automação, validações de qualidade mais robustas e criação de dashboards. Também seria interessante testar modelos preditivos ou explorar análises mais profundas usando bibliotecas avançadas.