### ****MVP – Engenharia de Dados****

**Marcos Fernando de Moraes Siliprandi**

**Descrição**

Este trabalho tem como objetivo a construção de um pipeline de dados utilizando tecnologias na nuvem, com foco na plataforma Databricks Community Edition. O pipeline abrangerá todo o ciclo de vida dos dados, incluindo busca, coleta, modelagem, carga e análise. A base de dados utilizada contém informações sobre partidas do Campeonato Brasileiro, estatísticas dos clubes, gols e cartões, permitindo a extração de insights relevantes sobre o desempenho dos times e jogadores ao longo da competição.

### ****Objetivo****

O objetivo deste MVP é estruturar e transformar dados brutos do Campeonato Brasileiro em um formato otimizado para análise, possibilitando a extração de insights que possam responder a perguntas como:

* Quais times apresentam melhor desempenho ao longo das rodadas com base em estatísticas como posse de bola, chutes e passes?
* Há relação entre a quantidade de cartões recebidos e o resultado final da partida?
* Quais jogadores se destacam na artilharia e em quais momentos da partida ocorrem mais gols?
* Como o fator "mandante" influencia nos resultados das partidas?
* Existe uma correlação entre estatísticas de jogo e a vitória de um time?

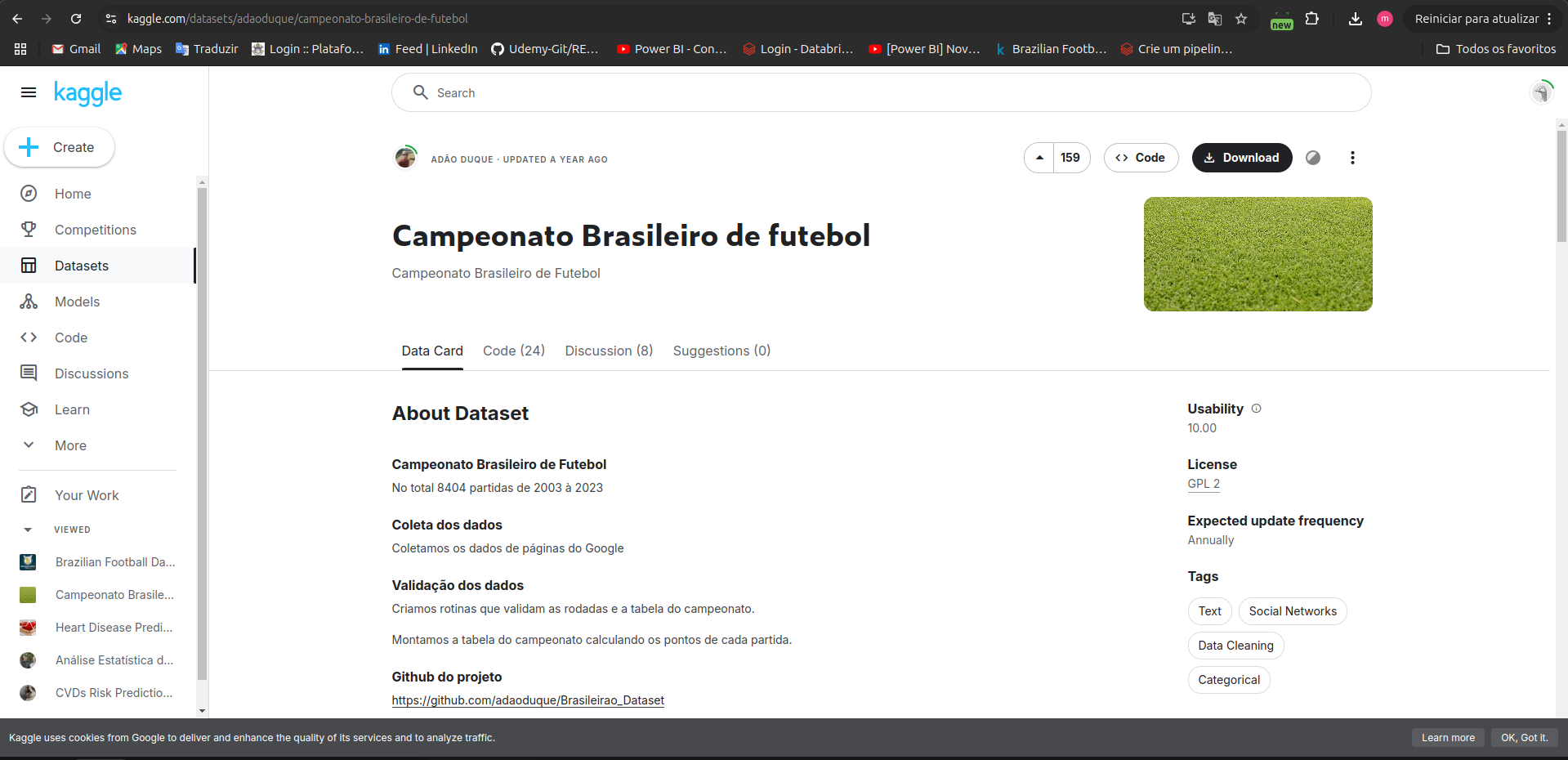
A partir dessa estruturação, será possível garantir a qualidade, consistência e acessibilidade dos dados para futuras análises, facilitando a tomada de decisão baseada em dados.

**Detalhamento**

**1 - Busca pelos dados**

**Para atender aos objetivos deste MVP, optei por utilizar a base de dados "Campeonato Brasileiro de Futebol", disponível gratuitamente no Kaggle. Essa base contém informações detalhadas sobre partidas, estatísticas dos clubes, gols e cartões, permitindo uma análise abrangente do desempenho das equipes ao longo do campeonato.**

A base pode ser acessada através do link:  
[Campeonato Brasileiro de Futebol - Kaggle](https://www.kaggle.com/datasets/adaoduque/campeonato-brasileiro-de-futebol)



**2 – Coleta**

**Esta etapa será dividida entre as etapas Criação de Cluster, Criação de Notebook, Upload dos dados, Leitura dos arquivos e Armazenamento inicial**