Prévia Projeto Final - MC536

Grupo DDD (FUT)
Gustavo Mantellato Elias - 169366
Leandro Garcia Pereira - 178258

Tema do dataset

O Futebol é o esporte mais popular do mundo e sem dúvida representa parte importante da cultura brasileira.

Por esse motivo decidimos tentar consolidar e tratar registros de todas as partidas ocorridas em Campeonatos Brasileiros desde 1959 no nosso dataset. Esse conteúdo não é encontrado de forma centralizada em um mesmo banco de dados, nem mesmo por parte da Confederação Brasileira de Futebol (CBF) por motivos que descreveremos a seguir.

Relevância do Projeto

O que faz desse projeto relevante?

O Campeonato Brasileiro de Futebol é a principal competição de futebol entre clubes no país. Porém o modelo de disputa atual, entre 20 equipes e com pontos corridos não é nem de perto o primeiro formato da competição.

Além do formato, o Campeonato Brasileiro teve diversos nomes desde que passou a ser disputado e recentemente a Confederação Brasileira de Futebol decidiu adicionar mais alguns a lista

Relevância do Projeto

Em 2010 a Confederação Brasileira de Futebol (CBF) decidiu unificar títulos de alguns torneios disputados anteriormente à 1971 (até então o marco de início do Campeonato Brasileiro) e passou também a considerar também os seguintes campeonatos:

Taça Brasil

Torneio Roberto Gomes Pedrosa

Taça de Prata

Relevância do Projeto

Apesar da decisão pela unificação ter sido tomada, a CBF não demonstrou muito esforço em agregar e disponibilizar os dados de forma consistente dos referidos torneios recentemente adicionados.

Com isso, a nossa ideia é ter todas as partidas realizadas no Campeonato Brasileiro (1959-2021) no nosso dataset, unindo informações encontradas em diferentes repositórios e bancos de dados pela internet.

Métodos de extração

Os dados das partidas estão sendo reunidos de diferentes maneiras. Algumas informações foram encontradas já em forma de tabelas em repositórios, enquanto a grande maioria dos dados será coletado por meio de Web Scraping com auxílio da linguagem Python para extrair os dados de maneira automatizada.

Fontes de dados

título da base	link	breve descrição
Ogol	www.ogol.com.br	Um site com infomações de competições de futebol nacional e internacional. Mais importante, com o histórico de todos os torneios nacionais desde 1959. A técnica utilizada para extração de dados será o web scrapping com auxílio de bibliotecas do python.
Brasileirão Dataset	github.com/adaoduque/Brasileirao_Dataset	Dataset aberto independente com as partidas do campeonato brasileiro no período de pontos corridos. Os dados precisam passar por tratamento para se adequar ao modelo proposto pelo grupo, que acredita ser mais adequado pois simplifica as tabelas ao mesmo tempo que acomoda mais informações.
Wikipedia	wikipedia.com	Informações dos clubes e suas participações nos torneios nacionais serão extraídas aqui utilizando web scrapping e bibliotecas do python.

Possíveis análises

Ter uma base de dados com todas as partidas em Campeonatos Brasileiros além de algumas informações como Estado do clube, número de títulos, número de participações entre outras permite inúmeras possibilidades de análises e agregação de dados, como por exemplo:

Qual região do país tem o maior número de títulos de Campeonato Brasileiro antes e depois da unificação?

Para essa análise precisamos do número de conquistas de cada estado além de realizar uma transformação da nossa rede para definir a composição de uma região e assim, comparar os resultados.

Possíveis análises

Considerando a edição do campeonato deste ano - ainda em andamento - e que determinado time possui X pontos em uma determinada rodada, qual a probabilidade desse time ser campeão do torneio?

Para isso devemos fazer uma análise de predição, usando como base times de edições anteriores (considerando o mesmo formato de torneio) que possuíam a mesma, maior e menor quantidade de pontos na mesma rodada do torneio e a colocação que esses times terminaram a competição.