Projeto Final de Aprendizado Descritivo

Luís Felipe Ramos Ferreira
Igor Lacerda iFaria da Silva
Matheus Tiago Pimenta de Souza

1 Introdução

O uso de ciência de dados e estatística para analisar esportes é algo que vem crescendo cada vez mais nos últimos anos. Em particular, o futebol têm sido um desses esportes [1]. A própria UFMG ofertou no ano passado e novamente neste semestre a disciplina 'Ciência de Dados aplicada ao futebol', o que mostra a relevância do tema. Diversas empresas que atuam na área surgem a cada dia, e os times de futebol, no Brasil e no resto do mundo, estão investimento em seus departamentos de dados e estatística.

Nesse sentido, nosso grupo optou por estudar e compreender melhor como funciona o uso de análises estatísticas no futebol, dado o interesse geral pelo esporte, e, para isso, nos propusemos a aplicar algoritmos de mineração de dados em dados futebolísticos, sendo eles dados de súmula, dados de eventos ou até mesmo dados de *tracking* dos jogadores, para compreender como as informações acerca do jogo estão contidas dentro dos dados coletados e como isso pode ser utilizado a favor das equipes.

Os dados de eventos, especialmente, costumam ser mais fáceis de lidar e mais fáceis de acessar do que dados de *tracking*, enquanto trazem muito mais informações do que dados de súmula. Existem, atualmente, algumas bases gratuitas de dados de evento de partidas, disponibilizadas por diferentes empresas como *Wyscout* e *StasBomb*. Como a ideia é ter um panorama geral de diversas partidas, campeonatos e jogadores, iremos utilizar as bases de dados disponibilizadas sobre as 5 grandes ligas de futebol europeu das temporadas 17/18 da empresa *Wyscout*.

1.1 Base de dados

As bases de dados utilizadas na ferramenta consistirá na base principal disponibilizada pela empresa *Wyscout*, consistindo em uma base de dados de evento das 5 grandes ligas europeias na temporada 17/18.

Cada empresa fornecedora de dados possuem seu próprio formato de representação dos dados de evento. De modo a facilitar a mesclagem entre as bases de dados utilizadas, iremos converter os dados coletados para uma representação geral proposta por pesquisadores denominada SPADL. A SPADL é uma boa escolha por ser uma representação concisa e fácil de utilizar. Ela é uma representação tabular de cada evento da partida, onde cada linha possui 12 colunas. A tabela abaixo ilustra o esquema de representação de um evento segundo o formato SPADL.

Atributo	Descrição
game_id	O ID do jogo no qual a ação foi realizada
period_id	O ID do período do jogo no qual a ação foi realizada
seconds	O tempo de início da ação
player	O jogador que realizou a ação
team	O time do jogador
start_x	A localização x onde a ação começou
start_y	A localização y onde a ação começou
end_x	A localização x onde a ação terminou
end_y	A localização y onde a ação terminou
action_type	O tipo de ação (por exemplo, passe, chute, drible)
result	O resultado da ação (por exemplo, sucesso ou falha)
bodypart	A parte do corpo do jogador usada para a ação

Table 1: Descrição dos dados no formato SPADL

2 Implementação

A linguagem escolhida para o desenvolvimento do trabalho foi `Python` (versão 3.10.12), devida a seu vasto ecossistema para ciência de dados e mineração de dados.

A manipulação dos dados foi feita com o uso de bibliotecas de análise numérica como `NumPy` e manipulação de *dataframes* como `Polars` e `Pandas`, uma vez que se tratam de ferramentas extremamente completas que facilitaram o desenvolvimento do projeto como um todo.

Para aplicar os algoritmos de descobertas de subgrupos, foi utilizado o pacote `pysubgroup`, que fornece uma aglomeração de algoritmos do estado da arte de descoberta de subgrupos em um formato simples e leve para serem utilizados.

Para organizar o ambiente de desenvolvimento, que englobava vários pacotes diferentes, foi utilizado o gerenciador de pacotes `Anaconda`, o que facilitou o trabalho com os pacotes de ciência de dados citados. O projeto final foi salvo em um `repositório` no GitHub para fácil versionamento e organização de código. As instruções de como utilizar o que foi implementado estão descritas no arquivo *README.md* do repositório.

3 Resultados

gfdfd

4 Conclusão

fefre

Referências Bibliográficas

[1] Aristotelis Takvorian. *The Beautiful (Computer) Game: How Data Science Will Revolutionize the World's Most Popular Sport*. PhD thesis, 2021.