Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Projeto Final de Aprendizado Descritivo

Mineração de Dados de Eventos em Futebol

Alunos: Luís Felipe Ramos Ferreira, Igor Lacerda Faria da Silva, Matheus Tiago Pimenta de Souza

Professor: Renato Vimieiro

Belo Horizonte - Minas Gerais
2024

Sumário

Mineração de Dados de Evento em Futebol

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

July 30, 2024

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 tem sido um desses esportes [12]. Diversas empresas que atuam na área surgem a cada dia, e os times de futebol, no Brasil e no resto do mundo, estão investindo em seus departamentos de dados e estatística.

Nesse sentido, este trabalho propõe estudar e compreender melhor como funciona o uso de análises estatísticas no futebol, dado o interesse do público geral pelo esporte, e, para isso, propomos a aplicação de algoritmos de mineração de dados em dados futebolísticos, sendo eles dados de súmula e dados de eventos das partidas, 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*.

1.1 Base de dados

A principal base disponibilizada pela empresa *Wyscout* foi usada como base de dados. Ela contém dados de evento das 5 grandes ligas europeias na temporada 17/18.

Cada empresa fornecedora de dados possui 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 os pacote pysubgroup, que fornecem uma aglomeração de algoritmos do estado da arte de descoberta de subgrupos em um formato simples e leve para serem utilizados. Também foi utilizada a implementação do algoritmo SSD++ para valores numéricos, um dos métodos mais recentes utilizado nos artigos [**?**] e [**?**].

falar dos pacotes de mineração de seq aqui

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 *README* do repositório.

3 Referencial Teórico

No decorrer do trabalho, duas importantes métricas de análise ofensiva no futebol serão utilizadas. Esta seção aglomera os conhecimentos necessários sobre elas para a compreensão do projeto e o como elas foram utilizadas.

3.1 Gols Esperados (xG)

Intuitivamente, existe a noção de que, quanto mais próximo um jogador está do gol, mais chance ele tem de conseguir marcar um ponto para sua equipe. Uma noção similar existe para o ângulo entre o gol e jogador: é mais difícil um atacante acertar se ele está em uma das laterais. Na área de *analytics* de futebol, essa noção é formalizada através da métrica de *expected goals*, ou xG. A ideia central é construir um modelo do que um jogador médio faria em dado estado de jogo, que, além de incluir os fatores mencionados anteriormente, pode ser mais (ou menos) extensivo.

A métrica de Gols Esperados captura um estado de jogo e retorna uma probabilidade estimada de um jogador marcar. Ela pode ser vista como um *framework*, cujos detalhes de implementação são decididos pelo usuário. Outros fatores considerados importantes são: parte do corpo associada à ação (pé dominante ou não; de cabeça), origem da assistência (cruzamento, passe) e a posição dos defensores. Para avaliar este último quesito, é necessário ter dados de *tracking*, o que limita consideravelmente sua adoção. Além disso, mesmo questões como a liga em análise podem ser relevantes.

O uso do xG é mais amplo do que se pode imaginar a princípio. Em um primeiro momento, partidas podem ser analisadas: é possível avaliar se um time fez mais "pressão" pela soma do seu xG na partida. No entanto, como com qualquer outro indicador, o xG não define absolutamente o resultado de um jogo, podendo um time aproveitar muito bem situações menos favoráveis. Outra aplicação é avaliação de jogadores: se um jogador consistentemente faz gols em situações desfavoráveis, isso é um grande indício de que ele é jogador de destaque, ou age muito bem sob pressão. O xG também pode ser usado taticamente, para proporcionar melhorias no posicionamento dos jogadores da *defesa* e para embasar estratégias.

É comum usar modelos de regressão logística para aplicar o xG, em que se considera uma combinação dos fatores mencionados anteriormente. Essa combinação não é necessariamente linear: considerar o quadrado da distância pode ser valioso, por exemplo. Na prática, no entanto, qualquer modelo de classificação binária pode ser usado. Um exemplo muito prevalente é o xG do *Statsbomb*, que usa um *XGBoost* e considera a quantidade e a posição dos jogadores entre o gol e o jogador que faz o lance.

3.2 Valuing Actions by Estimating Probabilities (VAEP)

O xG deixa a desejar em alguns aspectos. Principalmente pelo fato de focar exclusivamente nas chances de se fazer gols, sem considerar outros aspectos importantes do jogo. Por exemplo, existem passes que são cruciais para uma equipe se aproximar do gol. É justamente para suprir a necessidade de avaliar as ações dos jogadores de forma mais ampla que foi desenvolvido o VAEP [5]. O VAEP avalia cada ação do jogo em termos do quanto ela aumenta (ou diminui) a chance de um time fazer (ou sofrer) um gol.

Com base no estado atual do jogo (placar, tempo restante, ações anteriores, posição da bola, etc), o VAEP estima a chance de um time fazer ou sofrer um gol, antes e após cada ação. Desse modo, é possível estimar a utilidade de uma ação com base na diferença entre as mudanças de se fazer e de se sofrer um gol. Formalizando, temos que, para um dado time t, uma ação a_i e um estado de jogo S_i:

$$\Delta P_{score}(a_i, t) = P^k_{score}(S_i, t) - P^k_{score}(S_{i-1}, t) \tag{1}$$

$$\Delta P_{concede}(a_i, t) = P^k_{concede}(S_i, t) - P^k_{concede}(S_{i-1}, t) \tag{2}$$

$$V_{\text{VAEP}}(a_i) = \Delta P_{score}(a_i, t) - \Delta P_{concede}(a_i, t) \tag{3}$$

Nas equações, k é um parâmetro que indica quantas jogadas futuras estão sendo consideradas para se avaliar a jogada atual. Assim, quando uma jogada aumenta muito a chance de um gol acontecer, seu VAEP é um número positivo e, caso a jogada ofereça mais risco do que recompensa, o VAEP será um valor negativo.

Então, para se calcular o VAEP, é necessário estimar as probabilidades de marcar ou sofrer um gol. Essa tarefa pode ser resolvida com Aprendizado de Máquina, com a criação de dois modelos: um para estimar a probabilidade da equipe com a posse da bola fazer um gol até as k ações após o estado atual S_i e outro para estimar a probabilidade de a equipe sofrer o gol no mesmo período. Um valor comummente utilizado para k é 3.

Tal como o xG, o VAEP pode ser utilizado para se avaliar a performance dos jogadores, ao se agregar as ações. Inclusive, os autores do *paper* usaram disso para validar que o VAEP é uma métrica consistente. Considerando os jogadores que participaram de jogos das ligas europeias nas temporadas de 17/18 (*coincidentemente* a mesma base de dados que foi usada nesse trabalho) por pelo menos 900 minutos, foram construídos gráficos mostrando o número de ações por 90 minutos pelo valor médio das ações. Liderando o *ranking*, nomes conhecidos como Messi, Bale e Cristiano Ronaldo aparecem, o que indica que o VAEP é condizente com outras avaliações de desempenho.

4 Análise Exploratória

A base de dados de eventos possui muitas informações interessantes que podem ser exploradas antes mesmo da aplicação de algoritmo de mineração de dados. Nesta seção, discutimos alguns *insights* interessantes observados na base das 5 grandes ligas europeias da temporada 17/18.

4.1 Distribuição de ações

A base de dados possui uma distribuição não uniforme de ações. Como pode-se analisar no histograma abaixo, os eventos de passes são extremamente mais frequentes do que qualquer outro. Isso está dentro do esperado, dado que o passe é o principal fundamento do futebol, mas deve ser levado em consideração quando modelos de mineração forem aplicados à base de dados.

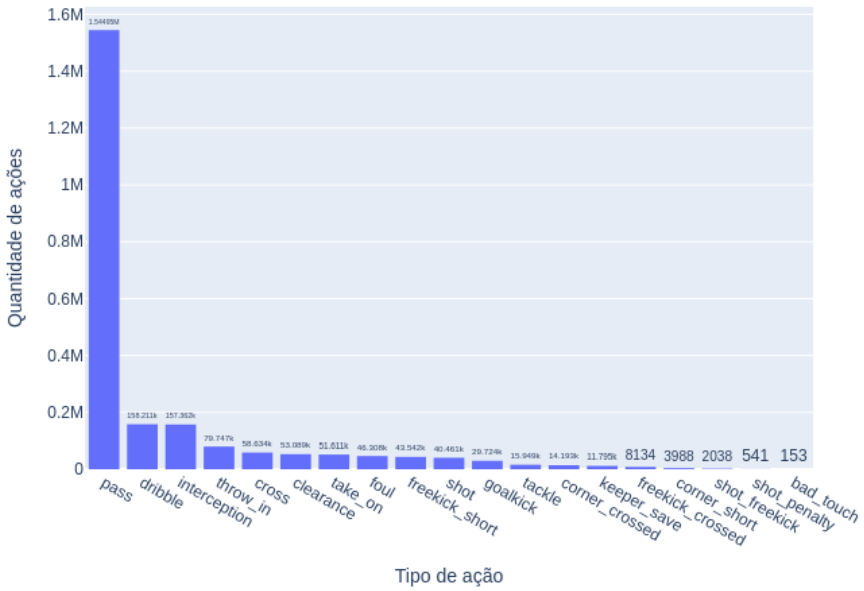

Figure 1: Distribuições dos tipos de ação

4.2 Distribuição da posição de chutes convertidos em gol

O mapa de calor abaixo permite que sejam analisadas a distribuição das posições dos chutes que se converteram em gol na base de dados.

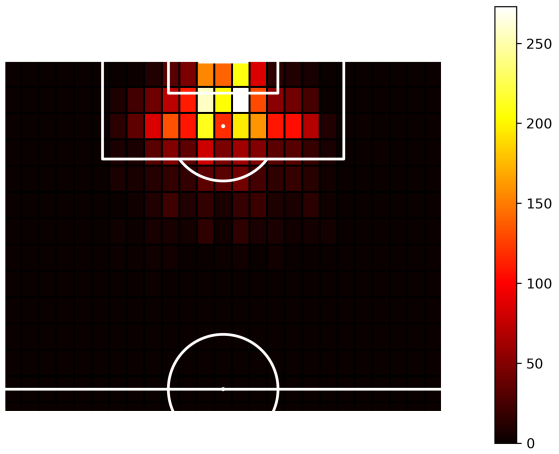

Figure 2: Distribuições dos tipos de ação

4.3 Distância média entre passes

O passe é um dos, se não o fundamento mais importante em um esporte coletivo como o futebol. Uma boa execução desse fundamento por parte dos jogadores de uma equipe indica um bom controle da posse de bola que está diretamente conectado com bons resultados [4]. Nesse sentido é interessante fazer um análise das distâncias entre os passes executados por cada equipe.

Em particular, foi coletada a distância entre todos os passes bem sucedidos de cada equipe e, posteriormente, coletadas estatísticas sobre essas distâncias. A distância média entre passes se destacou, uma vez que ela indica como funciona a dinâmica de troca de passes de uma equipe. Os resultados observados confirmaram suspeitas prévias acerca do assunto. Como pode-se notar nas

tabelas abaixo, para cada uma das grandes ligas, as principais equipes, isto é, as equipes com maior grandeza histórica e maior poderio financeiro figuram como aquelas que possuem a menor distância média entre passes bem sucedidos. Na Inglaterra, por exemplo, o Manchester City, equipe comandada pelo espanhol Pep Guardiola, foi a equipe com a menor média de distância entre passes. O estilo de jogo de Guardiola é muito focado no controle da posse de bola e na movimentação dos jogadores para receptar um passe [13], então o resultado está dentro do esperado.

É interessantes notar que os clubes com menor média de distância entre os passes da temporada 17/18 em cada liga mantiveram posições altas em seus respectivos campeonatos. O Manchester City e o Paris Saint Germain se sagraram campeões, enquanto Napoli e Atletico de Madrid ficaram com a segunda colocação. Na Alemanha, o RB Leipzig ficou com a sexta colocação. Pode-se notar então que existe uma correlação entre a o desempenho de um time no campeonato com a distância média entre os passes da equipe.

Equipe	**Distância média entre passes (m)**
Manchester City FC	17.208
Arsenal FC	17.729
Manchester United FC	17.823
AFC Bournemouth	18.417
Tottenham Hotspur FC	18.490
Crystal Palace FC	18.555
Chelsea FC	18.682
Southampton FC	18.801
Liverpool FC	18.808
West Ham United FC	18.832
Watford FC	18.967
Newcastle United FC	18.972
Swansea City AFC	19.077
Leicester City FC	19.189
Huddersfield Town FC	19.215
Stoke City FC	19.690
West Bromwich Albion FC	19.712
Everton FC	19.785
Brighton & Hove Albion FC	19.838
Burnley FC	20.634

Table 2: Premier League

5 Extração de *Features*

O foco das análises desse trabalho foram em avaliar eventos que levaram à chutes ao gol, na tentativa de tentar entender quais atributos que levam (ou não) a pontuar no futebol. Dessa forma, o primeiro pré-processamento foi filtrar os eventos de todas as partidas, para capturar as informações só do evento de chute, além de capturar outras informações.

Mais especialmente, a partir dos atributos listados na 1, realizamos um pré-processamento para computar métricas adicionais relevantes para a nossa análise. Alguns dos valores por sí só já são relevantes e portanto foram mantidos, como o nome do jogador, as localizações de onde a ação começou (nesse caso, o chute) e qual parte do corpo foi usada. Outras métricas foram calculadas, como distância percorrida da bola partindo do chute até o gol, e o ângulo do chute relativo ao gol. Já outros atributos foram computadas com base nos outros eventos que levaram ao chute: consideramos para a parte de SD desse trabalho todos os eventos desde que o time que chutou tem a posse de bola, por acreditar que tudo possa fazer parte da estratégia para chegar ao gol (e por ser muito difícil definir a partir de qual ponto começou uma jogada que levou ao chute). Com os eventos que levaram a tentativa de marcar, computamos *features* como total de eventos, passes e dribles, velocidade da jogada e duração total, dentre outros. Por fim, utilizamos a informação de avaliações (ranking) dos jogadores nas partidas para complementar a análise. Todos os atributos utilizados estão na tabela 3.

Atributo	Descrição
start_x	A localização x onde a ação começou
start_y	A localização y onde a ação começou
num_events	Contagem total de eventos (ações) na jogada
num_passes	Contagem total de passes na jogada
num_dribles	Contagem total de dribles na jogada
play_duration	Duração total da jogada
player_rank	Ranking do jogador
play_distance	Distância (euclidiana) total percorrida na jogada
play_mean_distance_to_the_goal	Distância (euclidiana) média de cada jogada até o gol
play_std_distance_to_the_goal	Desvio padrão da distância (euclidiana) de cada jogada até o gol
play_distance_towards_goal	Distância percorrida nas jogadas em direção ao gol (considerando só o eixo x)
bodypart_name	A parte do corpo do jogador usada para a ação
ratio_distance	Razão entre play_distance e play_distance_towards_goal
total_time_per_play	Desvio padrão da distância (euclidiana) de cada jogada até o gol
play_duration	Duração total das jogadas
total_time_per_play	Média do tempo por jogada (razão entre play_duration e num_events)
play_speed	Razão entre play_distance e play_duration
play_speed_towards_goal	Razão entre play_distance_towards_goal e play_duration

Table 3: Atributos utilizados na descoberta de subgrupos

6 Descoberta de Subgrupos

Os dados disponibilizados pela *Wyscout* possuem diversas granularidades: temos 5 ligas de 5 países diferentes, cada uma com suas características, e cada uma com seus respectivos times e estilo. Visando avaliar essas diferenças, a análise de Descoberta de Subgrupos foi dividida em 3 partes: na primeira, avaliamos os grupos descobertos de todo o dataset com os datasets de duas ligas expecíficas (da Inglaterra e da Espanha), para ver se encontramos diferenças entre uma liga VS toda a "população". Na segunda, escolhemos três times de uma mesma liga, um que terminou no topo da tabela, outro no meio e outro que ficou em último, para analisar se essa diferença de posição se reflete nos padrões. Por fim, avaliamos todo o dataset com duas estratégias diferentes de SD, para avaliar redundância e diversidade entre estratégias.

Em todas avaliações foi utilizado a estratégia de *Beam Search* do pacote do *pysubgroup*, com profundidade 3, buscando 100 subgrupos com largura do Beam de 250. Na terceira, utilizamos o *SSD++* com profundidade 3, largura do beam de 25 e máximo de regras 20 (o que gerou aproximadamente 100 resultados, para comparar com a outra abordagem). Também, em todas as etapas foram avaliados todos os alvos, sendo se foi realmente gol ou não (binário), o xg (binarizado na condição xG ¿= 0.5), e também a VAEP (numérico).

7 Mineração de Sequências

Também foram aplicadas técnicas de mineração de sequências para encontrar padrões de tipos de ações realizadas antes de um chute ao gol. Para realizar a mineração de sequências foram usadas algumas técnicas.

Uma delas foi inspirada em [3]: o *Safe Pattern Pruning* (SPP). Usado originalmente no rugby, com resultados promissores, avaliou-se a viabilidade de aplicar o mesmo algoritmo no contexto do futebol. Para isso, os dados foram adaptados ao formato de entrada do programa publicado pelos autores do artigo: cada linha possui um número para indicar a prosperidade da sequência de ações (1 para resultado promissor e -1 para caso contrário) e uma sequência de ações propriamente dita.

As sequências de ações são definidas com base na troca de posse de bola, até ocorrer algum chute ao gol (*shot*). As ações foram definidas como números, conforme a tabela 4.

ID	Ação
0	pass
1	interception
2	dribble
3	take on
4	tackle
5	foul
6	freekick short
7	cross
8	**shot**
9	clearance
10	throw in
11	goalkick
12	corner short
13	corner crossed
14	keeper save
15	freekick crossed
16	shot freekick
17	bad touch
18	shot penalty

Table 4: Mapeamento de Ações para IDs

Para definir se o resultados das ações foi promissor, foram usadas 3 métricas distintas: o xG, o VAEP e uma classificação binária que indica o resultado da ação (no caso de um chute ao gol, se a ação marcou um ponto). O xG foi estimado com uso de um *RandomForest*, considerando a parte do corpo associada ao chute (categoricamente), a distância até o gol e o ângulo entre o chute e o gol. Já o VAEP foi treinado com um *XGBoost*, discretizando-se com um valor de P_{scores} igual a $\frac{1}{2}$. Ou seja, apenas chutes cuja probabilidade de marcar foi maior que $\frac{1}{2}$ foram considerados promissores.

Os parâmetros usados na execução foram: tamanho máximo dos padrões igual a 5, suporte mínimo igual a 50, multiprocessamento habilitado. Traduzindo-se para as *flags* utilizadas no programa:

```
-L 5 -m 50 -M 1
```

Apesar dos bons resultados no rugby em [3], as sequências obtidas deixaram a desejar quanto aplicadas ao futebol, independentemente da métrica usada. Em todos os casos, as sequências de jogadas obtidas eram muito redundantes, continham passes (0) em excesso ou não eram muito surpreendentes. Por exemplo, uma sequência de passes seguida por um cruzamento (7). Os arquivos de saída gerados podem ser obtidos aqui para o xG, aqui para o binário e aqui para o VAEP.

Uma possível explicação para os resultados encontrados é o fato de o futebol ser um esporte muito dinâmico e que, por essa natureza, exigiria outra abordagem para discretização dos dados, para além das sequências de ações. De fato, muitos gols são (e não são) marcados a partir de uma sequência de passes, por exemplo. Desse modo, uma abordagem que consiga mapear o campo de maneira mais sutil é necessária para que se possa realizar uma mineração de sequências usando o algoritmo proposto por [3].

8 Resultados

8.1 Avaliação geral

Na tabela 5 são apresentados os resultados das métricas obtidas em cada uma das etapas, considerando cada um dos algoritmos, dataset utilizado e os diferentes alvos. Podemos ver que, de maneira geral, o *expected Goals* (xG) foi o que desempenhou pior dos três alvos, com tendências de WRAcc mais baixas, e coberturas também. O VAEP e o resultado real (se foi gol ou não) tiveram comportamento bem semelhante, sendo mais altos ou baixo juntos (indicando uma certa correlação). Por fim, apesar de algumas variações nos resultados, o SSD++ demonstrou ser uma estratégia bem mais abrangente

no quesito de cobertura, ao custo de um WRAcc menor. Esse resultado condiz com os resultados do artigo **??**, onde isso também foi percebido. Vale ressaltar que apesar dos testes envolvendo Beam Search terem coberturas altas (pois foram gerados 100 subgrupos), os resultados são bem redundantes (veja subseção 5.4).

Algoritmo	**Dado**	**Target**	**WRAcc do melhor**	**Cobertura Total**
Beam Search	Todo dataset	Gol	0.0312	0.5099
Beam Search	Todo dataset	xG	0.0187	0.5046
Beam Search	Todo dataset	VAEP	0.0307	0.5099
Beam Search	Inglaterra	Gol	0.03192	0.4980
Beam Search	Inglaterra	xG	0.0203	0.4933
Beam Search	Inglaterra	VAEP	0.0317	0.4977
Beam Search	Espanha	Gol	0.0307	0.5110
Beam Search	Espanha	xG	0.0195	0.3983
Beam Search	Espanha	VAEP	0.0304	0.5120
Beam Search	Man City	Gol	0.0483	0.5721
Beam Search	Man City	xG	0.0303	0.4959
Beam Search	Man City	VAEP	0.0487	0.5656
Beam Search	Newcastle	Gol	0.0405	0.6561
Beam Search	Newcastle	xG	0.0247	0.3878
Beam Search	Newcastle	VAEP	0.0420	0.6561
Beam Search	West Bromwich	Gol	0.0293	0.5499
Beam Search	West Bromwich	xG	0.0307	0.3989
Beam Search	West Bromwich	VAEP	0.0287	0.5254
SSD++	Todo Dataset	Gol	0.0164	0.7818
SSD++	Todo Dataset	xG	0.0117	0.9956
SSD++	Todo Dataset	VAEP	0.0194	0.8114

Table 5: Resultados das métricas para todas as avaliações

8.2 Avaliação entre ligas

Na tabela 6 temos alguns subgrupos de cada uma das análises dos datasets, considerando os três diferentes alvos.

Dataset	**Target**	**Subgrupo**
Todo dataset	Gol	shot_angle_from_goal $\geq$ 0.60 AND shot_distance_from_goal $<$ 11.26
Todo dataset	xG	shot_angle_from_goal $\geq$ 0.60 AND shot_distance_from_goal $<$ 11.26 AND start_x $\geq$ 96.60
Todo dataset	VAEP	num_dribbles : [0 : 1[AND shot_angle_from_goal $\geq$ 0.60 AND shot_distance_from_goal $<$ 11.26
Inglaterra	Gol	shot_angle_from_goal $\geq$ 0.61 AND shot_distance_from_goal $<$ 11.26
Inglaterra	xG	num_dribbles : [0 : 1[AND shot_distance_from_goal $<$ 11.26 AND start_x $\geq$ 96.60
Inglaterra	VAEP	play_duration $<$ 1.39 AND shot_distance_from_goal $<$ 11.26 AND total_time_per_play $<$ 0.65
Espanha	Gol	shot_angle_from_goal $\geq$ 0.61 AND shot_distance_from_goal $<$ 11.04 AND start_x $\geq$ 96.60
Espanha	xG	shot_angle_from_goal $\geq$ 0.61
Espanha	VAEP	play_distance_towards_goal : [0.0 : 7.35[AND ratio_distance : [0.0 : 0.12[AND shot_distance_from_goal $<$ 11.04

Table 6: Resultados das avaliações de diferentes ligas com todo o dataset

Os subgrupos foram escolhidos dos top-10 considerando WRAcc, com exceção dos do alvo VAEP para os datasets de ligas isoladas, para ter maior variedade. O mais perceptível é a redundância dos

subgrupos, onde todos são arranjos de cinco ou seis atributos principais. Contudo, podemos perceber pequenas diferenças, como nas distâncias e ângulos para o chute. A medida que descemos as listas dos subgrupos obtidos também temos mais variações, aparecendo features como ranking do jogador, quantidade de passes, dentre outras. Outro fato interessante é de que as métricas concordam entre si, nos subgrupos e nos valores.

8.3 Avaliação entre times

Para avaliar os times de uma mesma liga dentro do dataset, escolhemos a *Premier League* (liga inglesa) e três times baseados na sua colocação final, sendo eles o *Manchester City* (1º na temporada 17/18), *Newcastle* (10º) e *West Bromwich* (20º). O intuito foi avaliar se as estratégias de SD conseguiriam identificar diferenças que possam indicar o motivo ou reflexos dele que fizeram os times ficarem nessa posição. Na tabela ?? temos alguns resultados dos top-30 subgrupos (com base em qualidade):

Dataset	**Target**	**Subgrupo**
Man City	Gol	player_rank : [0.02:0.03[AND shot_distance_from_goal < 10.90 AND start_x > 96.60
Man City	xG	num_dribbles : [0:2[AND shot_angle_from_goal ≥ 0.62 AND start_x ≥ 96.60
Man City	VAEP	bodypart_name =foot_left AND shot_distance_from_goal < 11.03
Newcastle	Gol	shot_angle_from_goal ≥ 0.61 AND shot_distance_from_goal < 11.04
Newcastle	xG	num_dribbles : [1:3[AND start_x ≥ 96.60 AND start_y : [33.32:38.08[
Newcastle	VAEP	bodypart_name = foot_right AND shot_distance_from_goal < 11.04
West Bromwich	Gol	bodypart_name = head/other AND shot_distance_from_goal < 10.52 AND start_x ≥ 96.60
West Bromwich	xG	num_dribbles : [0:2[AND start_x ≥ 96.60
West Bromwich	VAEP	bodypart_name = head/other AND play_distance_towards_goal : [39.90:56.70[AND start_x

Table 7: Resultados das avaliações entre times

Novamente, os diferentes alvos concordaram entre si, encontrando praticamente os mesmos subgrupos, em ordens um pouco diferentes (especialmente no top-10). Podemos perceber diferenças interessante entre os times. O Manchester City, por ser um time com mais jogadores caros e famosos, tem o atributo do ranking do jogador alto (no dataset a maioria das notas são menores que 0.2, e ainda ponderamos e normalizamos pelos minutos jogados). Também percebemos uma maior tendência de chutes com a perna esquerda, poucos dribles e ataques mais perto do gol. Já para o Newcastle, temos mais chutes com a perna direita, e chutes de uma distância maior, o que pode indicar maiores chutes de fora da área. Temos também uma taxa de dribles maior. Por fim, o West Bromwich tem um grande destaque para lances de cabeça, o que pode indicar problemas de construção de jogadas, precisando assim recorrer a cruzamentos e passes longos.

Avaliando além dos top-30 subgrupos, percebemos que a contagem de passes nas jogadas é menor no Newcastle, sendo menor que 5, ressaltando a ideia dos chutes de fora da área, enquanto do Manchester City fica entre 6 e 18, e a do West Bromwich ficou entre 26 e 51. Considerando nossa abordagem de seleção dos eventos, concluímos que o Man City é um time bem mais efetivo no ataque, já que tem mais chutes e com menos passes. Contudo, é um time que toca bastante a bola, característica famosa do seu técnico. Já o WB se mostra menos efetivo na criação, tocando muito a bola, o que pode indicar recuos e passes infrutíferos.

8.4 Avaliação entre algoritmos de SD

Até o momento, todas as análises foram realizadas com Beam Search, e foi percebida uma grande redundância nos resultados. Visando combater esse fator, e também comparar a abordagem do Beam Search como um todo e a granularidade de todo o dataset, aplicamos o algoritmo do SSD++. Alguns dos top-10 subgrupos encontrados estão na tabela ??.

Dataset	Target	Subgrupo
Beam Search	Gol	$start_x \geq$ 96.60
Beam Search	Gol	num_dribbles : [0:1[AND shot_angle_from_goal $\geq$ 0.60 AND shot_distance_from_goal $<$ 11.26
Beam Search	xG	$start_x \geq$ 96.60 AND start_y : [31.96:37.40[
Beam Search	xG	shot_angle_from_goal $\geq$ 0.60 AND shot_distance_from_goal $<$ 11.26
Beam Search	VAEP	$tart_x \geq$ 96.60
Beam Search	VAEP	num_dribbles : [0:1[AND shot_distance_from_goal $<$ 11.26
SSD++	Gol	$shot\_angle\_from_goal \geq$ 0.5591 AND bodypart_name = foot_right AND start_x $\geq$ 95.55
SSD++	Gol	player_rank ¿= 0.016 AND 2.0 ¡= num_passes ¡= 3.0 AND 0.359 ¡= shot_angle_from_goal ¡=
SSD++	xG	27.2 $\leq start_y \leq$ 34.0 AND 4.64 $\leq play_speed \leq$ 6.87 AND $num_dribbles \geq$ 1.0
SSD++	xG	play_speed $\geq$ 8.98 AND play_duration $\geq$ 2.07 AND 0.004 $\geq$ player_rank $\geq$ 0.016
SSD++	VAEP	shot_angle_from_goal $\geq$ 0.56 AND 0.004 $\leq$ player_rank $\leq$ 0.016 AND 0.93 $\leq$ total_time_per_p
SSD++	VAEP	shot_angle_from_goal $\geq$ 0.56 AND play_speed_towards_goal $\geq$ 1.52 AND 12.39 $\leq$ play_mean_d

Table 8: Descrição dos dados no formato SPADL

Avaliando os resultados, conseguimos notar indiretamente o resultado da métrica de cobertura do SSD++ a partir dos vários atributos que não eram comummente vistos no Beam Search, como tempo total por jogada, média e variança da distância para o gol, etc. Essas métricas também são bem relevantes, e ajudam a dar mais informação sobre o comportamento das jogadas. Por exemplo, existem muitas regras que limitam o ranking do jogador para baixo, o que pode ser relevante para uma contratação de um atacante, onde se o restante do time consegue colocar esse jogador em determinada posição, terá uma chance interessante de marcar, e pode não ser um jogador tão caro.

Diferente do que foi visto em **??**, aqui não obtivemos métricas de qualidade negativas, o que atribuímos ao nosso processamento de dados, que filtra as jogadas para só momentos de gols, dificultando avaliações de lances de não-gol. Contudo, acreditamos que essa é também uma análise interessante, e está no radar para trabalhos futuros. Por fim, comparando os dois algoritmos, o Beam Search se mostrou mais direto em suas análises, focando sempre nos mesmos conjuntos de atributos que considerou mais importantes. Já o SSD++ trouxe uma variedade bem maior de atributos, o que podem ser relevantes para análises mais específicas, com o preço de uma qualidade média menor. Assim, acreditamos que a melhor abordagem é avaliar os dois (e outras técnicas) juntas, para ter o melhor das duas vertentes.

9 Conclusão

Neste trabalho avaliamos estratégias de Descoberta de Subgrupos e Mineração de Frequências para dados de ações de futebol. Na parte de SD, apresentamos um framework de tratamento e extração de features, e realizamos uma avaliação sistemática de diferentes granularidade do dataset, para diferentes alvos e diferentes estratégias.

Do ponto de vista dos alvos, a VAEP se mostrou coerente com os resultados obtidos, em especial com o valor binário de se o lance resultou em gol ou não. Usando a VAEP com Beam Search levou a praticamente os mesmos subgrupos do xG e do Gol/Não Gol, já usando SSD++ houveram algumas variações no top-10 melhores grupos (referente a qualidade), mas analisando todos os grupos também são parecidos. O xG levou a valores médios de qualidade menor, mas isso não refletiu muito nos subgrupos em si. Acreditamos que isso se deve ao cálculo e discretização do xG, que leva a menos incidências positivas do que o resultado da ação em si. Contudo, isso levou a uma maior cobertura no SSD++. Concluímos portanto que nenhuma métrica se sobressai em todos os quesitos, e reforçamos que todas sejam utilizadas para avaliações.

Já no ponto de vista dos algoritmos, ocorreu o esperado do SSD++ ter subgrupos bem mais diversos e abrangentes (levando a maior cobertura), ao ponto de ter subgrupos com qualidade menor. Acreditamos que os dois resultados são complementares, e uma melhor análise deve levar em conta os dois resultados.

Por fim, avaliando a utilidade dos subgrupos gerados, temos que os grupos do Beam Search são mais simples e diretos, mas podem ter utilidade, em especial para comparar times distintos. O SSD++ traz análises mais exóticas e com parâmetros que não costumam aparecer no Beam Search, o que pode

ser útil para estratégias específicas. Acreditamos que a melhor maneira é usar os dois modelos, para entender o quão bom é a análise exótica do SSD++, e ter uma ideia geral com o Beam Search.

Referências Bibliográficas

[1] Jay Ayres, Jason Flannick, Johannes Gehrke, and Tomi Yiu. Sequential pattern mining using a bitmap representation. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 429–435, 2002.

[2] Clive B. Beggs, Alexander J. Bond, Stacey Emmonds, and Ben Jones. Hidden dynamics of soccer leagues: The predictive 'power' of partial standings. *PLOS ONE*, 14(12):1–28, 12 2019.

[3] Rory Bunker, Keisuke Fujii, Hiroyuki Hanada, and Ichiro Takeuchi. Supervised sequential pattern mining of event sequences in sport to identify important patterns of play: an application to rugby union. *PloS one*, 16(9):e0256329, 2021.

[4] M. Cox and T. Benjamin. *Entre Linhas: de Ajax a Zidane, a construção do futebol moderno nos gramados da Europa*. Editora Grande Área, 2022.

[5] Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '19, pages 1851–1861, New York, NY, USA, 2019. ACM.

[6] Pappalardo et al. A public data set of spatio-temporal match events in soccer competitions. *Nature Scientific Data*, 2019.

[7] J. Van Haaren, V. Dzyuba, S. Hannosset, and J. Davis. Automatically discovering offensive patterns in soccer match data. 2015.

[8] Miguel Paulo Martins Marques. Subgroup discovery in soccer data. Master's thesis, 2022.

[9] Jian Pei, Jiawei Han, B. Mortazavi-Asl, H. Pinto, Qiming Chen, U. Dayal, and Mei-Chun Hsu. Prefixspan,: mining sequential patterns efficiently by prefix-projected pattern growth. In *Proceedings 17th International Conference on Data Engineering*, pages 215–224, 2001.

[10] Hugo M Proença, Peter Grünwald, Thomas Bäck, and Matthijs van Leeuwen. Discovering outstanding subgroup lists for numeric targets using mdl. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020 , Proceedings, Part I*, pages 19–35. Springer, 2021.

[11] Leszek Szczecinski and Aymen Djebbi. Understanding draws in elo rating algorithm. *Journal of Quantitative Analysis in Sports*, 2020.

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

[13] A. Terzis. *Pep Guardiola - Coaching High Pressing Tactics & Sessions Against Different Formations*. SoccerTutor.com, 2023.

[14] Iacopo Vagliano, Maurice Y Kingma, Dave A Dongelmans, Dylan W De Lange, Nicolette F de Keizer, Martijn C Schut, MS Arbous, DP Verbiest, LF te Velde, EM van Driel, et al. Automated identification of patient subgroups: A case-study on mortality of covid-19 patients admitted to the icu. *Computers in Biology and Medicine*, 163:107146, 2023.

[15] Maaike Van Roy, Pieter Robberechts, Tom Decroos, and Jesse Davis. Valuing on-the-ball actions in soccer: A critical comparison of xt and vaep. In *Proceedings of the AAAI-20 Workshop on Artifical Intelligence in Team Sports*, AITS. AI in Team Sports Organising Committee, dec 2020.