

Trabalho Final

Masters em Big Data Aplicado ao Futebol

Sports Data Campus

Tiago Alves, Vinicius Incrocci, Guilherme Ferreira

Introdução

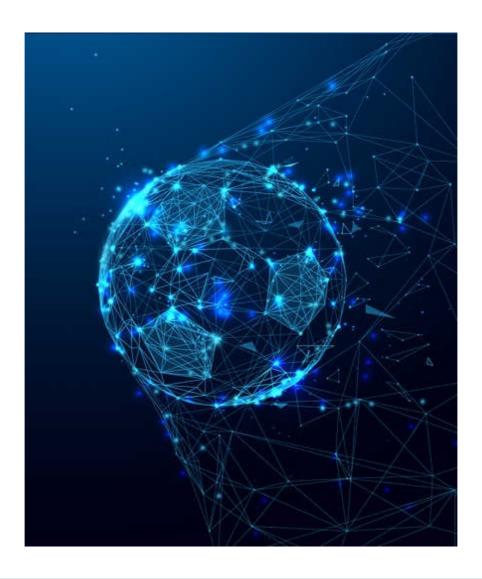
Nosso trabalho final no Masters em Big Data Aplicado ao Futebol consistiu na criação de um modelo de machine learning para a identificação de estilos de jogo a partir de métricas de desempenho coletivo.

O objetivo era encontrar correspondências entre a equipe que usa essa ferramenta (nesse caso, o Sporting) e o time com a qual se negocia o empréstimo de um jogador.

Para potencializar o desenvolvimento do atleta, é interessante que ele esteja inserido em um modelo de jogo semelhante ao do time que detém os direitos dele. Portanto, esse trabalho ajudaria a filtrar bons destinos para o empréstimo.

Usamos métricas de desempenho coletivo provenientes do Wyscout para criar o nosso modelo de clustering, que separou as equipes das cinco principais ligas nacionais da Europa em cinco grupos.

Além de apresentar essa divisão, também buscamos analisar e interpretar quais as características mais marcantes dos times cada cluster.



Documento de Opções

Iniciamos o nosso trabalho final buscando materiais que falassem sobre o uso de dados para a identificação de modelos de jogo de diferentes equipes.

Encontramos textos e pesquisas acadêmicas que nos ofereceram um direcionamento sobre quais caminhos podemos seguir e quais métricas podem ser mais importantes no nosso trabalho.

Nesta parte do documento, vamos apresentar de forma resumida os materiais que consideramos mais úteis para esta etapa. Em seguida, vamos explicar como eles podem servir de inspiração para o nosso trabalho e o que projetamos fazer a partir disso.

1. Attacking and defensive styles of play in soccer

A pesquisa conduzida por Javier Fernandez-Navarro, Luis Fraduab, Asier Zubillagac, Paul R. Forda e Allistair P. McRoberta e publicada no Journal of Sports Sciences em abri lde 2016 coletou dados de 97 partidas da Premier League e da La Liga nas temporadas 2006/07 e 2010/11 para analisar e classificar o estilo de jogo de 37 equipes.

Ao todo, 19 indicadores de performance foram usados para classificar os times de acordo com seis fatores listados abaixo:

Fator 1 (objetividade da posse de bola): indica se uma equipe ataca de forma direta ou com trocas de passes longas

Fator 2 (amplitude da recuperação de bola): indica quais times pressionam e recuperam a bola em áreas centrais e quais fazem isso perto das laterais

Fator 3 (uso de cruzamentos): distingue as equipes entre aquelas que usam muitos cruzamentos e as que usam pouco esse recurso

Fator 4 (amplitude da posse de bola): aponta quais times usam mais a amplitude enquanto tem a bola e quais usam mais as regiões centrais do campo

Fator 5 (pressão no adversário com a bola): indica quais times pressionam mais no campo de ataque e quais pressionam mais no campo de defesa

Fator 6 (progressão do ataque): velocidade que a equipe sai da defesa para o ataque

Os pesquisadores usaram um método estatístico chamado Kaiser-Meyer-Olkin para identificar o número adequado de fatores e os indicadores de desempenho que deveriam ser associados a cada um deles.

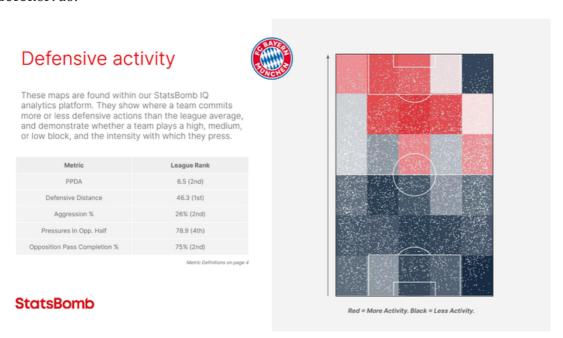
A pontuação das equipes em cada um desses fatores determina qual o estilo de jogo predominante nelas. O produto final do estudo foi uma tabela que apontava as características de maior destaque em cada time.

2. Defensive styles in the Bundesliga

A StatsBomb produziu um relatório focado na análise do estilo de jogo defensivo das equipes da Bundesliga na temporada 2021/22. Eles analisaram todos os 18 times a partir de um mesmo grupo de métricas que definia os principais comportamentos (especialmente os coletivos) sem bola.

O relatório do time começava com uma tabela apresentando o desempenho dele de acordo com cinco indicadores: PPDA, distância do time em relação à própria meta, % de agressividade do oponente que recebe a bola, pressões no campo adversário e % de acerto de passes do adversário.

Ao lado o mapa de um campo de futebol apresenta em quais regiões o time teve mais ações defensivas.



Na página seguinte, a análise é mais individualizada. Um gráfico de dispersão aponta quais jogadores mais pressionam e recuperam a bola num espaço de cinco segundos após a perda da posse.

Counterpressures 1.2 1.1 Counterpressure: Pressures exerted within 5 seconds 0.9 Counterpressure Regain: 0.7 Times a team won the ball back within 5 seconds of a player 0.6 counterpressuring an opponent. 0.5 Data is per 90 minutes & only includes players with a minimum 600 minutes played. 0.3 0.1 3.5 4.0

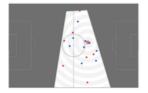
StatsBomb

Por fim, o relatório apresenta outro mapa de um campo de futebol indicando em quais regiões do campo o time mais concentra jogadores a seis jardas do oponente com a bola. Essas informações são adquiridas a partir de dados de tracking do StatsBomb 360.



Using our StatsBomb 360 event data (freeze frame player location data), we can identify how many players each team tend to have positioned around the ball when the opposition is in possession - an indicator of press intensity.

This is done by calculating how many players are within six yards of each opposition ball receipt - an event that logs when a player has received the ball from a pass.



StatsBomb



3. The Five Kingdoms of Football

Esse texto do The Athletic, escrito por John Muller, usa fortes recursos literários para apresentar graficamente como os times das cinco principais ligas da Europa estão distribuídos em relação ao seu modelo de jogo.



Apesar de contar uma história fictícia, o texto é totalmente fundamentado em dados que são aplicados em um algoritmo de machine learning para separar os clubes no mapa dos cinco reinos.

O autor apresenta no fim do texto a "receita" usada para produzir o material. Ele usou 18 indicadores de desempenho a partir de dados do FBref e os separou em quatro categorias: posse de bola, passes, condução de bola, defesa.

A predominância de um time em uma determinada métrica faz o time "se mover" em meio aos 'clusters' no mapa para mais próximo de outras equipes que possuem a mesma característica.

4. Analysis of styles of play in soccer and their effectiveness

Fernández Navarro, Francisco Javier. "Analysis of styles of play in soccer and their effectiveness." (2019)

Esta tese de doutorado teve como objetivo identificar os estilos de jogo e as respetivas características, praticado pelas diferentes equipas de elite, para além de examinar de que forma determinadas variáveis contextuais que influenciam o estilo praticado e ainda a eficácia de destes diferentes estilos sob diferentes circunstâncias.

Foram analisados 97 jogos da Premier League e da La Liga das épocas 2006/07 e 2010/11, onde usando um sistema de análise de jogo computadorizado, foi possível demonstrar que os diferentes estilos atacantes/defensivos podem ser determinados através da análise das diferentes variáveis táticas de cada equipa, recorrendo a uma factor analysis, que permite agrupar (cluster) variáveis e assim identificar diferentes estilos de jogo.

Para tal, foram definidas 8 categorias de forma a caracterizar o estilo atacante de uma equipa e 4 categorias para definir o estilo de jogo defensivo.

Estilo de jogo atacante

D (direct)
P (possession)
C (crossing)
NC (no crossing)
WP (wide possession)
NP (narrow possession)
FP (fast progression)
SP (slow possession)

Estilo de jogo defensivo

PW (pressure on wide areas)
PC (pressure on central areas)
LP (low pressure)
HP (high pressure)

Para além disto são ainda sugeridas algumas métricas que poderão melhor identificar certos estilos de jogo.

Adicionalmente, foram também analisados 280 jogos da época 2015/16 da Premier League, que demonstraram que variáveis contextuais como a importância do encontro, o estádio e a qualidade do adversário influenciam o estilo de jogo de uma equipa.

5. The Analyst

O site The Analyst, especializado em análise de esportes através das estatísticas, possui uma área de dados abastecida pela Stats Perform.

Essa parte do site cobre dez ligas nacionais e apresenta tabelas e gráficos com estatísticas coletivas e individuais.

Entre as estatísticas coletivas, são mostrados o xG (a favor e contra), além páginas chamadas Team Sequences (Style) e Team Sequences (Pressure). A primeira possui métricas que indicam se uma equipe avança em campo de forma rápida e direta ou devagar e pausadamente. A segunda aponta o quão forte é a marcação do time e o que acontece quando ele recupera uma bola perto do gol adversário.

A primeira parte do trabalho final terá esses materiais citados acima como uma das fontes de inspiração, já que um dos nossos objetivos é organizar times de acordo com seus estilos de jogo e todos eles fazem isso, mesmo que de formas distintas.

As pesquisas e textos apresentados por nós neste documento não explicam com profundidade os motivos para escolher cada métrica usada. No entanto, o estudo publicado no Journal of Sports Sciences em 2016 indica qual caminho foi seguido nesse processo.

"Pesquisas anteriores relacionadas a tática foram consideradas na seleção dos seguintes indicadores de performance para o estudo: posse de bola, cruzamentos e finalizações. Os indicadores de performance restantes [...] foram considerados relevantes para determinar os estilos de jogo pela importância da ocorrência espacial dos eventos para medir aspectos táticos", diz o estudo.

Esse e outros materiais apresentados aqui seguem uma linha de raciocínio parecida. Os autores criam categorias focadas em uma característica marcante para o modelo de jogo (categoria posse de bola, por exemplo, que pode definir times como controladores do jogo ou como reativos, além do meio termo existente entre esses dois extremos) e em seguida selecionam as métricas que se encaixam em cada categoria e são relevantes para medir aspectos táticos.

Optamos por seguir esse caminho no nosso trabalho final escolhendo categorias que servem de parâmetro para explicar o modelo de jogo e em seguida buscando métricas que traduzem os eventos mais importantes dentro dessa categoria.

Nesta etapa, sugerimos a seguinte divisão (os números entre parênteses indicam qual das referências citadas foi usada ou serviu de inspiração para a métrica):

Controle da bola

- 3 % de toques na bola do time no terço final do campo (em relação ao total de toques na bola no terço final somando os dois times)
- 3 % de passes curtos (em relação ao total de passes curtos durante o jogo) Zonas de ataque
- 1 % de toques na bola nas pontas do campo (em relação ao total de toques na bola no terço de ataque)
- 3 % de cruzamentos para a grande área (em relação ao total de passes para a grande área)
- 3 % de through balls para a grande área (em relação ao total de passes para a grande área)

Construção

- % de passes longos (em relação ao total de passes do time)
- % de passes curtos (em relação ao total de passes do time)
- 3 % de passes longos (acima de 40 jardas) do goleiro (em relação ao total de passes do goleiro)
- 3 média de distância progressiva (distância total que um passe viaja em direção ao gol adversáio) a cada passe completo
- 3 média de carregadas progressivas (distância total que um jogador carrega a bola em direção ao gol adversáio) a cada carregada de bola
- 3 % de dribles (em relação ao número de toques na bola)
- 5 distância direta (em metros por segundo)

Pressão defensiva

- PPDA
- 2 % de agressividade (proporção de passes do adversário que antecedem um desarme, falta ou pressão dentro de dois segundos)

Altura da defesa

- 2 Distância defensiva
- 3 % de pressões no terço de ataque (em relação ao total de pressões)
- 3 % de pressões no terço defensivo (em relação ao total de pressões)

Além de usar indicadores que mensuram as ações que acontecem dentro de campo, pensamos também em considerar outros fatores que podem interferir diretamente na adaptação de um jogador emprestado a outro clube.

A função de um lateral, por exemplo, muda totalmente a depender do número de defensores. O seu trabalho também pode ser influenciado por uma tendência do treinador a usar mais ou menos jovens no elenco.

Temos também a intenção de avaliar o nível competitivo e o contexto das ligas para onde o atleta pode ser emprestado. Acreditamos que um jogador pode ter mais ou menos facilidade de adaptação em certos campeonatos do que outros.

Pontos como a força dos adversários (que pode ser medida pelo ranking da Uefa, por exemplo), média de público e atributos físicos e estilo de jogo (média de faltas, cartões, passes, etc.) merecem uma análise cuidadosa para embasar a decisão pelo empréstimo. Em resumo esses são os fatores extras que queremos considerar no trabalho final:

Formação tática do time Uso de jovens no clube Nível de dificuldade da liga Média de público da liga Estilo de jogo da liga em relação a faltas

Fontes:

Attacking and defensive styles of play in soccer (Javier Fernandez-Navarro, Luis Fraduab, Asier Zubillagac, Paul R. Forda, Allistair P. McRoberta)

Defensive Styles in the Bundesliga 2021/22 Season (StatsBomb)

The Five Kingdoms of Football (John Muller): https://theathletic.com/3044091/2022/01/02/the-five-kingdoms-of-football/

Analysis of styles of play in soccer and their effectiveness (Fernández Navarro, Francisco Javier)

The Analyst: https://dataviz.theanalyst.com/season-reviews/2021/

Plano de desenvolvimento

Dando sequência ao nosso trabalho final no Masters em Big Data Aplicado ao Futebol, apontamos os passos que precisamos seguir e nossas necessidades na fase de desenvolvimento do projeto.

Detalhamos a nossa ideia sobre como e quais tipos de dados pensamos em obter (focando especialmente no Wyscout) para aplicar ao modelo de machine learning.

1. Wyscout >> Advanced Stats >> Teams Advanced Stats (todas as métricas - filtrados por campeonatos europeus de primeira e segunda divisão)

Aqui queremos as métricas coletivas das equipes dos principais campeonatos da Europa. É com esses números que queremos identificar padrões e classificar os times de diferentes países.

Em um primeiro momento imaginamos usar as métricas citadas ao final deste planejamento, seguindo o padrão apresentado no Documento de Opções. No entanto, gostaríamos de ter todas as estatísticas para testar diversos modelos e simulações, a fim de fazer o melhor trabalho possível (além de não precisar pedir novas bases de dados no futuro).

Abaixo, colocamos as métricas iniciais que estão disponíveis no Wyscout e imaginamos utilizar para o modelo, seguindo a divisão por categorias apresentada no Documento de Opções:

Controle da bola

deepCompletedPasses (Deep completed passes): A non-cross Pass that is targeted to the zone within 20 meters of the opponent's goal deepCompletedPassesSuccessful (Successful deep completed passes)

passToFinalThirdsSuccessful (Successful pass to final thirds) touchesInBox (Touches in box): An action (a Pass or a Touch) that happens in the opponent penalty area. Duels are excluded from this definition

positional Attack (Positional attacks) total (Total attacks)

Zonas de ataque

centerAttacks (Attacks from center) leftFlankAttacks (Attacks from left flank) rightFlankAttacks (Attacks from right flank)

Construção

passes (Total passes): An attempt to pass the ball to a teammate shortMediumPasses (Short medium passes) longPasses (Long passes)

veryLong (Very long open plays) long (Long open plays) medium (Medium open plays) short (Short open plays) total (Total open plays)

progressiveRuns (Progressive runs): A continuous ball control by one player attempting to draw the team significantly closer to the opponent goal progressivePasses (Progressive passes): A forward pass that attempts to advance a team significantly closer to the opponent's goal

Pressão defensiva

ppda

counterpressingRecoveries (Counterpressing recoveries): Any Recovery that ends a Possession of the opposition team with length less than 5 seconds challengeIntensity (Challenge intensity): Number of defensive actions (defensive duels, loose ball duels, interceptions, tackles) per minute of opponent ball possession

Altura da defesa

recoveriesHigh (Recoveries in the final third) recoveriesLow (Recoveries in the own third) recoveriesTotal (Total recoveries)

Desenvolvimento

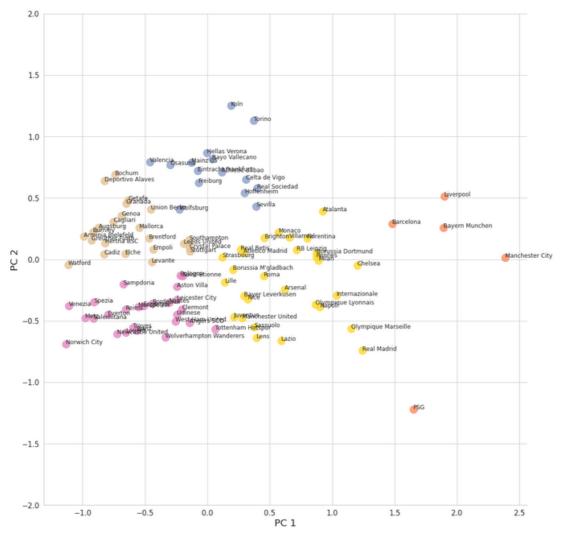
Recebemos as métricas de todos os times das cinco principais ligas nacionais da Europa e começamos a produzir o modelo de machine learning que nos levaria a agrupar as equipes em clusters de acordo com o estilo de jogo de cada uma delas.

Testamos diferentes conjuntos de métricas para aplicar ao modelo buscando conciliar dois pontos fundamentais: que o gráfico final tivesse fronteiras bem delimitadas entre um cluster e outro, e que as métricas escolhidas fossem capazes de traduzir bem um modelo de jogo.

Fizemos o primeiro teste com todas as estatísticas disponíveis no dataset para entender como o modelo responderia e depois aplicamos ajustes pensando nos indicadores que havíamos selecionado nas etapas anteriores do trabalho.

Chegamos ao produto final com 37 métricas colocadas no modelo, sendo que 20 dizem respeito ao momento com bola da equipe e 16 estão relacionadas ao momento sem bola. Uma delas, as disputas aéreas, pode caracterizar as duas fases do jogo.

No processo de elaboração do modelo, notamos que o número ideal de clusters era cinco e finalmente criamos a representação gráfica com os times distribuídos nesses grupos. Para mais detalhes sobre a elaboração do código, enviamos também o arquivo .ipynb.



Depois de criar o gráfico, chegou o momento de analisar quais as características mais marcantes do estilo de jogo dos clusters, que foram nomeados da seguinte forma: Norte, Noroeste, Sudoeste, Sudeste e Leste.

Importante ressaltar que, ao rodar o código mais uma vez para reproduzir o processo e gerar o gráfico novamente, é possível que alguns times que estão nas "margens" dos cluster mudem de grupo. No entanto, acreditamos que isso não compromete a avaliação sobre os modelos de jogo.

Para chegar às conclusões apresentadas abaixo, calculamos a média de cada cluster nas métricas que usamos dentro do modelo de machine learning. Os times do Leste, por exemplo, tiveram uma média de 16,2 finalizações, a maior entre os cinco clusters.

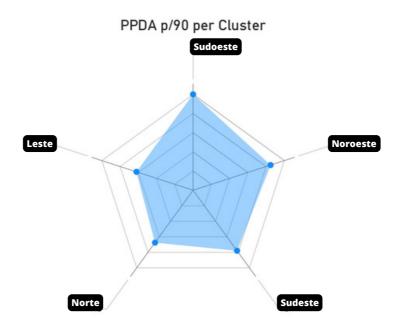
Comparamos os valores médios para cada métrica e anotamos quais clusters mais se destacavam (positivamente ou negativamente) nelas.

SUDOESTE

Destaques sem bola

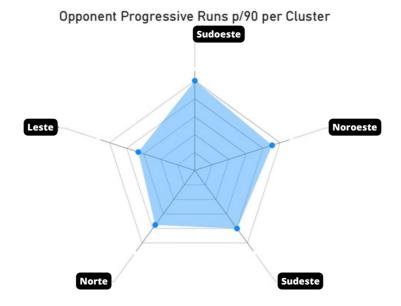
- Poucas recuperações de bola e alto PPDA
- Muitos gols sofridos e muita posse de bola cedida ao adversário

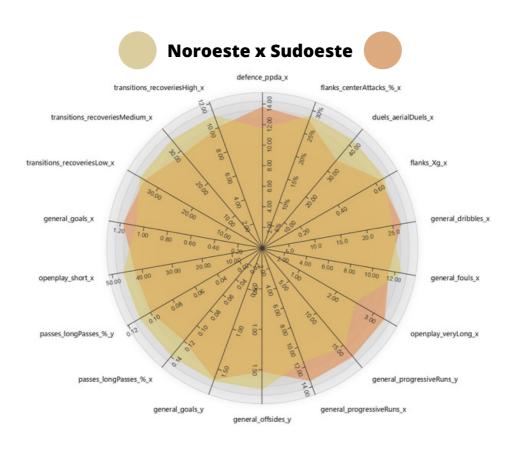
Também cede campo ao adversário e permite a ele avançar sem grandes obstáculos. Em comparação com o cluster Noroeste, é ainda menos combativo, pois não se destaca nem no terço inicial do campo.



Destaques com bola

- Poucos remates e alto número de passes por finalização
- Alto número de perdas de bola no próprio campo
- Baixa produção ofensiva
- Há uma ligeira diferença (23%) entre o Noroeste (último) e o Sudoeste (penúltimo) em posses de bolas longas





NOROESTE

Destaques sem bola

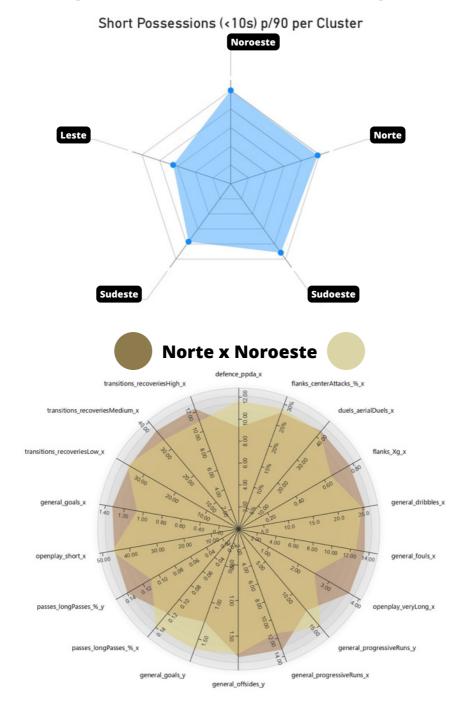
- Muitas faltas cometidas
- Muitos gols sofridos e muita posse de bola cedida ao adversário
- Recuperações de bola no primeiro terço do campo

Cede campo ao adversário e comete muitas faltas. É consideravelmente mais combativo no primeiro terço do gramado (em relação às outras zonas) e cede muitas oportunidades de gol.

Destaques com bola

- Poucos remates e gols marcados
- Posses de bola extremamente curtas
- Grande distância média de passes
- Muitos duelos aéreos
- Há uma grande diferença (113%) entre o Noroeste (1°) e Leste (5°) em porcentagem de passes longos

Aposta em passes longos para chegar ao ataque e termina suas jogadas muito rapidamente, seja com uma finalização ou devolvendo a bola ao adversário. Os lançamentos também provocam um número elevado de duelos pelo alto.

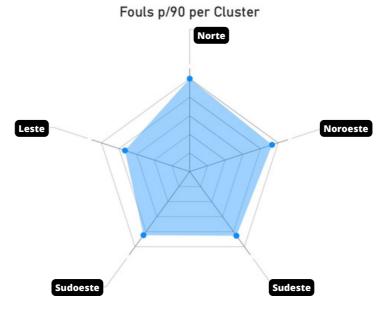


NORTE

Destaques sem bola

- Recuperações de bola no primeiro terço do campo e no terço médio
- Alto número de faltas cometidas
- Muitos duelos aéreos

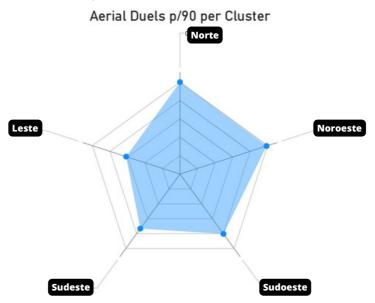
Físico, agressivo e competitivo na marcação, especialmente nas zonas baixas e intermediárias do campo. Incomoda o adversário a ponto de fazer ele perder a posse de bola rapidamente.

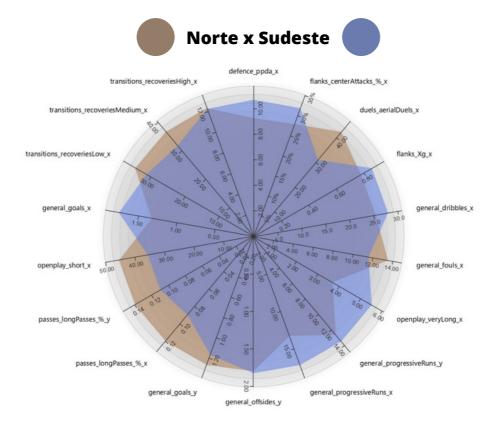


Destaques com bola

- Posses de bola curtas
- Longa distância média de passes e poucos passes por remate
- Alta porcentagem de ataques pelos lados do campo
- Muitos duelos aéreos

Tenta um jogo direto e pelos lados do campo. Chega ao ataque e finaliza com poucos passes (e normalmente de longa distância). O jogo aéreo parece ser importante para avançar no gramado e atacar a grande área.





SUDESTE

Destaques sem bola

- Recuperações de bola no terço final do campo
- Forçam posses de bola curtas do adversário
- Adversário chega pouco na área

Tenta uma marcação mais adiantada em campo e assim força posses de bola curtas do adversário, que em geral chega pouco na grande área.

Noroeste Sudeste Norte

Destaques com bola

- Boa produção ofensiva
- Muitos ataques pelo centro do campo
- Alto número de passes por finalização

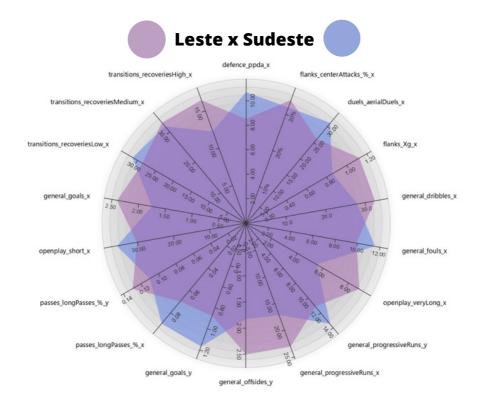
Sudoeste

- Alta porcentagem de contra-ataques

Jogo de posse de bola, usando. Trabalha muito a bola até concluir as jogadas e leva perigo constante ao oponente. No entanto, não apresenta um predomínio nesses quesitos como as equipes do cluster Leste.

Sudeste

Norte



LESTE

Destaques sem bola

- PPDA baixo
- Provoca muitos impedimentos do adversário
- Há uma grande diferença (50%) entre o Leste (1°) e o Sudeste (2°) em contra-ataques permitidos ao adversário

Retoma a posse de bola com muita frequência no terço final do campo (e rapidamente). Consequentemente, não costuma necessitar de recuperações perto da própria grande área. Joga com uma defesa adiantada, que faz os adversários arriscarem passes longos, muitas vezes resultando em impedimento.

Destaques com bola

- Altíssima produção ofensiva
- Poucas perdas de bola no campo de defesa

Sudoeste

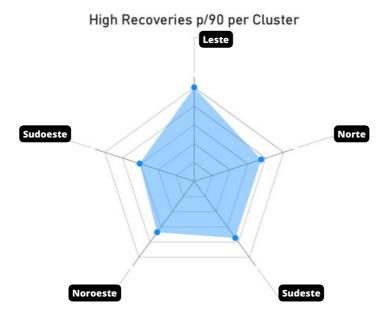
- Posses de bola longas, com passes curtos
- Há uma grande diferença (60%) entre o Leste (1°) e o Sudeste (2°) em posses de bola longas

Long Possessions (>30s) p/90 per Cluster

Leste

Sudeste

Norte



São os times mais propositivos e eficientes. Ocupa campo de ataque e predomínio na posse de bola, encontrando espaços para avançar com ela. Também apresenta os melhores números de produção ofensiva e não aposta no jogo aéreo.

Além das comparações das médias entre os cinco clusters, também observamos as diferenças mais nítidas entre os times que estavam nas extremidades dos dois eixos do gráfico (eixo x e eixo y).

Comparamos o time mais à direita na imagem com o time mais à esquerda e repetimos o processo usando o time mais acima com o time mais abaixo.

Essa análise serviu como complemento para entender quais fatores contribuem com maior peso para um time estar mais à direita ou acima no gráfico.

A presença nas grandes áreas é importante para o eixo x. Nossa conclusão é que quanto menor for a produção ofensiva do nosso time e quanto menor for a produção ofensiva do adversário, mais à direita o time vai se posicionar.

Para o eixo y, o fundamental é a combatividade. Quanto melhores forem os indicadores de marcação, mais acima no gráfico o time vai aparecer.

O cluster Norte se destaca no quesito combatividade. Poucos times de outros clusters chegam perto da altura alcançada por essas equipes.

O cluster Leste se distancia (e de forma considerável) dos demais quando se mede a presença nas grandes áreas. Os times mais à direita são os que mais produzem ofensivamente, os que mais controlam a partida com a bola e os que menos sofrem na defesa.

Sugestões para trabalhos futuros

Usando este trabalho como ponto de partida, é possível desenvolver outras ferramentas que ajudem ainda mais na busca de clubes que se encaixem na negociação de empréstimo de um jogador.

Abaixo, indicamos alguns pontos que podem ser desenvolvidos:

- 1. Além de filtrar as equipes que possuem um estilo de jogo parecido com o do Sporting, sugerimos observar com quais delas faz mais sentido fazer negócio. Podemos usar o Transfermark para obter o valor global dos elencos e retirar aqueles que são muito mais valiosos do que o nosso clube.
- 2. Além de analisar os clubes, é possível também obter métricas de desempenho individual para comparar os jogadores do Sporting com os atletas do clube alvo identificados e separados no item anterior.
- 3. Também pensamos em buscar dados da quantidade de jogadores portugueses em diferentes ligas. Queremos usar esse número como um critério simples e objetivo sobre a aceitação e adaptação de jogadores dessa nacionalidade nesses países. Podemos obter essa informação do FBref ou do CIES.
- 4. Usando algum rating de avaliação (WhoScored, InStat Index, Opta, etc.), é possível fazer uma comparação final para entender se o jogador do Sporting é melhor avaliado que o titular do clube alvo para empréstimo. O objetivo é observar se há uma tendência dele jogar como titular e ganhar os minutos que o nosso clube tem interesse.