

Anais do **FAME**

FOOTBALL ANALYTICS AND MODELING EXPERIENCE

4ª EDIÇÃO

2025



Patrocinador:



Organização:



DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

UF *m* **G**

Apresentação

O *Football Analytics and Modelling Experience* (FAME), em sua quarta edição, integrou em 2025 uma programação ainda mais ampla, contando com um workshop ministrado em parceria com a **Gradient Sports**. Esta expansão resultou em dois dias de atividades complementares: no primeiro dia, workshop, no qual os participantes tiveram a oportunidade de aprender e construir, na prática, um modelo de xG (Expected Goals); e, no segundo dia, o **FAME'25**, dedicado à ciência de dados, modelagem e aplicações analíticas no futebol. Este compêndio reúne os artigos apresentados no **FAME'25**. A diversidade dos temas, abordagens e autores reflete o crescente interesse pela interseção entre tecnologia, dados e o esporte mais popular do mundo.

Agradecemos especialmente à **GEMINI Sports Analytics** e a **Gradient Sports**, patrocinadoras do FAME'25, cujo apoio foi decisivo para viabilizar esta edição. Reconhecemos também o importante suporte institucional do **Departamento de Ciência da Computação da UFMG (DCC/UFMG)** e da equipe do **Centro de Atividades Didáticas 3 (CAD 3)**, onde os encontros presenciais ocorreram.

Por fim, estendemos nosso agradecimento aos membros do SALab, voluntários e a todas as pessoas que participaram, divulgaram e colaboraram na realização deste evento, conforme listado a seguir.

Comissão Organizadora

FAME'25 – Football Analytics and Modelling Experience

Time de voluntários da UFMG – Universidade Federal de Minas Gerais (em ordem alfabética):

- | | |
|--------------------------------|-------------------------------|
| • Adriano C. Machado Pereira | • Jalmir Ferreira |
| • Ana Clara Vasconcelos | • João Araldi |
| • Augusto Caramelli | • João Henrique Alves Martins |
| • Bernardo Ribeiro | • João Lucas Lage Gonçalves |
| • Bruno Martins | • Júlia Borba |
| • Caio Souza Grossi | • Larissa Dolabella Gomide |
| • Daniel Barreto | • Laura Froede |
| • Davi Ladeira Balsamão | • Leo M. Sá Freire |
| • Felipe Silva Fraga Damasceno | • Luca Kosfeld |
| • Gabriel Alkmim | • Lucas Dayrell |
| • Gabriel Carvalho Gomes | • Lucca Carvalho Augusto |
| • Gabriel Reis | • Luis Felipe Pimenta |
| • Gabriel Valadão Meira | • Luís Henrique Emediato de |
| • Guilherme Cunha | Rezende Lara |
| • Guilherme Henrique | • Mateus Arantes |

- Matheus Nogueira
- Pedro Bacelar
- Rafael Martins Gomes
- Raiza Wunsch Pinto
- Raphael Henrique
- Ricardo Furbino
- Victoria Estanislau
- Victoria Flores
- Wagner Meira Jr.

Comitê de Programa

A quarta edição do FAME - Football Analytics and Modelling Experience – teve uma chamada de trabalhos técnicos, que foram apresentados no evento e, posteriormente, convidados para publicação dos trabalhos neste compêndio (Anais/Proceedings) do evento.

Este comitê de programa foi composto pelos seguintes colaboradores:

- Adriano C. Machado Pereira
- Bernardo Ribeiro
- Bruno Martins
- Daniel Barreto
- Gabriel Reis
- Gabriel Valadão Meira
- Lucca Carvalho Augusto
- Ricardo Furbino
- Wagner Meira Jr.

Os trabalhos submetidos constituíram de textos ou artigos acadêmicos, aplicações industriais e produtos tecnológicos de *startups* que utilizassem métodos quantitativos para solucionar problemas no esporte, preferencialmente no futebol. Sendo que foram sugeridos os seguintes temas, com possibilidade de outros temas relacionados:

- Análise tática e estratégica
- Análise de desempenho de jogadores e equipes
- Recrutamento e *valuation* de atletas
- Previsão e análise de resultados de jogos e competições
- Prevenção de lesões
- Otimização de treinamento e *performance* física

Cada trabalho enviou um resumo de até 1000 palavras, descrevendo o problema tratado e o objetivo geral, a metodologia, os resultados e as conclusões. E posteriormente convidados os(as) autores(as) a gerar uma versão final, com possibilidade de ser um resumo ou um artigo com mais detalhes do trabalho, a ser publicado em um compêndio do evento, que é este documento aqui apresentado.

Esperamos que possam desfrutar da leitura dos trabalhos e aprender mais sobre essa temática do evento.

Cordialmente,

O Comitê de Programa – FAME'25.

Sumário

Apresentamos, a seguir, o conjunto de trabalhos que compõem este compêndio, fruto das contribuições submetidas e selecionadas no âmbito do FAME'25 – *Football Analytics and Modelling Experience*. Os trabalhos aqui publicados passaram por um processo de revisão final após o evento, e incluem artigos completos e resumos expandidos que refletem a diversidade de temas e abordagens presentes na interseção entre ciência de dados e futebol.

Nesta seção, listamos os trabalhos que aceitaram o convite para publicação neste compêndio.

	Título do Trabalho	Pág.
Resumo		
1	Aprimorando o Atomic VAEP com Classificador de Sucesso Avelar Ribeiro Hostalácio, Enzo Pinheiro Pierazolli, Fernando Vilela Brandão, Iago Gabino Goncalves, Victor de Almeida Nunes Murta	6
2	Análise e Predição de Química entre Jogadores de Futebol com Base em Dados de Evento Gabriel Castelo Branco Rocha Alencar Pinto, Victor Kenji Pawlowski Abeki, Vinicius Silva Gomes, Matheus Vaz Leal Lima	18
3	Correlação entre Dados de Evento e Tracking com Lesões no Futebol Profissional Júlio Guerra Domingues, Leandro Luiz Duarte Teixeira, Victor Yuji Yano, Rodrigo Luiz Macedo Ferreira, Denilson Martins Vieira Santos	29
4	Análise da Relação entre Estilos Defensivos, Controle de Jogo e Desempenho Competitivo no Futebol Murilo Jorge de Figueiredo, João Victor de Miranda Gadelha, Balthazar Paixão, Jorge Soares, Glauco Amorim, Pedro Henrique González	35
5	Futmetria: identificando padrões de jogo e assinatura dos times Bernardo Vale dos Santos Bento, Filipe Mauro da Terra Caldeira, Gabriel Martins Miranda, Lucas Augusto da Silva Gonçalves	47
6	Decidindo o Jogo: Uma Nova Perspectiva sobre Finalizações no Futebol Davi Sakamoto Lamounier, Bernardo Viggiano Campos, Gabriel Liberato Matias de Souza, Juan Vinicius Coelho, Theo Lopes Mesquita Pedras	55
7	Entendendo Posicionismo e Relacionismo através de Métricas de Centralidade Daniel Rebouças de Sousa Barros, Hugo Leandro Antunes, Balthazar Paixão, Jorge de Abreu Soares, Glauco Fiorott Amorim, Pedro Henrique González Silva	63

Título do Trabalho		Pág.
Resumo		
8	Otimização Logística no Futebol Brasileiro: uma Análise do Impacto das Viagens no Desempenho dos Times e Proposta de Calendário Ajustado Paulo Henrique Sendas Resende, André Jacinto, Daniel Capanema, Roney Lira	67
9	Agrupamento de Jogadores de Futebol Similares do Campeonato Brasileiro de 2023 Utilizando Algoritmos de Aprendizado de Máquina Não-Supervisionados Gabriel R. Araujo, Natan B. Andrade, José Antônio Ferreira, Jorge de A. Soares, Pedro Henrique González, Glauco F. Amorim	81
10	Análise das Características da Posse de Bola e sua Relação com os Gols Esperados Rodrigo Marques Pau Ferro, Lucas Carneiro Calmon, Lucas Giusti Tavares, Pedro Henrique González, Glauco Fiorott Amorim, Jorge de Abreu Soares	88
11	Predição de Sucesso para Situações de Pressão no Futebol Luiza de Melo Gomes, Caio Souza Grossi, Júlia Souza Moura, Kaique de Oliveira e Silva, Victor Vieira Brito Amaral Pessoa	92
12	Variáveis Táticas Espaciais Associadas à Vitória no Futebol Lenizio Moraes de Oliveira, Gibson Moreira Praça, André Gustavo Pereira de Andrade	97
13	Impacto do Calendário no Desempenho das equipes da Série A do Campeonato Brasileiro Leonardo Pessoa Miranda, João Vitor Bicalho da Silva, João Pedro Reis Teclo de Miranda, Gabriel Chaves Ferreira	101
14	Mapping and Characterization of Defensive Pressing in Football Júlia Paes de Viterbo, Lucas de Oliveira Siqueira, Lucas Affonso, Bruno Butti	112
15	Análise de Mercado e Clusterização na Busca por um Substituto para Trent Alexander-Arnold no Liverpool João Vitor Takahashi Chemin	116
Sessão de Best Papers		119
16	Shared Goal Value: Uma Nova Métrica para Avaliar Contribuições em Gols no Futebol Lucas Carneiro Calmon, Pedro Carvalho da Rocha, Balthazar Paixão, Pedro Henrique González, Glauco Fiorott Amorim, Jorge de Abreu Soares	120
17	xCross: Um Modelo Probabilístico Para Avaliação de Cruzamentos através da Entropia Jalmir de Jesus Ferreira da Silva Junior, Vitor A. R. Azevedo, Lucas V. S. Ramos, Eduardo A. Tomich, Bernardo S. B. de Oliveira	127
18	Analysis of Off-Ball Movement in a Counterattack Using Pitch Control and Deep Learning Models Luca Florentino Kosfeld, Bernardo Venancio Cunha Oliveira, Bruno Soares e Silva, Lucas Albuquerque Santos Costa	138

Aprimorando o Atomic VAEP com Classificador de Sucesso

Avelar Hostalácio¹, Enzo Pierazolli¹, Fernando Vilela¹,
Iago Gonçalves¹, Victor Murta¹

¹Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brazil

avelarrh@ufmg.br, eppierazolli@ufmg.br, fvb@ufmg.br,

iagogabino@ufmg.br, victormurta@ufmg.br

Abstract. *In this work, we propose an enhancement to the Atomic VAEP model by incorporating a success classifier trained with tactical features derived from tracking data. Using detailed spatiotemporal data from the 2023–2024 Premier League season, we developed five contextual variables that better reflect the in-game dynamics at the moment of each action. These variables include defensive compactness, goal-side triangulation, player speed, local pressure and support, and goalkeeper angular deviation. The results show specific improvements over the original model in metrics such as Brier Score, Log Loss, and ROC AUC. The proposed integration enables a more contextualized evaluation of player actions and contributes to more robust and interpretable performance analysis tools in football.*

Resumo. *Neste trabalho, propomos um aprimoramento do modelo Atomic VAEP por meio da incorporação de um classificador de sucesso treinado com features táticas derivadas de dados de tracking. Utilizando dados espaço-temporais detalhados da temporada 2023–2024 da Premier League, desenvolvemos cinco variáveis contextuais que refletem melhor a dinâmica em campo no momento de cada ação. As variáveis incluem compactação defensiva, triangulação para o gol, velocidade do jogador, pressão e suporte locais, e desvio angular do goleiro. Os resultados obtidos mostram melhorias pontuais em relação ao modelo original nas métricas de Brier Score, Log Loss e ROC AUC. A integração proposta permite uma avaliação mais contextualizada das ações dos jogadores e contribui para ferramentas de análise de desempenho mais robustas e interpretáveis no futebol.*

1. Introdução

No futebol moderno, a análise quantitativa tem desempenhado um papel fundamental na avaliação de desempenho de jogadores e equipes, auxiliando técnicos, analistas e departamentos de recrutamento na tomada de decisões baseadas em dados. Com o avanço de tecnologias de coleta e processamento de informações, como sistemas de tracking e modelos preditivos, tornou-se possível quantificar aspectos cada vez mais complexos do jogo.

Entre as abordagens mais promissoras nesse cenário está o modelo Atomic VAEP (Valuing Actions by Estimating Probabilities) [Decroos et al. 2019], que busca atribuir

valor a cada ação individual em campo com base em sua contribuição esperada para aumentar ou diminuir a probabilidade de gol. A proposta central do modelo é oferecer uma métrica que vá além de estatísticas tradicionais como gols e assistências, capturando o impacto tático de ações como passes, finalizações, interceptações e dribles independentemente do seu resultado imediato. Dessa forma, o VAEP permite uma análise mais granular e justa do desempenho dos atletas, valorizando ações de construção e prevenção que muitas vezes passam despercebidas.

No entanto, apesar de sua robustez e ampla aceitação, o modelo original do Atomic VAEP apresenta limitações importantes. Por depender exclusivamente de dados de eventos (event data), ele não capta nuances contextuais e espaciais do jogo que influenciam diretamente o valor real de uma ação. Elementos como a pressão exercida pela defesa, o posicionamento do goleiro, o suporte de companheiros próximos ou a compactação do sistema defensivo adversário são ignorados, o que compromete a fidelidade da avaliação em momentos críticos da partida.

Diante dessas limitações, este trabalho propõe um aprimoramento do modelo Atomic VAEP por meio da incorporação de features táticas derivadas de dados de tracking: dados espaço-temporais que descrevem a movimentação contínua de todos os jogadores e da bola ao longo da partida. Com base nesses dados, desenvolvemos cinco variáveis contextuais que capturam com maior fidelidade a dinâmica do jogo no instante de cada ação: compactação defensiva, triangulação para o gol, velocidade do jogador, pressão e suporte locais, e desvio angular do goleiro.

Essas variáveis alimentam um classificador de sucesso, responsável por estimar a probabilidade de que uma ação tenha êxito sob determinadas condições táticas e espaciais. Ao substituir a estimativa empírica baseada em frequência por uma predição modelada a partir do contexto real de jogo, buscamos aumentar a precisão e a interpretabilidade do valor atribuído a cada ação.

Nosso objetivo é contribuir com o desenvolvimento de ferramentas analíticas mais robustas, que combinem o poder explicativo do modelo VAEP com a riqueza informacional dos dados de tracking, permitindo uma análise de desempenho mais precisa, personalizada e alinhada com os princípios táticos do jogo.

2. Metodologia

A metodologia deste trabalho está organizada em um pipeline experimental dividido em três etapas principais: (1) definição de um modelo de referência com base no Atomic VAEP original, (2) desenvolvimento de um classificador de sucesso utilizando *features* táticas extraídas de dados de tracking, e (3) integração das probabilidades estimadas pelo classificador ao cálculo de valor das ações no modelo VAEP. Cada etapa foi projetada para avaliar o impacto incremental da introdução de informações contextuais espaciais e temporais na modelagem da probabilidade de sucesso de uma ação.

2.1. Fonte e Pré-processamento dos Dados

2.1.1. Fonte dos Dados

Utilizamos dados fornecidos pela **PFF (Pro Football Focus)** referentes à temporada 2023/2024 da **Premier League**, abrangendo informações espaço-temporais detalhadas

de partidas completas. O conjunto de dados é composto por três fontes principais:

- **Metadados de eventos:** descrevem cada ação registrada durante os jogos, incluindo o tipo (por exemplo, passe, chute, interceptação), o resultado (sucesso ou falha) e a localização no campo no momento da execução;
- **Dados de tracking:** fornecem as posições x, y de todos os jogadores e da bola em cada *frame* da partida, com frequência de 60 Hz, permitindo reconstruir com alta precisão os deslocamentos e interações em campo;
- **Informações táticas adicionais:** incluem indicadores binários de posse de bola, identificação do time atacante no momento da ação, direção de ataque e outras variáveis auxiliares para segmentar períodos ofensivos e defensivos.

Para viabilizar a integração com o modelo Atomic VAEP, realizamos a conversão dos dados de evento para o formato SPADL (*Soccer Player Action Description Language*), utilizando o módulo *socceraction*. Este formato padroniza a representação das ações em campo e facilita a aplicação de modelos de valoração. Em seguida, transformamos as ações em **Atomic SPADL**, que descreve as ações em unidades atômicas, que melhor se adequam à granularidade exigida pelo modelo.

Durante o processo de conversão, mantivemos apenas a primeira ocorrência de cada `event_id`, a fim de evitar duplicidades oriundas da sobreposição de eventos derivados (por exemplo, um passe que é também registrado como uma perda de posse). Essa filtragem garante a consistência e a não redundância do conjunto de ações analisadas.

Essa estrutura de dados combinada, enriquecida com coordenadas de tracking e variáveis contextuais táticas, serviu como base para a construção das features e treinamento do classificador de sucesso, conforme descrito nas próximas seções.

2.1.2. Pré-processamento

O processo de pré-processamento foi fundamental para garantir a integridade, padronização e alinhamento temporal dos diferentes conjuntos de dados utilizados no modelo. As transformações realizadas abrangeram dados de elenco, eventos (SPADL), metadados e tracking, conforme descrito a seguir.

Limpeza e Estruturação dos Dados de Elenco. O arquivo `rosters.csv`, contendo informações aninhadas sobre jogadores e equipes, foi convertido em um *DataFrame* estruturado com os seguintes passos:

- Leitura do arquivo bruto com *parsing* das colunas `player` e `team` para extração dos respectivos identificadores (`player_id` e `team_id`);
- Seleção e renomeação das colunas relevantes: `game_id`, `player_id`, `shirt`, `team_id`;
- Conversão de todos os campos para tipos numéricos apropriados, garantindo a consistência das chaves de junção;
- Salvamento do resultado final em formato *Parquet* como `rosters.parquet`, visando eficiência de leitura nos próximos estágios do pipeline.

Padronização Direcional dos Dados SPADL. Os dados de evento convertidos para o formato SPADL passaram por uma etapa de padronização direcional para garantir a coerência na orientação das ações. A metodologia consistiu em:

- Carregamento dos dados SPADL brutos e do mapa de times, contendo a informação `homeTeamStartLeft`;
- Aplicação da função `play_left_to_right`, que transforma as coordenadas das ações para que todas fluam da esquerda para a direita, independentemente do tempo ou lado de campo em que ocorreram;
- Validação visual da padronização por meio de gráficos “Antes vs. Depois”, assegurando a integridade espacial das ações;
- Armazenamento do resultado final como `spadl_ltr.parquet`.

Essa padronização é essencial para garantir consistência espacial no treinamento do classificador e na construção das *features* contextuais.

Verificação Visual da Posição dos Goleiros. Como parte da verificação da consistência dos dados de tracking, foi realizada uma inspeção visual da posição dos goleiros ao longo dos jogos. Esse procedimento teve como objetivos:

- Confirmar que os dados de tracking representavam corretamente a posição dos atletas ao longo da partida;
- Validar a lógica de identificação do goleiro em campo, aspecto crítico para a correta construção da *feature* de desvio angular.

As visualizações geradas corroboraram a integridade dos dados e reforçaram a confiabilidade do pipeline de tracking.

Processamento em Lote dos Dados de Tracking. Dada a estrutura por jogo dos arquivos de tracking e metadados, foi desenvolvido um pipeline de pré-processamento em lote com os seguintes passos para cada partida:

1. Leitura dos arquivos de metadados e filtragem de frames com `possession_id` válido;
2. Armazenamento do *DataFrame* de metadados limpo;
3. Utilização dos `frame_ids` resultantes para filtrar os dados de tracking dos jogadores, assegurando sincronização precisa entre os dois conjuntos;
4. Aplicação de rotinas de limpeza, incluindo tratamento de inconsistências e valores extremos na variável `speed`;
5. Armazenamento do *DataFrame* de tracking limpo em disco.

Esse processamento permitiu criar uma base de tracking estruturada e sincronizada, pronta para a extração eficiente de variáveis contextuais associadas a cada ação do modelo.

2.2. Etapa 1: Modelo Atomic VAEP Baseline

Nesta etapa inicial, foi implementado um modelo de referência (*baseline*) utilizando exclusivamente as *features* tradicionais extraídas do SPADL (Soccer Player Action Description Language). Essas features incluem:

- **Tipo de Ação:** codificação categórica de eventos como passe, chute, desarme, entre outros;

- **Localização da Ação:** coordenadas (x, y) normalizadas do início e fim da ação no campo;
- **Distância até o Gol:** métrica escalar da distância entre o ponto final da ação e o centro da baliza adversária;
- **Ângulo até o Gol:** ângulo formado entre a linha do ponto final da ação e as duas traves;
- **Contexto Imediato da Jogada:** informações básicas como posse de bola e sequência de ações do jogador.

A motivação para essa etapa é estabelecer uma linha de base justa e interpretável para avaliar o impacto das futuras incorporações de contexto posicional e tático. Por não utilizar dados de tracking, o modelo representa o estado da arte com base apenas em eventos, sendo ideal para comparação com métodos que integram dados espaço-temporais. Além disso, o desempenho deste baseline também permite quantificar o ganho marginal proporcionado por cada grupo de novas variáveis adicionadas nas etapas seguintes.

2.3. Etapa 2: Classificador de Sucesso com Features Táticas

Nesta etapa, desenvolvemos um classificador binário para estimar a probabilidade de sucesso de cada ação registrada, com base em variáveis contextuais extraídas de dados de tracking. O objetivo é substituir a simples frequência empírica de acertos por uma estimativa modelada do contexto tático real de cada jogada. O classificador foi treinado sobre rótulos de sucesso oriundos do SPADL, definidos com base no resultado direto da ação (por exemplo, passe completo, drible bem-sucedido, finalização no alvo).

As features utilizadas foram construídas a partir dos dados de tracking descritos na seção anterior e incluem: compactação defensiva, triangulação para o gol, velocidade do portador da bola e companheiros próximos, pressão e suporte locais, número de linhas de passe abertas, e desvio angular do goleiro. Essas variáveis fornecem uma visão contextual abrangente do entorno da ação, incluindo tanto a organização defensiva adversária quanto os recursos ofensivos disponíveis.

O modelo utilizado foi o `XGBoostClassifier`, devido à sua robustez para dados tabulares, capacidade de interpretar relações não lineares e tolerância a dados ruidosos ou parcialmente correlacionados. As previsões do classificador foram posteriormente utilizadas como uma nova feature contínua (`success_score`) no modelo VAEP final, conforme descrito na próxima seção.

2.3.1. Otimização de Hiperparâmetros

Para garantir a melhor configuração do classificador, foi aplicada uma busca aleatória de hiperparâmetros (*Randomized Search*) com validação cruzada estratificada por partidas (*GroupKFold*), a fim de evitar vazamento de informações entre os conjuntos de treino e teste. Utilizou-se como grupos os identificadores de partida (`game_id`) do conjunto de dados.

A busca explorou 20 combinações de hiperparâmetros, variando valores de `n_estimators`, `max_depth`, `learning_rate`, `subsample`, `colsample_bytree`, `reg_alpha`, `reg_lambda` e `scale_pos_weight`. A

métrica de desempenho utilizada foi a área sob a curva ROC (AUC). O melhor modelo encontrado por essa busca foi utilizado como classificador final, fornecendo a estimativa probabilística de sucesso para cada ação no conjunto de dados.

2.3.2. Compactação Defensiva

A feature de compactação defensiva quantifica o quão agrupada está a linha de defesa adversária no momento da posse de bola do time atacante. Para cada posse, identificamos os jogadores do time que estão defendendo (isto é, que não pertencem ao time atacante) e calculamos o desvio padrão das coordenadas x e y desses jogadores. Valores baixos de desvio indicam uma defesa mais compacta, enquanto valores altos representam uma defesa mais espaçada e vulnerável.

Motivação: uma linha defensiva compacta tende a dificultar a progressão de jogadas, restringindo espaços e linhas de passe. Por outro lado, uma defesa desorganizada ou espaçada geralmente permite maior liberdade para infiltrações, cruzamentos e finalizações. Portanto, essa feature capta de forma agregada o estado coletivo da organização defensiva.

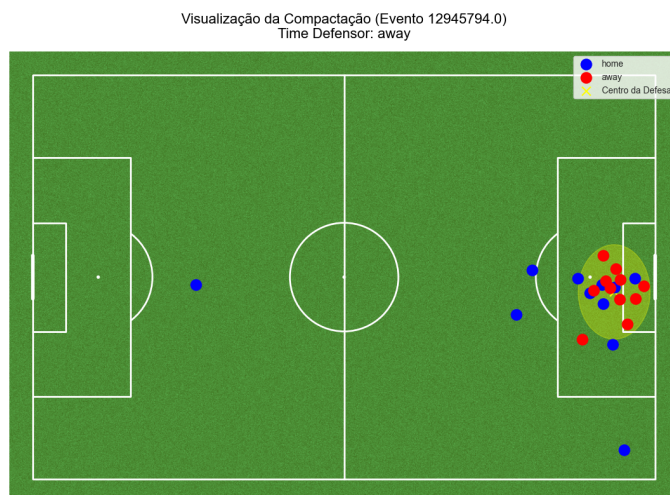


Figure 1. Exemplo Visual da Feature de Compactação Defensiva

2.3.3. Triângulo para o Gol

Essa feature busca medir o número de defensores que estão dentro do triângulo formado pelas traves do gol e o ponto da ação do jogador atacante. Para cada evento, construímos um polígono a partir das coordenadas das traves (ajustadas conforme o lado de ataque) e a posição do jogador com a bola. Em seguida, verificamos quantos jogadores do time adversário estão dentro desse triângulo (Figura 2).

Motivação: a presença de defensores dentro desse polígono reduz diretamente o campo de visão e o espaço útil entre o atacante e o gol, funcionando como um indicador da obstrução do caminho ao gol. Esse tipo de obstrução é particularmente relevante para prever finalizações, cruzamentos e ações decisivas em zonas perigosas.

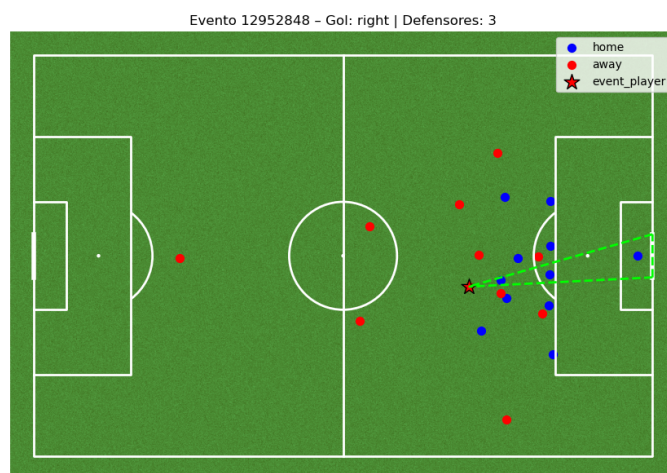


Figure 2. Exemplo Visual da Feature Triângulo para o Gol

2.3.4. Velocidade do Jogador e dos Companheiros

A feature de velocidade captura o estado dinâmico do jogador com a bola e de seus companheiros em um raio de 10 metros. Utilizando os dados de tracking sincronizados com cada posse, estimamos a velocidade instantânea do portador da bola e a média das velocidades de companheiros próximos.

Motivação: em situações de transição ofensiva ou contra-ataque, a velocidade é um indicador crítico da intensidade e do potencial de desorganização defensiva. Jogadas realizadas em alta velocidade têm maior probabilidade de romper linhas e gerar oportunidades antes que a defesa se reorganize. Além disso, jogadores em movimento próximo ao portador da bola aumentam o leque de opções táticas disponíveis (passes em profundidade, tabelas, arrastamento de marcação etc.).

2.3.5. Pressão, Suporte e Linhas de Passe

Esta *feature* (Figura 3) captura aspectos táticos do entorno do jogador com a bola. Para cada posse:

- Calculamos o número de defensores em raios de 3m e 6m (pressão imediata);
- Contamos companheiros de equipe em um raio de 8m (suporte ofensivo);
- Estimamos a pressão ponderada, levando em conta distância e o ângulo de abordagem dos defensores (maior peso para pressão frontal);
- Contamos o número de linhas de passe abertas, isto é, companheiros em posição avançada cuja rota de passe não esteja bloqueada por defensores.

Motivação: essas subfeatures fornecem uma leitura granular do contexto tático imediato. Alta pressão indica maior dificuldade de execução e maior risco de perda da posse, enquanto suporte e linhas de passe abertas aumentam as chances de sucesso da jogada. Juntas, essas medidas expressam o equilíbrio entre risco e oportunidade, que é central para a tomada de decisão no futebol.

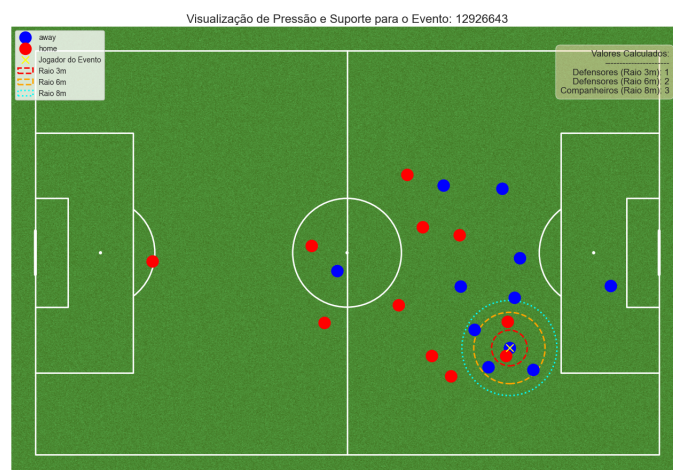


Figure 3. Exemplo Visual das Features de Pressão

2.3.6. Desvio Angular do Goleiro

Essa feature estima o quanto o goleiro está fora de posição no momento de uma finalização. Para cada evento com possibilidade de chute, comparamos o ângulo formado entre a posição do jogador e o centro do gol com o ângulo entre o jogador e a posição real do goleiro. A diferença angular (em graus) representa o desposicionamento do goleiro (Figura 4).

Motivação: um goleiro bem posicionado reduz significativamente o espaço útil para o chute. Já um goleiro ligeiramente fora de centro — seja por antecipação, deslocamento tardio ou desatenção — oferece brechas que aumentam a probabilidade de sucesso da finalização. Essa feature contribui diretamente para modelar o valor esperado de chutes em diferentes cenários, com base em pequenas variações posicionais do arqueiro.

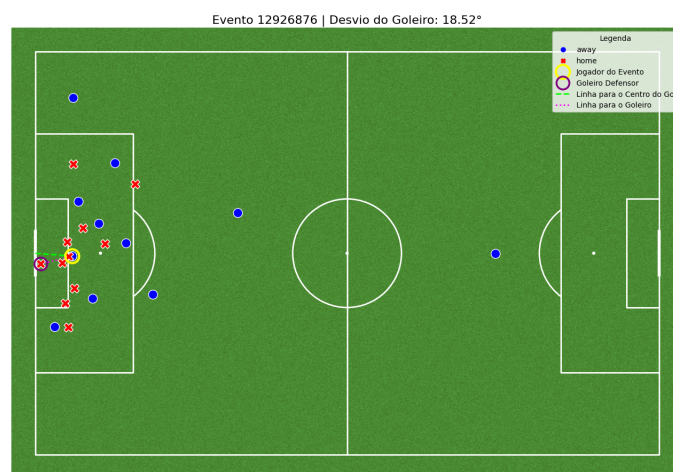


Figure 4. Exemplo Visual da Feature Desvio Angular do Goleiro

2.4. Etapa 3: Modelo VAEP com `success_score`

A probabilidade predita pelo classificador da Etapa 2 (`success_score`) foi então incorporada como uma feature adicional no modelo VAEP, que agora é treinado com o

algoritmo XGBoost e utiliza todas as features do modelo baseline (Etapa 1) mais a `success_score`. Em outras palavras, o novo modelo de valor de ação se beneficia tanto das informações estruturais dos eventos SPADL quanto da inferência tática derivada dos dados de tracking, sintetizada em forma de probabilidade predita.

Motivação: a inclusão da `success_score` permite ao modelo VAEP capturar, de forma indireta, nuances espaciais e comportamentais que influenciam a chance de êxito de uma ação, mas que não estão diretamente representadas nas features tradicionais. Como o classificador da Etapa 2 é alimentado com variáveis de contexto tático (como pressão, compactação defensiva, velocidade do setor etc.), a `success_score` atua como um vetor sintético dessas condições contextuais.

Vantagem metodológica: esse acoplamento modular entre modelos permite separar a modelagem do contexto tático (Etapa 2) da estimativa de valor das ações (Etapa 3), resultando em uma abordagem mais interpretável e extensível. Em vez de inflar o modelo VAEP com dezenas de features táticas adicionais, a complexidade contextual é absorvida e traduzida em uma única variável probabilística.

Considerações técnicas: o modelo VAEP com `success_score` foi treinado e validado utilizando o mesmo protocolo da Etapa 1 (validação estratificada por partidas). A performance foi avaliada com as mesmas métricas — Brier Score, AUC e correlação com expected goals — permitindo uma comparação direta com o modelo baseline. O impacto da `success_score` também foi analisado via *feature importance*, confirmando sua contribuição significativa para a melhora do modelo final.

Otimização hiperparamétrica: para o modelo VAEP com `success_score`, realizamos também uma *Randomized Search* para seleção dos melhores hiperparâmetros do algoritmo XGBoost, incorporando as mesmas variáveis táticas da Etapa 2. A busca foi conduzida com *GroupKFold* para garantir validação por partidas, com foco na métrica de AUC. Esta otimização foi aplicada apenas ao modelo com variáveis de tracking, e não ao baseline, de forma a preservar a comparabilidade e isolar os efeitos das novas variáveis táticas sobre o desempenho preditivo..

Ambos os modelos – baseline e enriquecido – foram avaliados com base em três métricas principais:

- **Brier Score;**
- **Log Loss;**
- **ROC AUC.**

A escolha das métricas e a estrutura de validação cruzada adotadas seguem as boas práticas recomendadas por trabalhos recentes na área de sports analytics [Fernández et al. 2024].

As métricas foram computadas separadamente para a probabilidade de marcar (*scores*) e de sofrer gols (*concedes*). O impacto da integração do `success_score` foi quantificado comparando os dois modelos com os mesmos dados de teste.

3. Análise dos resultados

Os gráficos de importância apresentados nas próximas subseções foram gerados com base na métrica de *total_gain* do XGBoost, que quantifica o ganho de informação acumulado

por cada variável nas divisões das árvores. Essa medida permite comparar a relevância relativa das features dentro de um mesmo modelo.

3.1. Efeito das Features do Atomic Vaep Baseline

O modelo baseline demonstra forte dependência de features espaciais fundamentais, com `dist_to_goal_a0` dominando claramente a importância relativa. As coordenadas de posição (`x_a0`, `y_a0`) e o tipo de ação (`actiontype_shot_a0`) também apresentam contribuições significativas. Essa distribuição (Figura 5) confirma que o modelo tradicional se apoia principalmente em métricas geométricas diretas e na classificação básica das ações, sem capturar nuances táticas do contexto de jogo.

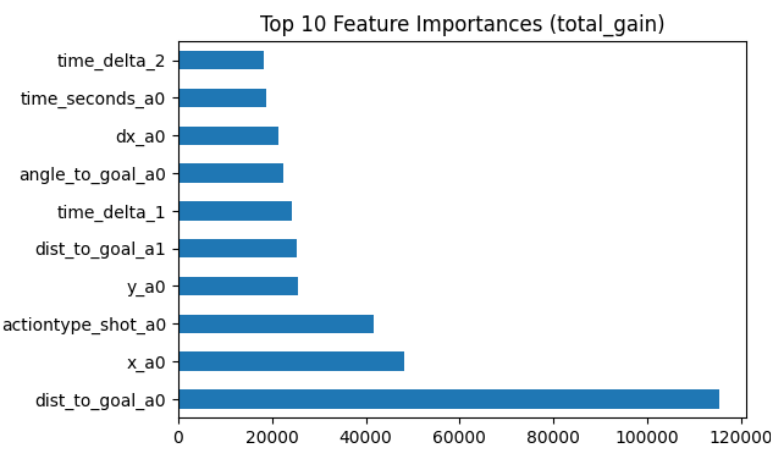


Figure 5. Importância das features no modelo VAEP baseline

3.2. Efeito das Features do Classificador

O classificador de sucesso revela uma distribuição (Figura 6) mais equilibrada entre as features, com destaque para variáveis de movimento (`dx_a0`, `dx_a1`) e métricas táticas como `pressao_posicional_ponderada`. A presença de `defensive_compaction_x` entre as principais features valida a relevância das variáveis contextuais derivadas do tracking. Notavelmente, o modelo valoriza tanto aspectos dinâmicos quanto posicionais, sugerindo que o sucesso de uma ação depende fortemente do contexto tático instantâneo.

3.3. Efeito das Features do Atomic Vaep Final

A incorporação do `success_score` como feature principal no modelo final demonstra o impacto substancial da integração do classificador tático. Com importância total superior a 25.000, essa variável sintética captura efetivamente o contexto espacial e comportamental modelado na Etapa 2. As features tradicionais como `actiontype_shot_a0` e `dist_to_goal_a0` mantêm relevância, mas agora complementam a informação contextual agregada, resultando em um modelo mais robusto e contextualmente informado (Figura 7).

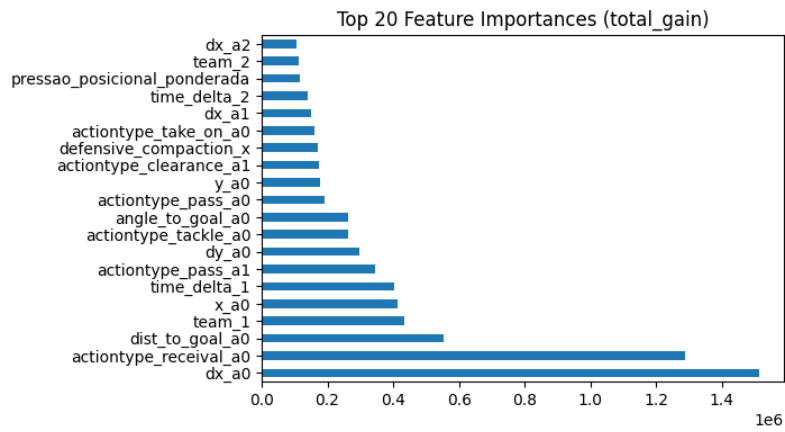


Figure 6. Importância das features no classificador de sucesso

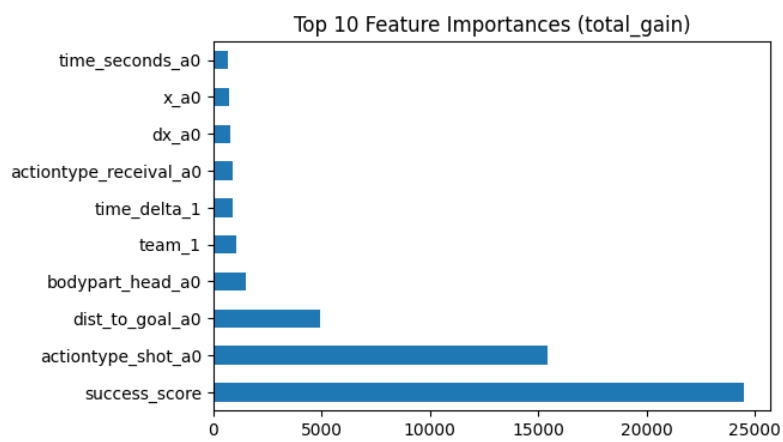


Figure 7. Importância das features no modelo VAEP com success_score

4. Conclusão

Ao incorporar um classificador de sucesso alimentado por *features* táticas extraídas de dados de tracking da Premier League 2023–2024, conseguimos enriquecer o contexto das ações individuais realizadas pelos jogadores.

Os resultados (Tabela 1) demonstram ganhos pontuais nas principais métricas de avaliação. Na de previsão de *scores*, o modelo aprimorado apresentou melhora em todas as métricas: o *Brier Score* reduziu de 0,01008 para 0,00954, o *Log Loss* de 0,05153 para 0,04847, e a área sob a curva ROC (AUC) aumentou de 0,79694 para 0,81630. Da mesma forma, para a tarefa de *concedes*, observou-se melhora em todas as métricas, com *Brier Score* de 0,00176 para 0,00175, *Log Loss* de 0,01166 para 0,01153 e AUC de 0,81841 para 0,82452.

Table 1. Comparação das métricas entre os modelos A-VAEP e A-VAEP com `success_score`

Métrica	Baseline A-VAEP	A-VAEP + <code>success_score</code>
<i>Scores</i>		
Brier Score	0.01008	0.00954
Log Loss	0.05153	0.04847
ROC AUC	0.79694	0.81630
<i>Concedes</i>		
Brier Score	0.00176	0.00175
Log Loss	0.01166	0.01153
ROC AUC	0.81841	0.82452

Esses avanços indicam que a adição de informações posicionais e táticas em tempo real pode aumentar a capacidade do modelo de estimar o valor real de uma ação no futebol. Além de melhorar o desempenho preditivo, o modelo proposto fornece interpretações mais alinhadas com a análise qualitativa realizada por analistas táticos, departamentos de desempenho e equipes técnicas. Assim, este estudo representa um passo importante em direção a sistemas analíticos mais completos, contextualmente ricos e aplicáveis ao ambiente competitivo do futebol profissional.

References

Decroos, T., Bransen, L., Haaren, J. V., and Davis, J. (2019). Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1851–1861.

Fernández, J., Cervone, D., and Bornn, L. (2024). Methodology and evaluation in sports analytics: challenges, approaches and lessons learned. *ResearchGate Preprint*.

Análise e Predição de Química entre Jogadores de Futebol com Base em Dados de Evento

Gabriel Castelo, Matheus Vaz, Victor Kenji, Vinicius Gomes

Universidade Federal de Minas Gerais, Belo Horizonte, Brasil
{gabriel-castelo, vazleal, victorkpa, vinisilvag}@ufmg.br

Resumo A química entre jogadores de futebol, frequentemente ignorada nas análises convencionais de desempenho, exerce papel fundamental na eficácia coletiva das equipes. Este trabalho investiga essa dimensão coletiva por meio da replicação e extensão de métricas propostas no artigo *Player Chemistry: Striving for a Perfectly Balanced Soccer Team*, utilizando dados públicos de eventos das principais ligas europeias da temporada 2017/2018. Foram desenvolvidas as métricas *Joint Offensive Impact* (JOI) e *Joint Defensive Impact* (JDI), que quantificam, respectivamente, a sinergia ofensiva e a efetividade defensiva de duplas de jogadores, a partir da granularidade do modelo Atomic SPADL e valores calculados via VAEP. Apesar das limitações de hardware, propôs-se a estruturação de um modelo preditivo para estimar a química entre jogadores que nunca atuaram juntos, abrindo caminho para aplicações em processos de contratação. Os resultados mostram que é possível calcular de forma robusta métricas de entrosamento com dados públicos, oferecendo uma alternativa viável e acessível para clubes com menor estrutura analítica.

Palavras-chave: futebol, entrosamento, química entre jogadores, dados de evento, Atomic VAEP, análise esportiva

1 Introdução

No futebol moderno, a avaliação de desempenho ultrapassa os limites da análise estatística individual de cada jogador. Embora métricas como número de passes certos, gols marcados ou desarmes efetuados sejam fundamentais, elas não capturam totalmente a complexidade do jogo coletivo. Um aspecto muitas vezes negligenciado, mas de extrema relevância, é a interação entre os atletas – o que denominamos *química*. Essa conexão e tática entre os jogadores pode influenciar significativamente o rendimento da equipe como um todo. A forma como atletas se comunicam, se posicionam e se antecipam às jogadas uns dos outros é fruto de um entrosamento que vai além das habilidades técnicas individuais.

Apesar de sua importância, a dimensão coletiva do jogo ainda é frequentemente negligenciada em processos de recrutamento e formação de elencos. Um exemplo marcante dessa falha ocorreu na janela de transferências de inverno da temporada 2010/2011, quando o Liverpool contratou simultaneamente Andy

Carroll e Luis Suárez. Carroll, um centroavante de referência com forte presença aérea, teve dificuldades de adaptação ao estilo da equipe, resultando em apenas 6 gols marcados em 44 partidas. Em contraste, Suárez – um atacante mais móvel e com características mais alinhadas ao perfil do time – alcançou 69 gols em 110 jogos. Esse contraste evidencia como a falta de química entre jogador e equipe pode comprometer o desempenho, mesmo quando se trata de atletas tecnicamente qualificados.

Nesse sentido, este trabalho propõe investigar os aspectos que influenciam a qualidade do entrosamento entre jogadores de futebol, buscando compreender em que medida atletas conseguem se complementar em campo, através da replicação da metodologia adotada no artigo "*Player chemistry: Striving for a perfectly balanced soccer team*" [1]. O foco está na análise de duplas de jogadores que atuaram juntos, com o objetivo de classificar o nível de química observado entre eles a partir de dados de desempenho coletivo e individual. A partir dessa classificação, queremos também treinar um modelo preditivo capaz de estimar o grau de entrosamento potencial entre jogadores que nunca atuaram juntos.

2 Motivação e Justificativa

A crescente utilização de dados no futebol tem transformado as decisões tomadas pelos grandes times, tanto intra quanto extra campo. Seja em termos de performance, contratações ou táticas de jogo, um grande número de equipes possui um setor voltado para a geração de estatísticas úteis. No entanto, a maioria das análises está centrada em indicadores e jogadores de forma individual, negligenciando aspectos coletivos, como a sinergia entre os atletas.

Alguns estudos, faça vista [1], porém, têm buscado uma abordagem que se estende para além da análise individual, incorporando perspectivas coletivas ao considerar a efetividade de grupos de jogadores na construção de jogadas de impacto. Essas investigações evidenciam um movimento crescente dentro da comunidade científica e da indústria esportiva em direção à valorização de aspectos coletivos do desempenho, como o entrosamento, a coordenação tática e a complementaridade de funções em campo.

Esse interesse tem se intensificado especialmente com a crescente profissionalização e estruturação dos departamentos de *sport analytics* nos clubes de futebol, que agora contam com equipes multidisciplinares e acesso a dados cada vez mais granulares. Nesse cenário, torna-se cada vez mais relevante e pertinente estudar métricas que capturem a dinâmica coletiva das equipes, não apenas para a avaliação de desempenho durante as competições, mas também como critério estratégico na contratação de novos jogadores e na montagem de elencos mais coesos e eficientes. A tendência é que análises voltadas à dimensão coletiva do jogo ganhem espaço e se consolidem como parte fundamental da tomada de decisão no futebol profissional.

Nesse contexto, o principal problema investigado neste trabalho é: quão bem um time de futebol consegue se entrosar? Mais especificamente, busca-se desenvolver uma forma de quantificar esse entrosamento – ou química – e analisá-lo de

maneira sistemática, tanto no aspecto ofensivo quanto defensivo, com base em dados históricos de partidas. O objetivo é construir métricas que permitam avaliar o nível de integração entre os jogadores, indo além da performance individual para capturar elementos do jogo coletivo.

Além disso, este trabalho propõe averiguar a viabilidade de treinar um modelo preditivo capaz de estimar o entrosamento entre pares de jogadores que nunca atuaram juntos. A ideia é avaliar se, à partir de dados prévios sobre estilo de jogo, características físicas e histórico de desempenho, é possível antecipar a compatibilidade entre dois atletas. Caso bem-sucedido, esse modelo poderia ser utilizado como ferramenta auxiliar em processos de montagem de elencos, oferecendo previsões sobre o potencial de química entre jogadores antes mesmo de sua atuação conjunta em campo.

3 Dados utilizados

Para o cálculo das métricas e o treinamento dos modelos de química ofensiva e defensiva, foram utilizados dados referentes aos campeonatos: Francês, Inglês, Alemão e Italiano, nos anos 2017/2018. Os dados foram obtidos através do dataset Público "*A public dataset of spatio-temporal match events in soccer competitions* [5].

Os conjuntos de dados utilizados descrevem, em nível detalhado, as ações e eventos realizados pelos jogadores com a bola ao longo das partidas – os chamados dados de evento. Informações adicionais sobre os atletas, como posição, nacionalidade, idade e altura, também estarão disponíveis. Quando necessário, os dados serão simplificados para facilitar a modelagem, além de serem enriquecidos com variáveis complementares, como perfis de estilo de jogo e indicadores de desempenho físico, a fim de capturar com maior precisão os fatores que influenciam o entrosamento entre jogadores.

Além disso, visando enriquecer as informações disponíveis, foi realizada uma operação de *Web Scrapping* para extrair dados do site *FBref* [4], de modo a integrar estatísticas agregadas sobre o desempenho dos atletas individualmente ao longo da temporada analisada.

4 Metodologia

A execução deste trabalho foi dividida em três etapas basilares: A preparação e o pré processamento dos dados, a programação e cálculo das métricas JOI (*Joint Offensive Impact*) e JDI (*Joint Defensive Impact*) e a criação de um modelo de aprendizado de máquina que fosse capaz de prever ditas métricas para um par de jogadores que nunca antes atuaram em conjunto.

4.1 Preparação dos dados

Para o cálculo das métricas de impacto conjunto, são necessários dados do tipo *dados de eventos*. Conforme explicitado na sessão 3, dados de evento descrevem

detalhadamente as ações realizadas em campo pelos jogadores. Esses dados foram baixados diretamente do dataset [5], e convertidos para o formato *Atomic SPADL (Soccer Player Action Description Language)* [2] [7], tido como referência no que tange ao uso de dados de evento no contexto futebolístico. A variante "*Atomic*" quebra as ações normalmente tratadas pelo modelo SPADL em ações granularmente mais finas, permitindo mais precisão para qualquer modelo estatístico que utilizar os dados.

Para trazer informações adicionais sobre os jogadores, buscando enriquecer a capacidade preditiva do modelo postulado na sessão 4.4, foi realizada uma operação de *Web Scrapping* do site *FBRef* [4], onde foram extraídas, para cada um dos campeonatos supracitados, as seguintes tabelas

- *Player Passing*: Traz informações referentes aos passes realizados por um jogador ao longo da temporada.
- *Player Pass Types*: Agrega dados sobre os diferentes tipos de passes realizados por um jogador ao longo de uma temporada, *e.g.*: Escanteios, laterais, faltas.
- *Player Defense*: Traz estatísticas relativas às ações defensivas tomadas por um jogador ao longo da temporada
- *Player Miscellaneous Stats*: Apresenta informações diversas, que não puderam ser categorizadas em nenhuma das tabelas anteriores, mas são relevantes para o estudo em questão, tais como número de cartões amarelos e vermelhos, cruzamentos, interceptações, pênaltis bem sucedidos, duelos bem sucedidos, etc.

Os dados extraídos foram, então, conciliados com as tabelas obtidas do dataset público de onde os dados de evento foram extraídos, enriquecendo as informações acerca dos jogadores.

4.2 Adoção do Atomic SPADL e a Melhoria com o Atomic VAEP

Uma decisão metodológica fundamental neste trabalho foi a utilização do formato **Atomic SPADL** para representar os dados de evento. Essa escolha, conforme descrito na seção de preparação de dados, impacta diretamente a forma como o valor de cada ação é calculado e, conseqüentemente, a precisão das métricas de química (JOI e JDI). A utilização deste formato leva ao emprego do **atomic VAEP** em vez do VAEP tradicional.

A principal vantagem do modelo atômico reside na sua granularidade. O modelo SPADL padrão trata ações complexas, como um passe longo ou um drible, como eventos únicos e monolíticos. Essa abordagem pode simplificar excessivamente a análise, pois não diferencia as várias fases que compõem a ação. Em contrapartida, a variante "*Atomic*" quebra as ações [...] em ações granularmente mais finas". Por exemplo, uma única ação de "passe" no modelo padrão é decomposta em até três micro-ações no Atomic SPADL:

- O recebimento da bola antes do passe.

- O carregamento ou condução da bola.
- A execução do passe em si.

Ao calcular um valor de VAEP para cada uma dessas sub-ações (o *atomic VAEP*), o modelo consegue atribuir valor de forma muito mais precisa. Ele pode, por exemplo, reconhecer uma condução de bola valiosa que quebra linhas defensivas antes de um passe simples, algo que o modelo padrão não capturaria.

Essa abordagem "permite mais precisão para qualquer modelo estatístico que utilizar os dados". Para as métricas de JOI e JDI, essa precisão é crucial. Como o JOI é calculado a partir da soma dos valores de ações consecutivas, a maior fidelidade do *atomic VAEP* resulta em uma medida de sinergia ofensiva mais robusta e representativa da realidade do jogo. Da mesma forma, o JDI se beneficia ao avaliar o impacto dos adversários com uma métrica de valor mais apurada. Portanto, a adoção do *atomic VAEP* não é apenas uma escolha técnica, mas uma melhoria metodológica que eleva a qualidade e a validade das métricas de química propostas.

4.3 Cálculo de métricas

Conforme a metodologia proposta no artigo de referência [1] e os objetivos deste trabalho, foram implementadas duas métricas centrais para quantificar a química entre duplas de jogadores: o Impacto Ofensivo Conjunto (JOI) e o Impacto Defensivo Conjunto (JDI). Ambas as métricas foram normalizadas por 90 minutos de jogo para permitir comparações justas.

JOI (Joint Offensive Impact) O Impacto Ofensivo Conjunto (JOI) quantifica a sinergia ofensiva entre uma dupla de jogadores, medindo o valor gerado por suas interações diretas e consecutivas durante uma partida. A implementação desta função segue uma abordagem que valoriza a continuidade e a qualidade das jogadas.

O processo de cálculo para uma partida envolve:

1. **Identificação de Interações Ofensivas:** O algoritmo analisa o fluxo de eventos da partida para encontrar todas as sequências onde um jogador p executa uma ação ofensiva (como passe, cruzamento, drible) e, imediatamente a seguir, o jogador q realiza a próxima ação registrada.
2. **Avaliação da Interação:** Cada interação identificada é avaliada pela soma dos valores *atomic VAEP* das duas ações consecutivas. Esta soma, denominada `extended_vaep`, representa o valor total agregado pela jogada da dupla.
3. **Cálculo do JOI Total:** O JOI da dupla para a partida é o somatório dos valores de todas as interações, considerando tanto as sequências de $p \rightarrow q$ quanto as de $q \rightarrow p$, refletindo a reciprocidade da parceria.

A fórmula matemática para o JOI de uma dupla (p, q) em uma partida g é:

$$JOI(p, q, g) = \sum_{i \in I_{p \rightarrow q, g}} (V_i + V_{i+1}) + \sum_{j \in I_{q \rightarrow p, g}} (V_j + V_{j+1}) \quad (1)$$

onde $I_{p \rightarrow q, g}$ representa o conjunto de interações de p para q na partida g , e V_k é o valor *atomic VAEP* da ação k .

JDI (Joint Defensive Impact) O Impacto Defensivo Conjunto (JDI) avalia a eficácia de uma dupla de jogadores em suprimir o desempenho ofensivo dos adversários. Diferentemente do JOI, esta métrica não foca nas ações diretas dos defensores, mas no seu impacto indireto sobre os oponentes, medindo o quanto a presença da dupla contribuiu para reduzir o valor ofensivo gerado por eles em comparação com seu desempenho esperado. O cálculo está implementado na função `joint_defensive_impact`.

Os passos para o cálculo em uma partida são:

1. **Análise do Desempenho do Oponente:** Para cada jogador oponente o , calcula-se o seu Impacto Ofensivo Real (AOI) na partida, que é a soma dos valores *atomic VAEP* de suas ações ofensivas. Em paralelo, calcula-se seu Impacto Ofensivo Esperado (EOI), que corresponde à sua média histórica de performance ofensiva por 90 minutos em jogos anteriores.
2. **Cálculo da Redução de Impacto:** A contribuição defensiva da equipe contra o oponente o é a diferença $\Delta OI = EOI(o) - AOI(o)$. Um valor positivo indica que o oponente teve um desempenho abaixo do seu padrão.
3. **Atribuição de Responsabilidade Espacial:** O mérito por essa redução de impacto é distribuído à dupla defensiva (p, q) por meio de uma função de responsabilidade compartilhada. Esta função atribui uma cota de responsabilidade com base na proximidade posicional da dupla em relação ao oponente, utilizando a distância euclidiana inversa. Quanto mais próximos os defensores estão do atacante, maior sua responsabilidade.
4. **Cálculo do JDI Total:** O JDI da dupla na partida é o somatório das reduções de impacto (ΔOI) de todos os oponentes, ponderado pela cota de responsabilidade da dupla sobre cada um deles.

A fórmula do JDI para a dupla (p, q) na partida g é:

$$JDI(p, q, g) = \sum_{o \in O_g} (EOI(o, g) - AOI(o, g)) \times R(p, q, o) \quad (2)$$

onde O_g é o conjunto de oponentes na partida g , e $R(p, q, o)$ é a parcela de responsabilidade da dupla sobre o oponente o .

4.4 Criação do modelo

Visando seguir os moldes propostos por [1], este trabalho almejou a criação de um modelo de *Machine Learning* que fosse capaz de prever as métricas JOI e JDI, detalhadas nas seções 4.3 e 4.3. No artigo citado, foram treinados dois modelos de regressão separados, utilizando o framework *CatBoost* [3] [6]. Na teoria proposta, os dados foram estruturados da seguinte forma:

- Representação dos jogadores
Cada jogador foi representado como um vetor de características individuais, tanto de caráter pessoal, como idade, altura, peso, nacionalidade, região, língua materna e posição, quanto de caráter qualitativo de desempenho, como atributos físicos (velocidade, capacidade em duelos aéreos e terrestres) e táticos.
- Pares de Jogadores
Os pares de jogadores foram estruturados considerando indicadores culturais comuns e quantidade de partidas históricas disputadas.

5 Resultados

5.1 JOI_{90} e JDI_{90}

As métricas finais para avaliação da química entre jogadores são o JOI_{90} e o JDI_{90} , que correspondem às normalizações dos valores de JOI e JDI por 90 minutos de jogo. Essa normalização, implementada nas funções `calculate_joi90` e `calculate_jdi90`, é fundamental para permitir uma comparação justa entre duplas que atuaram por durações de tempo distintas.

O cálculo do JOI_{90} para uma dupla (p, q) é dado por:

$$JOI_{90}(p, q) = \frac{\sum_{g \in G_{p,q}} JOI(p, q, g)}{\sum_{g \in G_{p,q}} M_{min}(p, q, g)} \times 90 \quad (3)$$

Analogamente, o JDI_{90} é calculado como:

$$JDI_{90}(p, q) = \frac{\sum_{g \in G_{p,q}} JDI(p, q, g)}{\sum_{g \in G_{p,q}} M_{min}(p, q, g)} \times 90 \quad (4)$$

Onde:

- $G_{p,q}$ é o conjunto de todos os jogos em que a dupla (p, q) atuou junta.
- $JOI(p, q, g)$ e $JDI(p, q, g)$ são os impactos da dupla na partida g .
- $M_{min}(p, q, g)$ representa o mínimo de minutos jogados entre p e q na partida g .

5.2 Modelo Preditivo

Apesar da metodologia clara e estruturada para a criação do modelo preditor, a conclusão da tarefa não foi possível. Tendo em vista a presença de métricas cuja caracterização não foi dissecada no *paper* original, e a falta de *hardware* adequado para treinamento do modelo, não houve possibilidade de realizar a etapa de treinamento e validação.

6 Estudos de Caso: Análise de Equipe

Com o objetivo de demonstrar a aplicação prática das métricas JOI90 e JDI90, elaboramos estudos de caso com equipes reconhecidas do cenário europeu. Esses exemplos visam ilustrar como os indicadores propostos podem auxiliar na identificação de duplas com alto (ou baixo) grau de entrosamento, tanto em contextos ofensivos quanto defensivos.

6.1 Impacto Ofensivo

Na Figura 1, analisamos o setor ofensivo do Manchester City, formado por Sergio Agüero, Leroy Sané, Raheem Sterling e Bernardo Silva. É possível notar uma forte sinergia entre Sterling e Agüero (2.76), bem como entre Sterling e Bernardo Silva (1.512), sugerindo que essas combinações foram especialmente eficazes na geração de ações ofensivas consecutivas de alto valor. Em contraste, a interação entre Sané e Bernardo Silva apresenta um JOI90 baixo (0.211), indicando menor fluidez nas jogadas entre esses atletas.

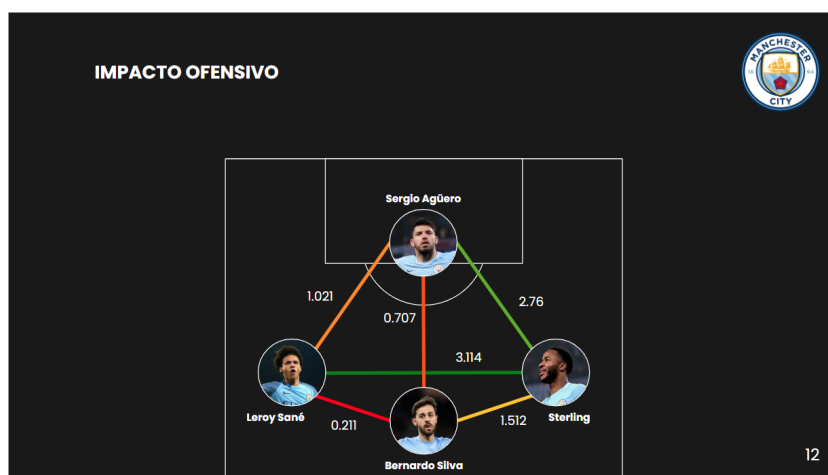


Figura 1. JOI90 entre atletas ofensivos do Manchester City.

A Figura 2 apresenta a análise ofensiva do Barcelona, com Messi, Suárez, Dembélé e Coutinho. A conexão entre Messi e Suárez se destaca com um JOI90 de 9.833, evidenciando uma química ofensiva muito superior às demais combinações do time. As demais interações — como Messi e Dembélé (2.806) ou Suárez e Coutinho (1.769) — são significativamente menos impactantes, reforçando o papel central da dupla Messi-Suárez na dinâmica ofensiva do time.



Figura 2. JOI90 entre atletas ofensivos do Barcelona.

6.2 Impacto Defensivo

A Figura 3 apresenta a análise defensiva do Manchester City. Observa-se que a dupla John Stones e Kompany obteve um JDI90 negativo (-2.300), indicando que os adversários, quando enfrentaram esses dois defensores, conseguiram desempenho acima do esperado. Por outro lado, a interação entre Fabian Delph e Stones apresentou um JDI90 de 0.894, apontando um desempenho defensivo coletivo mais efetivo.

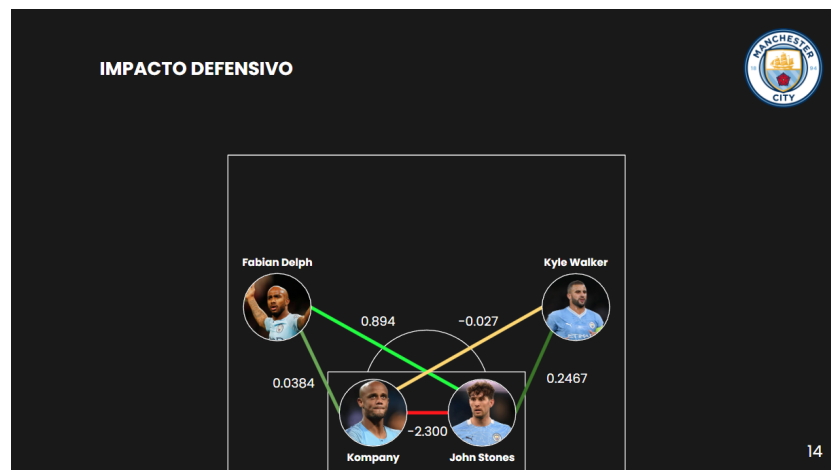


Figura 3. JDI90 entre defensores do Manchester City.

Por fim, a Figura 4 mostra a análise defensiva do Liverpool, com destaque para a dupla Van Dijk e Lovren, que obteve um JDI_{90} de 0.802, indicando forte sinergia defensiva. Os laterais Robertson e Alexander-Arnold também apresentaram bons valores nas combinações com os zagueiros centrais, reforçando a consistência do sistema defensivo da equipe.

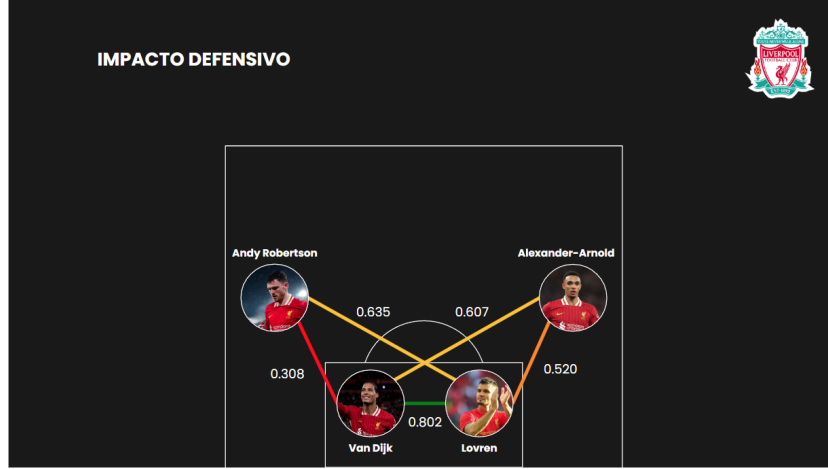


Figura 4. JDI_{90} entre defensores do Liverpool.

7 Conclusão

O estabelecimento de estatísticas e modelos capazes de mensurar a capacidade de interação entre jogadores, e o entendimento da dinâmica da denominada "química" de um time é um campo com grande potencial de impacto na dinâmica de contratações e gerência no cenário futebolístico moderno.

Neste trabalho, foi demonstrado ser possível obter resultados promissores das métricas JOI_{90} e JDI_{90} com a utilização de dados públicos, efetivamente calculando de maneira coerente ambos os indicadores. Este fato abre espaço potencial para utilização de uma metodologia de ponta para avaliação de jogadores por parte de clubes menores, apesar das limitações impostas por dados proprietários.

7.1 Trabalhos Futuros

Visando a continuidade do trabalho desenvolvido, postula-se que seja executado o desenvolvimento do modelo preditor de métricas de química, acompanhado do cálculo e/ou substituição das métricas faltantes. Ademais, a integração de dados de *tracking* para melhorar o cálculo das métricas, em especial o JDI , assim como a extensão destas para grupos de n jogadores, e não apenas 2, também são considerados trabalhos relevantes e caminhos naturais a serem seguidos.

Referências

1. Bransen, L., Haaren, J.V.: Player chemistry: Striving for a perfectly balanced soccer team. CoRR **abs/2003.01712** (2020), <https://arxiv.org/abs/2003.01712>
2. Decroos, T., Bransen, L., Van Haaren, J., Davis, J.: 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. pp. 1851–1861. KDD ’19, ACM, New York, NY, USA (2019). <https://doi.org/10.1145/3292500.3330758>, <http://doi.acm.org/10.1145/3292500.3330758>
3. Dorogush, A.V., Ershov, V., Gulin, A.: Catboost: Gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363 (2018), <https://arxiv.org/abs/1810.11363>
4. FBref.com: 2017–2018 premier league stats – passing. Página de estatísticas online (2025), <https://fbref.com/en/comps/9/2017-2018/passing/2017-2018-Premier-League-Stats>, acesso em: 5 jul. 2025
5. Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., Giannotti, F.: A public data set of spatio-temporal match events in soccer competitions. Scientific Data **6**(236) (Oct 2019). <https://doi.org/10.1038/s41597-019-0247-7>, <https://www.nature.com/articles/s41597-019-0247-7>
6. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: Catboost: Unbiased boosting with categorical features. In: Advances in Neural Information Processing Systems (NeurIPS) Workshop on Advances in Approximate Bayesian Inference (2017), <https://arxiv.org/abs/1706.09516>
7. Van Roy, M., Robberechts, P., Decroos, T., Davis, J.: Valuing on-the-ball actions in soccer: A critical comparison of xt and vaep. In: Proceedings of the AAAI-20 Workshop on Artificial Intelligence in Team Sports. AITS, AI in Team Sports Organising Committee (dec 2020)

Correlação entre Dados de Evento e Tracking com Lesões no Futebol Profissional

Júlio Guerra Domingues¹, Leandro Luiz Duarte Teixeira¹,
Victor Yuji Yano¹, Rodrigo Luiz Macedo Ferreira¹, Denilson Martins Vieira Santos¹

¹ Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, MG, Brasil

{juliogd, lteixeira, victory, rodrigoluizmf, denilsonmvs}@ufmg.br

Resumo

Lesões no futebol profissional são um problema crescente e de alto impacto financeiro e esportivo. Este trabalho investigou os fatores de risco para lesões musculares por meio da análise de dados de evento e de tracking. A metodologia baseou-se na integração de múltiplas fontes, incluindo históricos de lesão do Transfermarkt (validados por um médico), dados de jogos da Pro Football Focus e um dataset detalhado com dados de treino e jogos de um time brasileiro, no qual a ocorrência de lesão foi inferida a partir de ausências prolongadas. Os resultados indicam que, além de fatores como idade e histórico prévio, picos de carga de trabalho são preditores significativos. Métricas como a Carga Total de Aceleração, a Distância Total percorrida e os esforços de desaceleração foram significativamente maiores no grupo de jogadores lesionados no período pré-lesão. A análise evidencia que a inclusão de dados de treino é fundamental para a identificação de riscos, fornecendo uma base quantitativa para o desenvolvimento de programas de prevenção mais eficazes.

Abstract

Injuries in professional football are a growing problem with high financial and sporting impact. This study investigated the risk factors for muscle injuries by analyzing event and tracking data. The methodology was based on the integration of multiple sources, including injury histories from Transfermarkt, game data from PFF, and a detailed dataset with training and game data from a Brazilian team, in which injury occurrence was inferred from prolonged absences. The results indicate that, in addition to factors such as age and previous history, workload spikes are significant predictors. Metrics like Total Acceleration Load, Total Distance covered, and deceleration efforts were significantly higher in the group of injured players in the pre-injury period. The analysis highlights that including training data is crucial for risk identification, providing a quantitative basis for developing more effective prevention programs.

1. Introdução

Lesões são eventos de impacto significativo no futebol profissional, afetando negativamente a carreira dos jogadores[1], o desempenho das equipes[2] e as finanças dos clubes[3, 4]. Estudos apontam que lesões relacionadas ao futebol são uma das principais causas para o encerramento da carreira de atletas profissionais[1] e que sua frequência

e custo vêm aumentando nas últimas temporadas[4]. Diante deste cenário, entender os fatores de risco associados às lesões é um passo fundamental para desenvolver estratégias de prevenção eficazes.

O objetivo deste trabalho é investigar a relação entre métricas físicas e de desempenho dos atletas e a ocorrência de lesões musculares, avaliando como os dados disponíveis podem auxiliar na prevenção por meio de análise estatística e identificação de padrões.

2. Métodos

2.1. Desenho e amostra

A abordagem metodológica deste estudo baseou-se na integração de múltiplas fontes de dados para a extração de características relevantes e posterior análise estatística comparativa entre jogadores lesionados e não lesionados. O processo iniciou-se com o levantamento do histórico dos atletas a partir do Transfermarkt, que serviu como fonte primária para dados demográficos (idade, altura) e registros de lesões (datas e dias de afastamento). Cada lesão única registrada foi avaliada por um médico especialista, que as classificou quanto ao tipo (muscular, ligamentar, etc.) e região corporal específica.

Adicionalmente, o estudo utilizou dados de tracking e desempenho físico da PFF (Pro Football Focus) das temporadas 2022/23 e 2023/24 da Premier League, extraídos exclusivamente de partidas oficiais. Desta fonte foram quantificadas variáveis de carga externa, como a contagem e intensidade de corridas em alta velocidade (HSR) e sprints, o número de acelerações de média e alta intensidade[5, 6], os minutos jogados e o intervalo de dias entre as partidas.

Por fim, foi analisado um terceiro conjunto de dados de tracking, proveniente de um time da primeira divisão do campeonato brasileiro. Este dataset continha informações detalhadas tanto de sessões de treino quanto de jogos, incluindo não apenas dados de GPS, mas também métricas de Análise de Movimento Inercial (IMA), que capturam acelerações, desacelerações e mudanças de direção com alta precisão. Para possibilitar a análise, a variável de desfecho “lesão” foi definida a partir de uma premissa metodológica: um jogador era classificado como lesionado ao iniciar um período de ausência isolada e prolongada, superior a 7 dias, das atividades de treino. Após a aplicação desta classificação, os dados passaram por um rigoroso processo de higienização para remover erros e vieses antes da condução das análises comparativas.

2.2. Análise Estatística

A análise se baseou na comparação entre as métricas de jogadores que sofreram lesões e os que não se lesionaram em uma determinada temporada. Foram utilizadas análises visuais, como box plots e violin plots, para explorar as distribuições das variáveis entre os grupos. Para validar as hipóteses, foram empregados testes estatísticos, como o Teste t, para verificar se as diferenças observadas nas médias de carga eram estatisticamente significativas. O Teste t é uma ferramenta estatística utilizada para determinar se a diferença entre as médias de dois grupos independentes (neste caso, 'lesionados' e 'não lesionados') é real ou se poderia ter ocorrido por acaso.

3. Resultados

3.1. Análise Descritiva de Lesões

A análise do histórico de lesões do Transfermarkt mostrou que as regiões corporais mais afetadas são a coxa (29,9%) e o joelho (28,7%) (Figura 1). O tipo de lesão variou significativamente dependendo da região acometida: lesões ligamentares são mais frequentes nas articulações, enquanto lesões musculares afetam os demais segmentos, correspondendo a 99,7% das lesões na coxa (Figura 2).

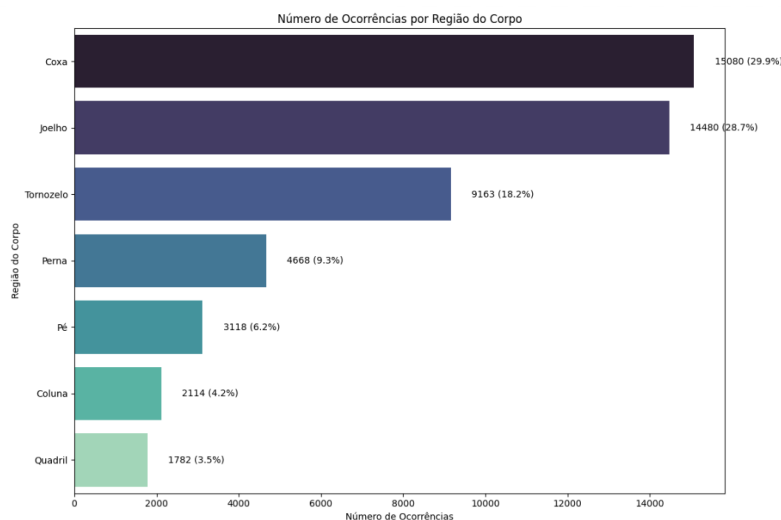


Figura 1. Número de ocorrências de lesão por região corporal, com dados extraídos do Transfermarkt.

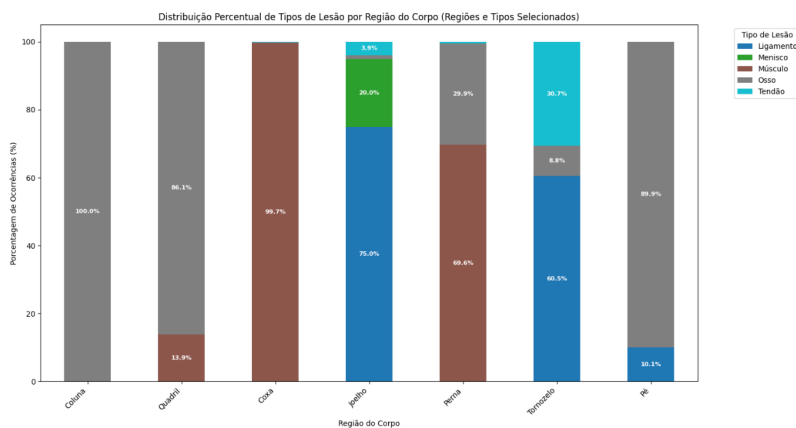


Figura 2. Distribuição percentual dos tipos de lesão para cada região corporal.

Com relação ao tempo de afastamento, observou-se que lesões ligamentares resultam em um tempo de afastamento significativamente maior em comparação com as lesões musculares (Figura 3).

3.2. Fatores de Risco para Lesões

A comparação entre os grupos de jogadores lesionados e não lesionados permitiu identificar diversos fatores de risco. A análise do perfil dos atletas revelou que jogadores

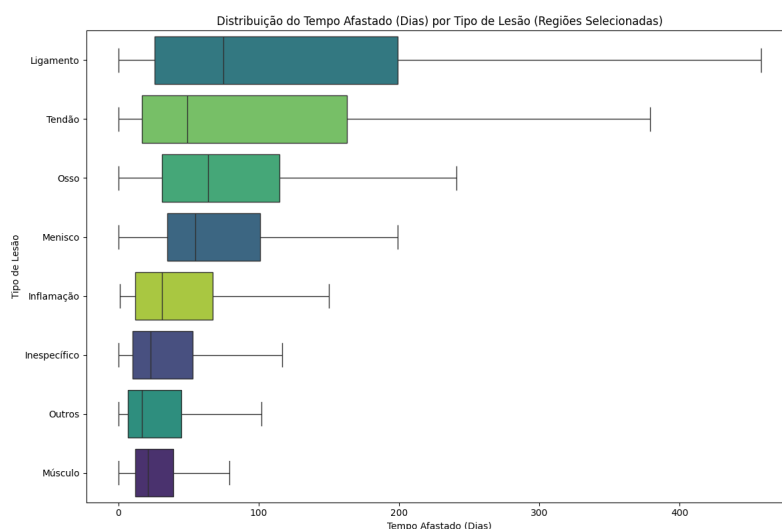


Figura 3. Comparação do tempo de afastamento (em dias) por tipo de lesão.

com idade mais avançada e com um maior número de lesões prévias em seu histórico apresentaram maior propensão a se lesionar (Figura 4).

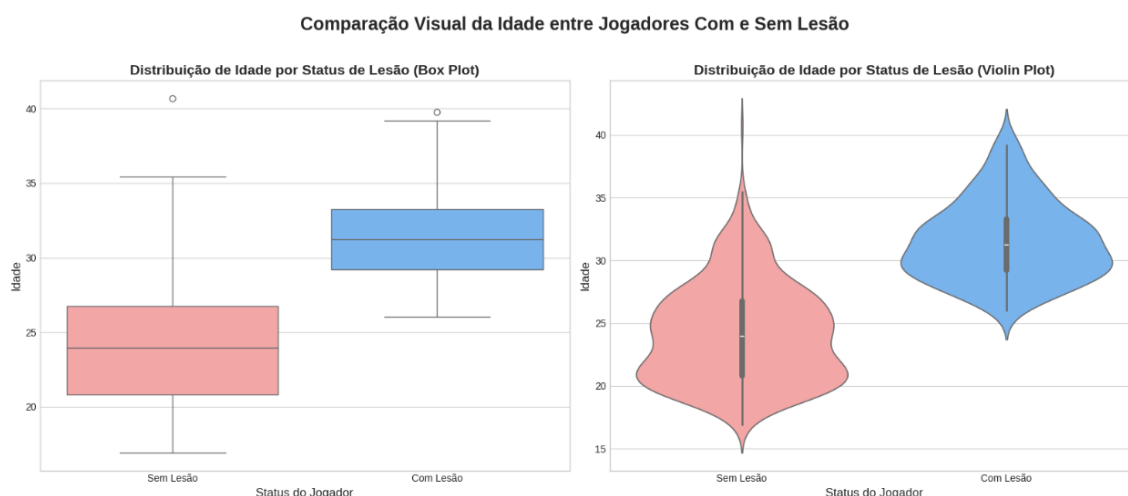


Figura 4. Distribuição da idade dos jogadores, comparando o grupo que sofreu lesão com o grupo que não se lesionou na temporada.

A análise da carga de trabalho, realizada a partir do dataset que combina treinos e jogos, demonstrou que, conforme validado pelo Teste t, os jogadores do grupo "Lesionado" exibiram valores médios de carga significativamente mais altos no período anterior à lesão. As métricas que apresentaram as diferenças mais expressivas foram a Carga Total de Aceleração (p-valor 3.39×10^{-2}), o número de acelerações de alta intensidade (p-valor 4.16×10^{-2}), a Distância Total percorrida (p-valor 4.15×10^{-2}) e o total de esforços de desaceleração (p-valor 4.06×10^{-2}) (Figura 5). Esses resultados sugerem que picos de carga, especialmente relacionados a esforços explosivos de aceleração e desaceleração, estão fortemente associados à ocorrência de lesões.

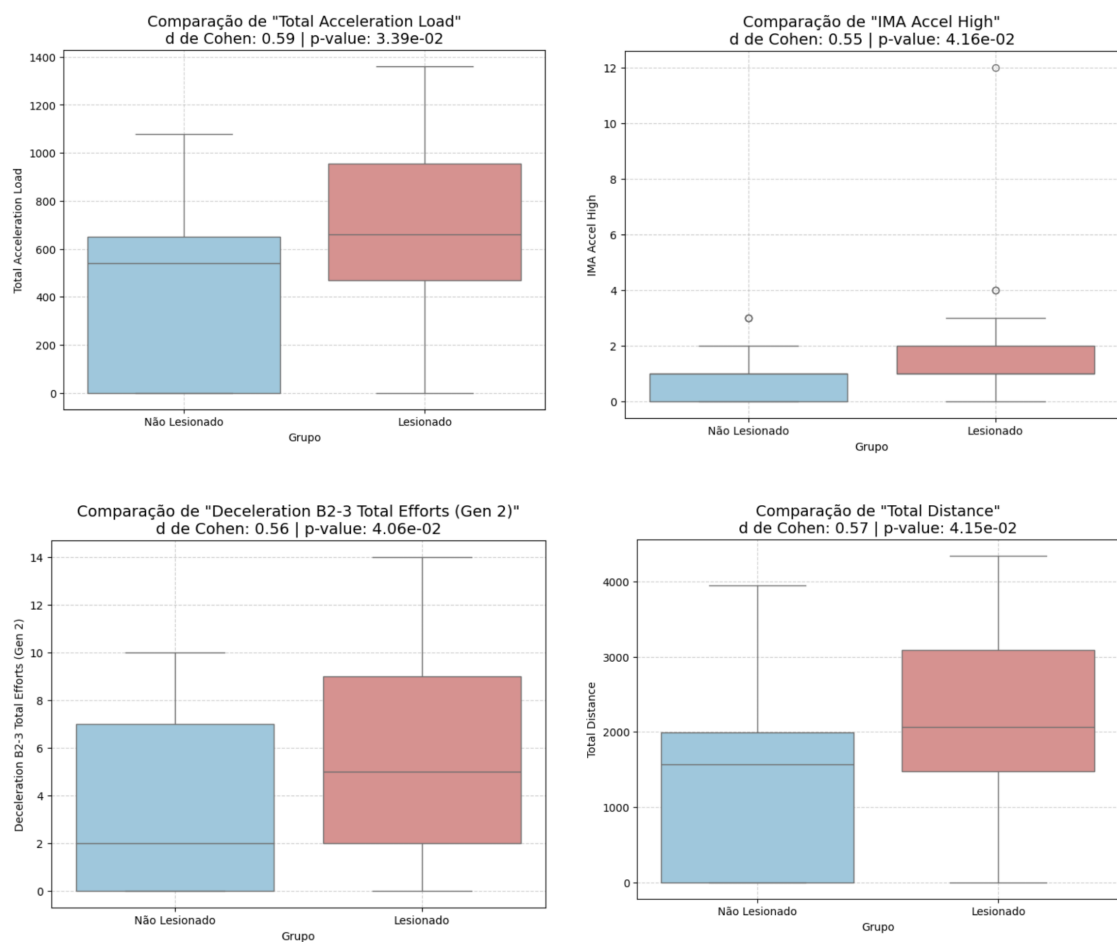


Figura 5. Comparação de métricas de carga de trabalho entre os grupos lesionado e não lesionado (apresentadas apenas as métricas em que houve diferença estatisticamente significativa entre os grupos pelo Teste t).

4. Conclusões

Este estudo conseguiu identificar fatores relevantes associados à ocorrência de lesões musculares no futebol com base em dados de evento e tracking. A análise demonstrou que as lesões musculares na coxa são as mais frequentes no esporte. Embora ocasionem afastamentos mais curtos em comparação com lesões ligamentares ou ósseas, elas apresentam uma maior tendência à reincidência. Os fatores preditivos mais fortes para a ocorrência dessas lesões foram a idade do atleta, seu histórico prévio de contusões e as métricas de carga de trabalho relacionadas a esforços explosivos. Notavelmente, observou-se que os jogadores lesionados tenderam a exibir picos agudos de carga — refletidos em acelerações, desacelerações e na distância total percorrida — nos períodos que antecederam o afastamento.

As principais limitações do trabalho incluíram o acesso restrito a dados clínicos detalhados e de treino, os desafios na integração de fontes de dados heterogêneas e a baixa frequência estatística das lesões, que dificulta a modelagem. Ainda assim, os achados reforçam a importância do monitoramento da carga de trabalho. As conclusões podem auxiliar clubes e comissões técnicas a estabelecer programas de treinamento mais individualizados e estratégias de prevenção mais eficazes, com foco nos fatores de risco identificados.

Referências

- [1] Koch M, Klügl M, Frankewycz B, Lang S, Worlicek M, Popp D, Alt V, Krutsch W. Football-related injuries are the major reason for the career end of professional male football players. *Knee Surg Sports Traumatol Arthrosc.* 2021 Nov;29(11):3560–68. doi:10.1007/s00167-021-06684-8.
- [2] Hägglund M, Waldén M, Magnusson H, Kristenson K, Bengtsson H, Ekstrand J. Injuries affect team performance negatively in professional football: an 11-year follow-up of the UEFA Champions League injury study. *Br J Sports Med.* 2013 Aug;47(12):738–42. doi:10.1136/bjsports-2013-092215.
- [3] Oliveira-Júnior O, Gabbett TJ, Bittencourt NFN, Quintão RC, Reis GF, Claudino JG, Lasmar RCP, Leopoldino AAO. Potential financial loss and risk factors for hamstring muscle injuries in elite male Brazilian soccer players: a season-long prospective cohort pilot study. *Front Sports Act Living.* 2024 Sep 24;6:1360452.
- [4] Howden. Men's European football injury index 2023/24 [Internet]. Howden Group; 2024 [citado em 7 jul. 2025]. Disponível em: https://www.howdengroupholdings.com/sites/default/files/2024-10/Howden%20Insurance%20Brokers%20Ltd%20-%20Men%27s%20European%20football%20injury%20index%202023_24.pdf.
- [5] SkillCorner. Making the leap: Quantifying the physical differences between the big-5 leagues and the second tiers [Internet]. SkillCorner; 2022 [citado em 7 jul. 2025]. Disponível em: <https://skillcorner.com/blog/making-the-leap-1>.
- [6] SkillCorner. New Feature: Explosive Accelerations [Internet]. SkillCorner; 2023 [citado em 7 jul. 2025]. Disponível em: <https://skillcorner.com/blog/new-feature-explosive-accelerations>.

Aprimorando o Atomic VAEP com Classificador de Sucesso

Avelar Hostalácio¹, Enzo Pierazolli¹, Fernando Vilela¹,
Iago Gonçalves¹, Victor Murta¹

¹Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brazil

avelarrh@ufmg.br, eppierazolli@ufmg.br, fvb@ufmg.br,

iagogabino@ufmg.br, victormurta@ufmg.br

Abstract. *In this work, we propose an enhancement to the Atomic VAEP model by incorporating a success classifier trained with tactical features derived from tracking data. Using detailed spatiotemporal data from the 2023–2024 Premier League season, we developed five contextual variables that better reflect the in-game dynamics at the moment of each action. These variables include defensive compactness, goal-side triangulation, player speed, local pressure and support, and goalkeeper angular deviation. The results show specific improvements over the original model in metrics such as Brier Score, Log Loss, and ROC AUC. The proposed integration enables a more contextualized evaluation of player actions and contributes to more robust and interpretable performance analysis tools in football.*

Resumo. *Neste trabalho, propomos um aprimoramento do modelo Atomic VAEP por meio da incorporação de um classificador de sucesso treinado com features táticas derivadas de dados de tracking. Utilizando dados espaço-temporais detalhados da temporada 2023–2024 da Premier League, desenvolvemos cinco variáveis contextuais que refletem melhor a dinâmica em campo no momento de cada ação. As variáveis incluem compactação defensiva, triangulação para o gol, velocidade do jogador, pressão e suporte locais, e desvio angular do goleiro. Os resultados obtidos mostram melhorias pontuais em relação ao modelo original nas métricas de Brier Score, Log Loss e ROC AUC. A integração proposta permite uma avaliação mais contextualizada das ações dos jogadores e contribui para ferramentas de análise de desempenho mais robustas e interpretáveis no futebol.*

1. Introdução

No futebol moderno, a análise quantitativa tem desempenhado um papel fundamental na avaliação de desempenho de jogadores e equipes, auxiliando técnicos, analistas e departamentos de recrutamento na tomada de decisões baseadas em dados. Com o avanço de tecnologias de coleta e processamento de informações, como sistemas de tracking e modelos preditivos, tornou-se possível quantificar aspectos cada vez mais complexos do jogo.

Entre as abordagens mais promissoras nesse cenário está o modelo Atomic VAEP (Valuing Actions by Estimating Probabilities) [Decroos et al. 2019], que busca atribuir

valor a cada ação individual em campo com base em sua contribuição esperada para aumentar ou diminuir a probabilidade de gol. A proposta central do modelo é oferecer uma métrica que vá além de estatísticas tradicionais como gols e assistências, capturando o impacto tático de ações como passes, finalizações, interceptações e dribles independentemente do seu resultado imediato. Dessa forma, o VAEP permite uma análise mais granular e justa do desempenho dos atletas, valorizando ações de construção e prevenção que muitas vezes passam despercebidas.

No entanto, apesar de sua robustez e ampla aceitação, o modelo original do Atomic VAEP apresenta limitações importantes. Por depender exclusivamente de dados de eventos (event data), ele não capta nuances contextuais e espaciais do jogo que influenciam diretamente o valor real de uma ação. Elementos como a pressão exercida pela defesa, o posicionamento do goleiro, o suporte de companheiros próximos ou a compactação do sistema defensivo adversário são ignorados, o que compromete a fidelidade da avaliação em momentos críticos da partida.

Diante dessas limitações, este trabalho propõe um aprimoramento do modelo Atomic VAEP por meio da incorporação de features táticas derivadas de dados de tracking: dados espaço-temporais que descrevem a movimentação contínua de todos os jogadores e da bola ao longo da partida. Com base nesses dados, desenvolvemos cinco variáveis contextuais que capturam com maior fidelidade a dinâmica do jogo no instante de cada ação: compactação defensiva, triangulação para o gol, velocidade do jogador, pressão e suporte locais, e desvio angular do goleiro.

Essas variáveis alimentam um classificador de sucesso, responsável por estimar a probabilidade de que uma ação tenha êxito sob determinadas condições táticas e espaciais. Ao substituir a estimativa empírica baseada em frequência por uma predição modelada a partir do contexto real de jogo, buscamos aumentar a precisão e a interpretabilidade do valor atribuído a cada ação.

Nosso objetivo é contribuir com o desenvolvimento de ferramentas analíticas mais robustas, que combinem o poder explicativo do modelo VAEP com a riqueza informacional dos dados de tracking, permitindo uma análise de desempenho mais precisa, personalizada e alinhada com os princípios táticos do jogo.

2. Metodologia

A metodologia deste trabalho está organizada em um pipeline experimental dividido em três etapas principais: (1) definição de um modelo de referência com base no Atomic VAEP original, (2) desenvolvimento de um classificador de sucesso utilizando *features* táticas extraídas de dados de tracking, e (3) integração das probabilidades estimadas pelo classificador ao cálculo de valor das ações no modelo VAEP. Cada etapa foi projetada para avaliar o impacto incremental da introdução de informações contextuais espaciais e temporais na modelagem da probabilidade de sucesso de uma ação.

2.1. Fonte e Pré-processamento dos Dados

2.1.1. Fonte dos Dados

Utilizamos dados fornecidos pela **PFF (Pro Football Focus)** referentes à temporada 2023/2024 da **Premier League**, abrangendo informações espaço-temporais detalhadas

de partidas completas. O conjunto de dados é composto por três fontes principais:

- **Metadados de eventos:** descrevem cada ação registrada durante os jogos, incluindo o tipo (por exemplo, passe, chute, interceptação), o resultado (sucesso ou falha) e a localização no campo no momento da execução;
- **Dados de tracking:** fornecem as posições x, y de todos os jogadores e da bola em cada *frame* da partida, com frequência de 60 Hz, permitindo reconstruir com alta precisão os deslocamentos e interações em campo;
- **Informações táticas adicionais:** incluem indicadores binários de posse de bola, identificação do time atacante no momento da ação, direção de ataque e outras variáveis auxiliares para segmentar períodos ofensivos e defensivos.

Para viabilizar a integração com o modelo Atomic VAEP, realizamos a conversão dos dados de evento para o formato SPADL (*Soccer Player Action Description Language*), utilizando o módulo *socceraction*. Este formato padroniza a representação das ações em campo e facilita a aplicação de modelos de valoração. Em seguida, transformamos as ações em **Atomic SPADL**, que descreve as ações em unidades atômicas, que melhor se adequam à granularidade exigida pelo modelo.

Durante o processo de conversão, mantivemos apenas a primeira ocorrência de cada `event_id`, a fim de evitar duplicidades oriundas da sobreposição de eventos derivados (por exemplo, um passe que é também registrado como uma perda de posse). Essa filtragem garante a consistência e a não redundância do conjunto de ações analisadas.

Essa estrutura de dados combinada, enriquecida com coordenadas de tracking e variáveis contextuais táticas, serviu como base para a construção das features e treinamento do classificador de sucesso, conforme descrito nas próximas seções.

2.1.2. Pré-processamento

O processo de pré-processamento foi fundamental para garantir a integridade, padronização e alinhamento temporal dos diferentes conjuntos de dados utilizados no modelo. As transformações realizadas abrangeram dados de elenco, eventos (SPADL), metadados e tracking, conforme descrito a seguir.

Limpeza e Estruturação dos Dados de Elenco. O arquivo `rosters.csv`, contendo informações aninhadas sobre jogadores e equipes, foi convertido em um *DataFrame* estruturado com os seguintes passos:

- Leitura do arquivo bruto com *parsing* das colunas `player` e `team` para extração dos respectivos identificadores (`player_id` e `team_id`);
- Seleção e renomeação das colunas relevantes: `game_id`, `player_id`, `shirt`, `team_id`;
- Conversão de todos os campos para tipos numéricos apropriados, garantindo a consistência das chaves de junção;
- Salvamento do resultado final em formato *Parquet* como `rosters.parquet`, visando eficiência de leitura nos próximos estágios do pipeline.

Padronização Direcional dos Dados SPADL. Os dados de evento convertidos para o formato SPADL passaram por uma etapa de padronização direcional para garantir a coerência na orientação das ações. A metodologia consistiu em:

- Carregamento dos dados SPADL brutos e do mapa de times, contendo a informação `homeTeamStartLeft`;
- Aplicação da função `play_left_to_right`, que transforma as coordenadas das ações para que todas fluam da esquerda para a direita, independentemente do tempo ou lado de campo em que ocorreram;
- Validação visual da padronização por meio de gráficos “Antes vs. Depois”, assegurando a integridade espacial das ações;
- Armazenamento do resultado final como `spadl_ltr.parquet`.

Essa padronização é essencial para garantir consistência espacial no treinamento do classificador e na construção das *features* contextuais.

Verificação Visual da Posição dos Goleiros. Como parte da verificação da consistência dos dados de tracking, foi realizada uma inspeção visual da posição dos goleiros ao longo dos jogos. Esse procedimento teve como objetivos:

- Confirmar que os dados de tracking representavam corretamente a posição dos atletas ao longo da partida;
- Validar a lógica de identificação do goleiro em campo, aspecto crítico para a correta construção da *feature* de desvio angular.

As visualizações geradas corroboraram a integridade dos dados e reforçaram a confiabilidade do pipeline de tracking.

Processamento em Lote dos Dados de Tracking. Dada a estrutura por jogo dos arquivos de tracking e metadados, foi desenvolvido um pipeline de pré-processamento em lote com os seguintes passos para cada partida:

1. Leitura dos arquivos de metadados e filtragem de frames com `possession_id` válido;
2. Armazenamento do *DataFrame* de metadados limpo;
3. Utilização dos `frame_ids` resultantes para filtrar os dados de tracking dos jogadores, assegurando sincronização precisa entre os dois conjuntos;
4. Aplicação de rotinas de limpeza, incluindo tratamento de inconsistências e valores extremos na variável `speed`;
5. Armazenamento do *DataFrame* de tracking limpo em disco.

Esse processamento permitiu criar uma base de tracking estruturada e sincronizada, pronta para a extração eficiente de variáveis contextuais associadas a cada ação do modelo.

2.2. Etapa 1: Modelo Atomic VAEP Baseline

Nesta etapa inicial, foi implementado um modelo de referência (*baseline*) utilizando exclusivamente as *features* tradicionais extraídas do SPADL (Soccer Player Action Description Language). Essas features incluem:

- **Tipo de Ação:** codificação categórica de eventos como passe, chute, desarme, entre outros;

- **Localização da Ação:** coordenadas (x, y) normalizadas do início e fim da ação no campo;
- **Distância até o Gol:** métrica escalar da distância entre o ponto final da ação e o centro da baliza adversária;
- **Ângulo até o Gol:** ângulo formado entre a linha do ponto final da ação e as duas traves;
- **Contexto Imediato da Jogada:** informações básicas como posse de bola e sequência de ações do jogador.

A motivação para essa etapa é estabelecer uma linha de base justa e interpretável para avaliar o impacto das futuras incorporações de contexto posicional e tático. Por não utilizar dados de tracking, o modelo representa o estado da arte com base apenas em eventos, sendo ideal para comparação com métodos que integram dados espaço-temporais. Além disso, o desempenho deste baseline também permite quantificar o ganho marginal proporcionado por cada grupo de novas variáveis adicionadas nas etapas seguintes.

2.3. Etapa 2: Classificador de Sucesso com Features Táticas

Nesta etapa, desenvolvemos um classificador binário para estimar a probabilidade de sucesso de cada ação registrada, com base em variáveis contextuais extraídas de dados de tracking. O objetivo é substituir a simples frequência empírica de acertos por uma estimativa modelada do contexto tático real de cada jogada. O classificador foi treinado sobre rótulos de sucesso oriundos do SPADL, definidos com base no resultado direto da ação (por exemplo, passe completo, drible bem-sucedido, finalização no alvo).

As features utilizadas foram construídas a partir dos dados de tracking descritos na seção anterior e incluem: compactação defensiva, triangulação para o gol, velocidade do portador da bola e companheiros próximos, pressão e suporte locais, número de linhas de passe abertas, e desvio angular do goleiro. Essas variáveis fornecem uma visão contextual abrangente do entorno da ação, incluindo tanto a organização defensiva adversária quanto os recursos ofensivos disponíveis.

O modelo utilizado foi o `XGBoostClassifier`, devido à sua robustez para dados tabulares, capacidade de interpretar relações não lineares e tolerância a dados ruidosos ou parcialmente correlacionados. As previsões do classificador foram posteriormente utilizadas como uma nova feature contínua (`success_score`) no modelo VAEP final, conforme descrito na próxima seção.

2.3.1. Otimização de Hiperparâmetros

Para garantir a melhor configuração do classificador, foi aplicada uma busca aleatória de hiperparâmetros (*Randomized Search*) com validação cruzada estratificada por partidas (*GroupKFold*), a fim de evitar vazamento de informações entre os conjuntos de treino e teste. Utilizou-se como grupos os identificadores de partida (`game_id`) do conjunto de dados.

A busca explorou 20 combinações de hiperparâmetros, variando valores de `n_estimators`, `max_depth`, `learning_rate`, `subsample`, `colsample_bytree`, `reg_alpha`, `reg_lambda` e `scale_pos_weight`. A

métrica de desempenho utilizada foi a área sob a curva ROC (AUC). O melhor modelo encontrado por essa busca foi utilizado como classificador final, fornecendo a estimativa probabilística de sucesso para cada ação no conjunto de dados.

2.3.2. Compactação Defensiva

A feature de compactação defensiva quantifica o quão agrupada está a linha de defesa adversária no momento da posse de bola do time atacante. Para cada posse, identificamos os jogadores do time que estão defendendo (isto é, que não pertencem ao time atacante) e calculamos o desvio padrão das coordenadas x e y desses jogadores. Valores baixos de desvio indicam uma defesa mais compacta, enquanto valores altos representam uma defesa mais espaçada e vulnerável.

Motivação: uma linha defensiva compacta tende a dificultar a progressão de jogadas, restringindo espaços e linhas de passe. Por outro lado, uma defesa desorganizada ou espaçada geralmente permite maior liberdade para infiltrações, cruzamentos e finalizações. Portanto, essa feature capta de forma agregada o estado coletivo da organização defensiva.

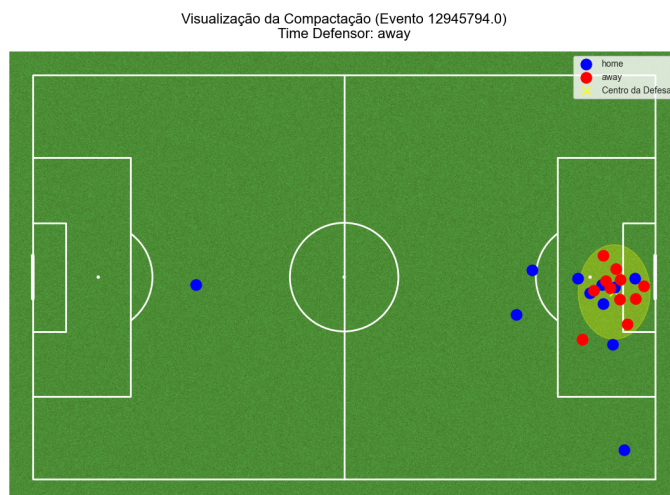


Figure 1. Exemplo Visual da Feature de Compactação Defensiva

2.3.3. Triângulo para o Gol

Essa feature busca medir o número de defensores que estão dentro do triângulo formado pelas traves do gol e o ponto da ação do jogador atacante. Para cada evento, construímos um polígono a partir das coordenadas das traves (ajustadas conforme o lado de ataque) e a posição do jogador com a bola. Em seguida, verificamos quantos jogadores do time adversário estão dentro desse triângulo (Figura 2).

Motivação: a presença de defensores dentro desse polígono reduz diretamente o campo de visão e o espaço útil entre o atacante e o gol, funcionando como um indicador da obstrução do caminho ao gol. Esse tipo de obstrução é particularmente relevante para prever finalizações, cruzamentos e ações decisivas em zonas perigosas.

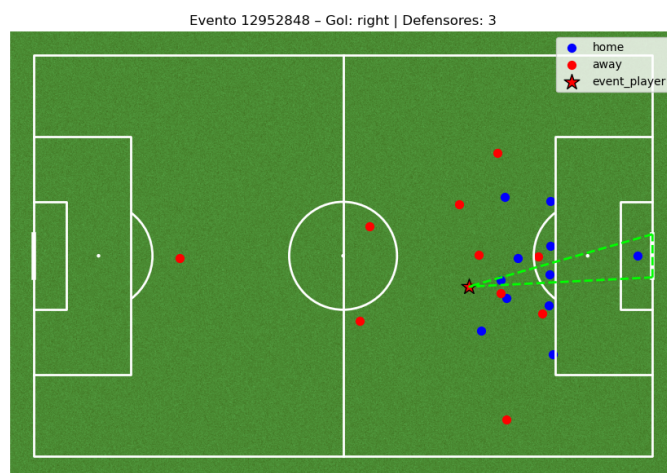


Figure 2. Exemplo Visual da Feature Triângulo para o Gol

2.3.4. Velocidade do Jogador e dos Companheiros

A feature de velocidade captura o estado dinâmico do jogador com a bola e de seus companheiros em um raio de 10 metros. Utilizando os dados de tracking sincronizados com cada posse, estimamos a velocidade instantânea do portador da bola e a média das velocidades de companheiros próximos.

Motivação: em situações de transição ofensiva ou contra-ataque, a velocidade é um indicador crítico da intensidade e do potencial de desorganização defensiva. Jogadas realizadas em alta velocidade têm maior probabilidade de romper linhas e gerar oportunidades antes que a defesa se reorganize. Além disso, jogadores em movimento próximo ao portador da bola aumentam o leque de opções táticas disponíveis (passes em profundidade, tabelas, arrastamento de marcação etc.).

2.3.5. Pressão, Suporte e Linhas de Passe

Esta *feature* (Figura 3) captura aspectos táticos do entorno do jogador com a bola. Para cada posse:

- Calculamos o número de defensores em raios de 3m e 6m (pressão imediata);
- Contamos companheiros de equipe em um raio de 8m (suporte ofensivo);
- Estimamos a pressão ponderada, levando em conta distância e o ângulo de abordagem dos defensores (maior peso para pressão frontal);
- Contamos o número de linhas de passe abertas, isto é, companheiros em posição avançada cuja rota de passe não esteja bloqueada por defensores.

Motivação: essas subfeatures fornecem uma leitura granular do contexto tático imediato. Alta pressão indica maior dificuldade de execução e maior risco de perda da posse, enquanto suporte e linhas de passe abertas aumentam as chances de sucesso da jogada. Juntas, essas medidas expressam o equilíbrio entre risco e oportunidade, que é central para a tomada de decisão no futebol.

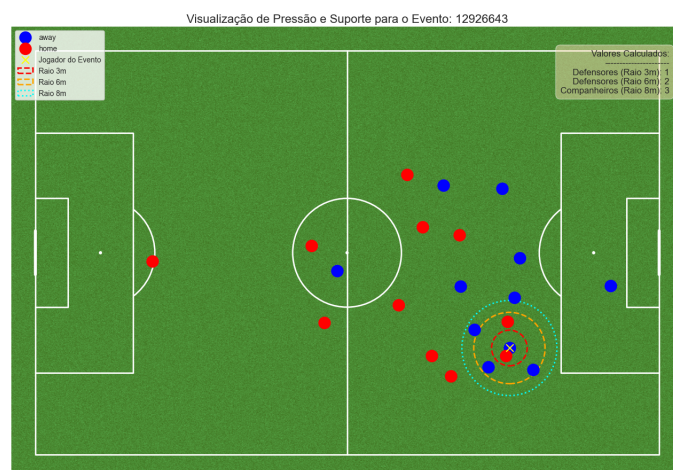


Figure 3. Exemplo Visual das Features de Pressão

2.3.6. Desvio Angular do Goleiro

Essa feature estima o quanto o goleiro está fora de posição no momento de uma finalização. Para cada evento com possibilidade de chute, comparamos o ângulo formado entre a posição do jogador e o centro do gol com o ângulo entre o jogador e a posição real do goleiro. A diferença angular (em graus) representa o desposicionamento do goleiro (Figura 4).

Motivação: um goleiro bem posicionado reduz significativamente o espaço útil para o chute. Já um goleiro ligeiramente fora de centro — seja por antecipação, deslocamento tardio ou desatenção — oferece brechas que aumentam a probabilidade de sucesso da finalização. Essa feature contribui diretamente para modelar o valor esperado de chutes em diferentes cenários, com base em pequenas variações posicionais do arqueiro.

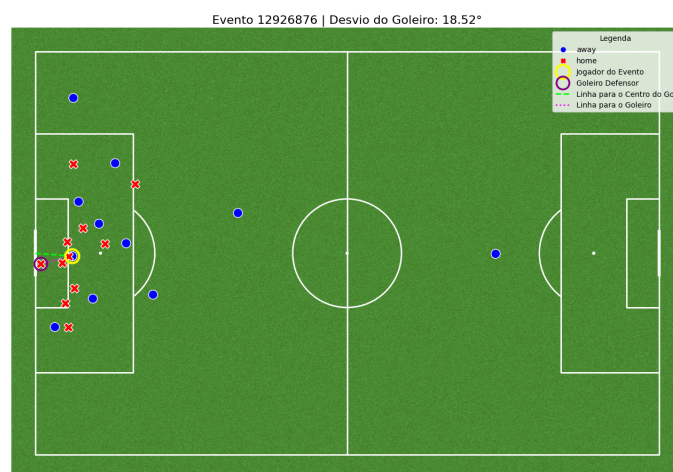


Figure 4. Exemplo Visual da Feature Desvio Angular do Goleiro

2.4. Etapa 3: Modelo VAEP com success_score

A probabilidade predita pelo classificador da Etapa 2 (`success_score`) foi então incorporada como uma feature adicional no modelo VAEP, que agora é treinado com o

algoritmo XGBoost e utiliza todas as features do modelo baseline (Etapa 1) mais a `success_score`. Em outras palavras, o novo modelo de valor de ação se beneficia tanto das informações estruturais dos eventos SPADL quanto da inferência tática derivada dos dados de tracking, sintetizada em forma de probabilidade predita.

Motivação: a inclusão da `success_score` permite ao modelo VAEP capturar, de forma indireta, nuances espaciais e comportamentais que influenciam a chance de êxito de uma ação, mas que não estão diretamente representadas nas features tradicionais. Como o classificador da Etapa 2 é alimentado com variáveis de contexto tático (como pressão, compactação defensiva, velocidade do setor etc.), a `success_score` atua como um vetor sintético dessas condições contextuais.

Vantagem metodológica: esse acoplamento modular entre modelos permite separar a modelagem do contexto tático (Etapa 2) da estimativa de valor das ações (Etapa 3), resultando em uma abordagem mais interpretável e extensível. Em vez de inflar o modelo VAEP com dezenas de features táticas adicionais, a complexidade contextual é absorvida e traduzida em uma única variável probabilística.

Considerações técnicas: o modelo VAEP com `success_score` foi treinado e validado utilizando o mesmo protocolo da Etapa 1 (validação estratificada por partidas). A performance foi avaliada com as mesmas métricas — Brier Score, AUC e correlação com expected goals — permitindo uma comparação direta com o modelo baseline. O impacto da `success_score` também foi analisado via *feature importance*, confirmando sua contribuição significativa para a melhora do modelo final.

Otimização hiperparamétrica: para o modelo VAEP com `success_score`, realizamos também uma *Randomized Search* para seleção dos melhores hiperparâmetros do algoritmo XGBoost, incorporando as mesmas variáveis táticas da Etapa 2. A busca foi conduzida com *GroupKFold* para garantir validação por partidas, com foco na métrica de AUC. Esta otimização foi aplicada apenas ao modelo com variáveis de tracking, e não ao baseline, de forma a preservar a comparabilidade e isolar os efeitos das novas variáveis táticas sobre o desempenho preditivo..

Ambos os modelos – baseline e enriquecido – foram avaliados com base em três métricas principais:

- **Brier Score;**
- **Log Loss;**
- **ROC AUC.**

A escolha das métricas e a estrutura de validação cruzada adotadas seguem as boas práticas recomendadas por trabalhos recentes na área de sports analytics [Fernández et al. 2024].

As métricas foram computadas separadamente para a probabilidade de marcar (*scores*) e de sofrer gols (*concedes*). O impacto da integração do `success_score` foi quantificado comparando os dois modelos com os mesmos dados de teste.

3. Análise dos resultados

Os gráficos de importância apresentados nas próximas subseções foram gerados com base na métrica de *total_gain* do XGBoost, que quantifica o ganho de informação acumulado

por cada variável nas divisões das árvores. Essa medida permite comparar a relevância relativa das features dentro de um mesmo modelo.

3.1. Efeito das Features do Atomic Vaep Baseline

O modelo baseline demonstra forte dependência de features espaciais fundamentais, com `dist_to_goal_a0` dominando claramente a importância relativa. As coordenadas de posição (`x_a0`, `y_a0`) e o tipo de ação (`actiontype_shot_a0`) também apresentam contribuições significativas. Essa distribuição (Figura 5) confirma que o modelo tradicional se apoia principalmente em métricas geométricas diretas e na classificação básica das ações, sem capturar nuances táticas do contexto de jogo.

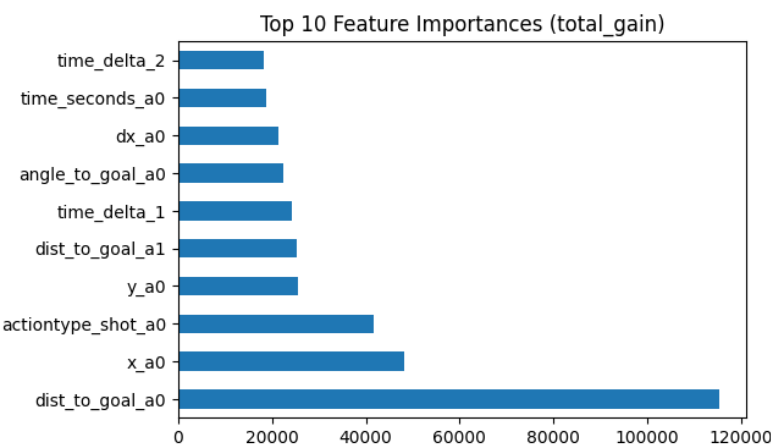


Figure 5. Importância das features no modelo VAEP baseline

3.2. Efeito das Features do Classificador

O classificador de sucesso revela uma distribuição (Figura 6) mais equilibrada entre as features, com destaque para variáveis de movimento (`dx_a0`, `dx_a1`) e métricas táticas como `pressao_posicional_ponderada`. A presença de `defensive_compaction_x` entre as principais features valida a relevância das variáveis contextuais derivadas do tracking. Notavelmente, o modelo valoriza tanto aspectos dinâmicos quanto posicionais, sugerindo que o sucesso de uma ação depende fortemente do contexto tático instantâneo.

3.3. Efeito das Features do Atomic Vaep Final

A incorporação do `success_score` como feature principal no modelo final demonstra o impacto substancial da integração do classificador tático. Com importância total superior a 25.000, essa variável sintética captura efetivamente o contexto espacial e comportamental modelado na Etapa 2. As features tradicionais como `actiontype_shot_a0` e `dist_to_goal_a0` mantêm relevância, mas agora complementam a informação contextual agregada, resultando em um modelo mais robusto e contextualmente informado (Figura 7).

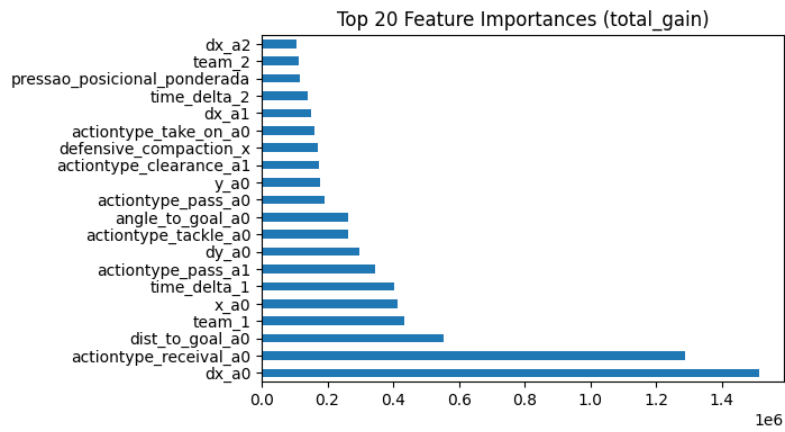


Figure 6. Importância das features no classificador de sucesso

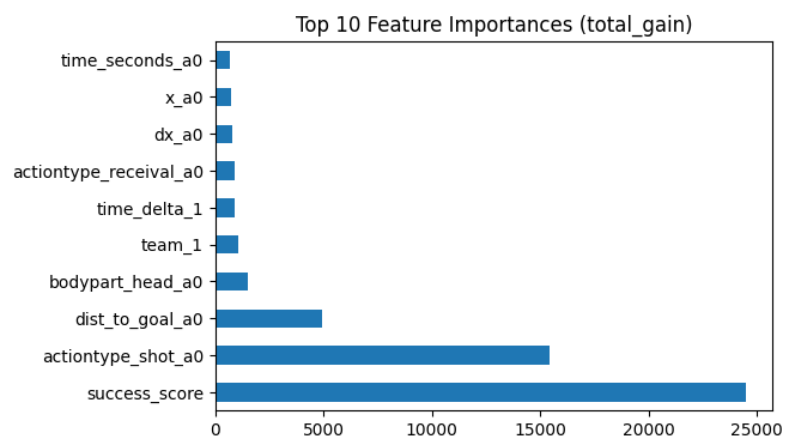


Figure 7. Importância das features no modelo VAE com success_score

4. Conclusão

Ao incorporar um classificador de sucesso alimentado por *features* táticas extraídas de dados de tracking da Premier League 2023–2024, conseguimos enriquecer o contexto das ações individuais realizadas pelos jogadores.

Os resultados (Tabela 1) demonstram ganhos pontuais nas principais métricas de avaliação. Na de previsão de *scores*, o modelo aprimorado apresentou melhora em todas as métricas: o *Brier Score* reduziu de 0,01008 para 0,00954, o *Log Loss* de 0,05153 para 0,04847, e a área sob a curva ROC (AUC) aumentou de 0,79694 para 0,81630. Da mesma forma, para a tarefa de *concedes*, observou-se melhora em todas as métricas, com *Brier Score* de 0,00176 para 0,00175, *Log Loss* de 0,01166 para 0,01153 e AUC de 0,81841 para 0,82452.

Table 1. Comparação das métricas entre os modelos A-VAEP e A-VAEP com `success_score`

Métrica	Baseline A-VAEP	A-VAEP + <code>success_score</code>
<i>Scores</i>		
Brier Score	0.01008	0.00954
Log Loss	0.05153	0.04847
ROC AUC	0.79694	0.81630
<i>Concedes</i>		
Brier Score	0.00176	0.00175
Log Loss	0.01166	0.01153
ROC AUC	0.81841	0.82452

Esses avanços indicam que a adição de informações posicionais e táticas em tempo real pode aumentar a capacidade do modelo de estimar o valor real de uma ação no futebol. Além de melhorar o desempenho preditivo, o modelo proposto fornece interpretações mais alinhadas com a análise qualitativa realizada por analistas táticos, departamentos de desempenho e equipes técnicas. Assim, este estudo representa um passo importante em direção a sistemas analíticos mais completos, contextualmente ricos e aplicáveis ao ambiente competitivo do futebol profissional.

References

- Decroos, T., Bransen, L., Haaren, J. V., and Davis, J. (2019). Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1851–1861.
- Fernández, J., Cervone, D., and Bornn, L. (2024). Methodology and evaluation in sports analytics: challenges, approaches and lessons learned. *ResearchGate Preprint*.

Futmetria: identificando padrões de jogo e assinatura dos times

Bernardo Vale dos Santos Bento, Filipe Mauro da Terra Caldeira,
Gabriel Martins Miranda, Lucas Augusto da Silva Gonçalves
{bernardovale, filipecaldeira, lucas.goncalves}@dcc.ufmg.br
gabrielmartinsmiranda800@gmail.com

Julho 2025

Abstract

Team analysis and opponent strategy identification is one of the key stages of pre-game preparation, as relevant as the training itself. Therefore, this paper aims to identify play patterns for the purpose of increasing the effectiveness of the defensive tactical preparation of a team. In order to do that, real events data from a complete season were utilized to cluster and evaluate the actions. The results demonstrate that it is possible to organize different actions in different groups and evaluate the quality of these actions.

Resumo

Análise de times e identificação de estratégias dos adversários é um dos estágios mais importantes na preparação pré-jogo, tão relevante quanto o próprio treino. Tendo isso em vista, este artigo tem como objetivo identificar padrões de jogo de modo a facilitar a preparação tática defensiva de um time. Para isso, são utilizados dados reais de eventos de uma temporada completa para agrupar e avaliar as ações. Os resultados demonstram que é possível organizar diferentes ações em diferentes conjuntos e avaliar a qualidade dessas ações.

Contents

1	Introdução	3
2	Trabalhos Relacionados	3
3	Caracterização e análise de dados	4
3.1	Dados	4
3.2	SPADL e Atomic-SPADL	4
3.3	Agrupamento de ações	4
3.4	VAEP	5
3.5	Identificação das jogadas principais dos times	5
4	Demonstração	5
4.1	Caso 1	5
4.2	Caso 2	6
5	Conclusão	7

1 Introdução

No cenário do futebol moderno, onde cada detalhe pode decidir um campeonato e entender como seu adversário joga pode proporcionar vantagens decisivas, a análise de dados transcendeu a simples contagem de gols e assistências para se tornar uma ferramenta indispensável na avaliação de performance para ser usada na tomada de decisões estratégicas cada vez mais assertivas. A capacidade de decifrar o que acontece em campo, valorizando cada passe, desarme ou movimento, oferece uma vantagem competitiva crucial. Por isso, a ideia deste trabalho é combinar a identificação de padrões táticos — a Identidade de um time — com a quantificação do valor de ações individuais. Ao combinar esses dois tipos de métricas, é possível criar várias análises valiosas, o que chamamos de Futmetria.

A utilização do SoccerMix (DECROSS, 2020), um algoritmo de clusterização que analisa milhares de ações e as agrupa em “padrões de jogo”. Ele nos ajuda a construir um vocabulário de jogadas-padrão de um clube, mostrando se esse clube prefere passes curtos ou lançamentos longos, por exemplo, e mapeia as áreas do campo onde esses padrões acontecem. Combinado ao VAEP (DECROSS, 2018) — Valuing Actions by Estimating Probabilities — representaram um salto qualitativo, permitindo atribuir um valor objetivo a cada ação com bola, medindo como ela altera as probabilidades de uma equipe marcar ou sofrer um gol. Essa abordagem deu peso a ações defensivas cruciais e a passes de construção que não resultam em uma assistência direta.

Neste contexto, foram utilizados dados de evento da Premier League temporada 2017/2018 provenientes da plataforma Wyscout para modelar as análises e desenvolver um aplicativo interativo que o usuário pode navegar e realizar análises e comparações entre clubes e jogadores.

A principal contribuição deste estudo é, portanto, um framework que permite não apenas a caracterização do estilo tático de uma equipe através dos seus padrões de jogo mais frequentes, mas também a avaliação objetiva da sua eficiência. Com esta metodologia, é possível responder a perguntas complexas como: “Qual jogador desta liga melhor desempenharia a função de cruzamentos e dribles no último quarto do campo?” ou “Este time é eficiente em sair com a bola da sua própria defesa e criar boas chances?”. O resultado é uma visão integrada da performance, que captura tanto o estilo quanto a eficiência das jogadas praticadas.

2 Trabalhos Relacionados

Um desafio comum na análise de dados de eventos de futebol é a esparsidade dos dados, onde um jogador raramente executa a mesma ação no mesmo local mais de uma vez. Para lidar com isso, Decroos et al. [2] propõem o SoccerMix, uma técnica de clusterização probabilística que agrupa ações semelhantes para capturar o estilo de jogo de times e jogadores.

Outra abordagem fundamental na análise de desempenho é a valoração de ações individuais. Em vez de focar apenas em eventos raros como gols, o trabalho de Decroos et al. [1] introduz um framework para quantificar o valor de cada ação com a bola, como passes e dribles. A metodologia estima como cada ação impacta as probabilidades de marcar e sofrer um gol a curto prazo, permitindo uma avaliação mais granular da contribuição de um jogador.

Essas ferramentas foram fundamentais para a metodologia desenvolvida neste artigo, cujo objetivo é detectar padrões de jogo de times de futebol com base nos valores das ações que desempenham.

Também é importante destacar que esse é um tópico bastante importante para a análise do futebol baseada em dados e ocupa um papel relevante na frente de pesquisa nessa área. Pode-se afirmar isso, dado a outros trabalhos similares ao em questão, como o "Caracterização de Estratégias de Futebol com base na Frequência, Importância e Efetividade de Jogadas" [4]

3 Caracterização e análise de dados

3.1 Dados

Para realização deste trabalho, foram utilizados os dados de evento da Premier League 2017/2018 disponibilizados em [3]. Apesar de se ter utilizado os dados do campeonato inglês, todos os outros fornecidos na mesma plataforma também são válidos.

3.2 SPADL e Atomic-SPADL

Para representar de forma padronizada e estruturada as ações realizadas pelos jogadores durante uma partida, este trabalho adota o SPADL (Soccer Player Action Description Language) [1]. O SPADL é uma linguagem desenvolvida para descrever ações em jogos de futebol com maior consistência e facilidade de análise computacional. Ao invés de depender dos formatos proprietários e heterogêneos das diferentes fornecedoras de dados, o SPADL transforma os eventos brutos em um conjunto fixo de ações com atributos padronizados, como tipo de ação (passe, chute, drible etc.), jogador, equipe, tempo, posição inicial e final no campo, entre outros. Essa padronização permite que os dados de diferentes fontes sejam analisados de maneira unificada e facilita o desenvolvimento de modelos estatísticos e de aprendizado de máquina para a avaliação de desempenho e estilo de jogo.

Desse modo, os dados eram inicialmente convertidos para o padrão SPADL, a fim de possibilitar sua utilização na geração do VAEP associado às jogadas. Optou-se por trabalhar com esse padrão devido à incerteza inicial sobre qual seria a fonte de dados e, portanto, a utilização do SPADL seria compatível com diferentes fornecedores. Por fim, os eventos eram convertidos para o padrão Atomic-SPADL antes do agrupamento realizado pelo SoccerMix, e a junção desses dados com os valores referentes ao VAEP foi realizada por questões de desempenho.

3.3 Agrupamento de ações

O SoccerMix utiliza uma abordagem de soft clustering baseada em modelos de mistura para agrupar ações de jogo com base em suas características de localização e direção [2]. Esse processo ocorre em duas etapas: na primeira etapa, para cada tipo de ação (por exemplo, passes, dribles, interceptações), ajusta-se um modelo de mistura de distribuições gaussianas bidimensionais (gaussianas multivariadas) sobre as coordenadas de início dessas ações. Isso resulta em zonas probabilísticas que refletem áreas do campo onde essas ações ocorrem com diferentes densidades. Na segunda etapa, para cada componente identificado na etapa anterior, ajusta-se outro modelo de mistura — desta vez com

distribuições de von Mises — sobre os ângulos das direções dos movimentos, capturando o fato de que ações similares em termos espaciais podem apresentar direcionalidades distintas (como passes para frente, pra lateral ou pra trás). A combinação dessas duas etapas permite representar cada ação por um vetor de responsabilidades probabilísticas — um tipo de assinatura suave (não exclusiva) que indica o grau de pertença a cada componente de localização e direção. Esse modelo possibilita agrupar ações semelhantes em termos de posicionamento e intenção, criando uma representação intertemporal e espacial rica, que suporta análises de estilo de jogo mais expressivas e detalhadas.

Utilizou-se essa ferramenta com o objetivo de agrupar as ações individuais em grandes grupos com alta similaridade interna, possibilitando uma análise mais completa das jogadas e, portanto, a identificação de padrões que são chave para o foco de estudo deste artigo.

3.4 VAEP

O **VAEP** (Valuing Actions by Estimating Probabilities) é um framework que avalia cada ação com bola no futebol estimando seu impacto nas probabilidades de marcar ou sofrer gol nas próximas N jogadas [1]. Formalmente, considere um estado de jogo S_i após a ação a_i ; a valoração ofensiva é definida como

$$\Delta P_{\text{score}}(a_i) = P_{\text{score}}(S_i) - P_{\text{score}}(S_{i-1}),$$

e a valoração defensiva, como

$$-\Delta P_{\text{concede}}(a_i) = -(P_{\text{concede}}(S_i) - P_{\text{concede}}(S_{i-1})).$$

A valoração total da ação é a soma desses componentes:

$$\text{VAEP}(a_i) = \Delta P_{\text{score}}(a_i) - \Delta P_{\text{concede}}(a_i).$$

Dessa forma, é possível avaliar o perigo e a qualidade das ações realizadas por qualquer time, o que é fundamental para identificar padrões ofensivos que têm maior possibilidade de levar o adversário ao gol.

3.5 Identificação das jogadas principais dos times

Com a junção dessas duas ferramentas, conseguimos identificar claramente o estilo e planejamento de ações ofensivas tanto de times quanto de jogadores. Com elas, é possível indicar uma tendência estatística de jogadas similares que podem ser abstraídas em conjuntos maiores, juntamente com o fato de elas serem avaliadas de acordo com o impacto que elas têm em um jogo.

Desse modo, podemos destacar grupos de atividades de um agente específico que levam mais perigo ao adversário e, portanto, se caracterizam como uma estratégia ou plano ofensivo que deve ser considerado ao se armar uma tática defensiva do time.

4 Demonstração

4.1 Caso 1

A imagem a seguir mostra a distribuição espacial dos passes de Kevin De Bruyne, segmentados por zonas do campo. Cada elipse representa uma região de atuação, indicando

a quantidade de passes realizados e o Z-Score médio associado ao VAEP do cluster correspondente.

O Z-Score revela o impacto relativo dos passes em cada zona: valores positivos (em verde) indicam acima da média, enquanto negativos (em tons alaranjados) indicam impacto inferior. Observa-se, por exemplo, que apesar da alta frequência de passes no setor central ofensivo (368 passes), o impacto médio é negativo (-0,17), enquanto zonas nas laterais e no terço final do campo mostram maior efetividade.

Essa visualização permite identificar padrões táticos, zonas de risco ou destaque e apoiar decisões estratégicas com base em dados.

Rankings do Jogador - Kevin De Bruyne - pass

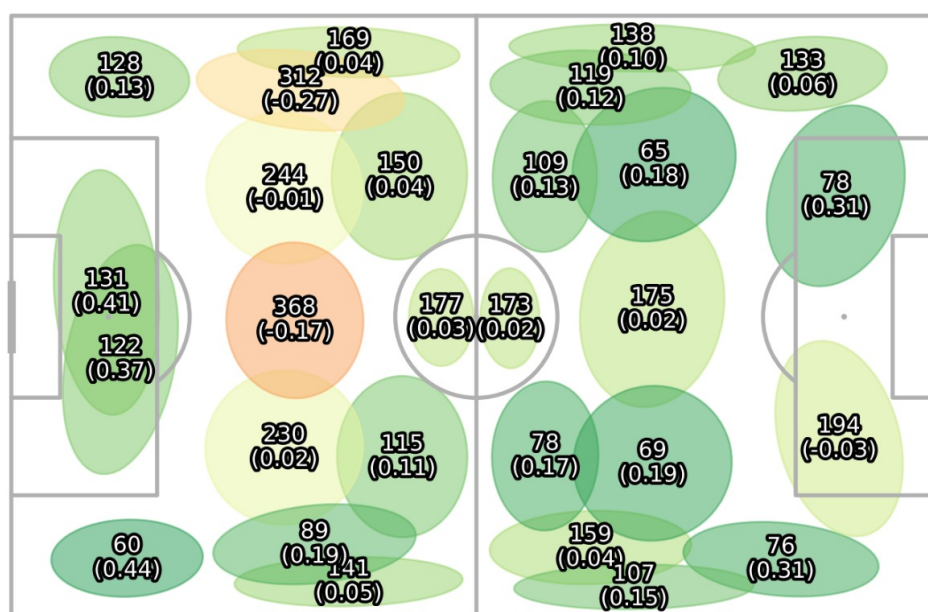


Figure 1: Distribuição dos passes de Kevin De Bruyne com Z-Score em relação ao VAEP do cluster.

4.2 Caso 2

A imagem a seguir apresenta a análise comparativa do clube Liverpool em relação aos demais times da liga, utilizando o VAEP aplicado a passes, segmentado por zonas do campo.

Cada região representa um *cluster* espacial, com coloração indicando o desempenho relativo do Liverpool em comparação com os demais clubes:

- **Verde escuro:** entre os melhores da liga (ranks baixos).
- **Amarelo a laranja:** desempenho mediano.
- **Vermelho:** entre os piores (ranks altos).

Os números indicam a posição do Liverpool no ranking da liga naquele cluster, enquanto as setas representam a direção média dos passes.

Ranking VAEP por Cluster - pass

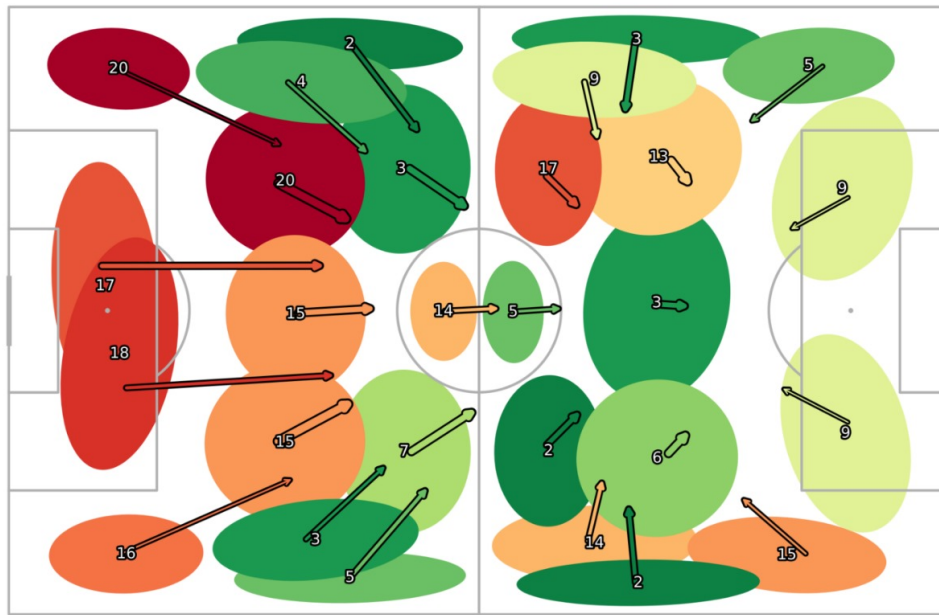


Figure 2: Ranking VAEP por cluster de passes do Liverpool em comparação com os demais times da liga.

A análise mostra que o Liverpool se destaca na construção ofensiva pelo lado direito e setor central do campo adversário (ranks 2, 3 e 5 em zonas verdes). Por outro lado, apresenta desempenho inferior na saída de bola pelo lado esquerdo e em clusters defensivos (ranks 17 a 20 em zonas vermelhas), sugerindo possíveis fragilidades nessas regiões.

Essa visualização fornece uma perspectiva clara sobre os pontos fortes e vulneráveis do modelo de jogo do Liverpool, apoiando decisões táticas e estratégias baseadas em dados.

5 Conclusão

Neste projeto, desenvolveu-se um framework de análise de padrões de estilo de jogo de times de futebol. No decorrer do processo, surgiram desafios como o de compreensão dos dados escolhidos, o de processamento destes, posto o volume disponível, além da tarefa de compreender e aplicar as ferramentas que proporcionaram a análise. Encontrar meios para contornar essas limitações contribuiu grandemente para a evolução e amadurecimento dos integrantes do grupo no âmbito de análise de dados aplicada.

Algumas melhorias podem ser feitas no que diz respeito a evitar viés e o impacto de outliers nos modelos gerados. A escolha das ações analisadas, bem como a filtragem de jogadores por tempo em campo, devem ser revisitadas com o fim de se aprimorar os resultados, garantindo-se maior consistência e confiabilidade nas análises produzidas.

De modo geral, a ferramenta desenvolvida apresentou bons resultados, e, sendo um framework de análise, pode ser empregada para diversos novos experimentos, explorando granularidade de eventos, clusterizações e valorações, por exemplo. Uma possibilidade a se explorar é analisar a frequência com que os eventos ocorrem, adicionando uma dimensão de quantidade associada à qualidade dos eventos. Ademais, outros meios de se

atribuir valor a diferentes ações podem ser aplicados, de modo a se tentar reduzir vieses existentes no VAEP.

References

- [1] Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Actions speak louder than goals: Valuing player actions in soccer. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.
- [2] Tom Decroos, Maaïke Van Roy, and Jesse Davis. Soccermix: Representing soccer actions with mixture models. In Yuxiao Dong, Georgiana Ifrim, Dunja Mladenić, Craig Saunders, and Sofie Van Hoecke, editors, *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track*, pages 459–474, Cham, 2021. Springer International Publishing.
- [3] Luca Pappalardo and Emanuele Massucco. Soccer match event dataset, 2019.
- [4] Gabriel Valadão, João L. L. Gonçalves, João L. Megale, Vinicius M. Paula, Hugo Rios-Neto, Adriano C. M. Pereira, and Wagner Meira Jr. Caracterização de estratégias de futebol com base na frequência, importância e efetividade de jogadas. In *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 1237–1251, Belo Horizonte, MG, Brasil, 2023. Sociedade Brasileira de Computação.

Decidindo o Jogo: Uma Nova Perspectiva sobre Finalizações no Futebol

Bernardo Viggiano, Davi Sakamoto, Gabriel Liberato, Juan Coelho, Theo Lopes

Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)

Resumo. *Este trabalho apresenta um modelo de expected goals (xG) que integra dados de eventos e tracking para avaliar com maior precisão as oportunidades de finalização no futebol. Ao incorporar variáveis contextuais como posicionamento dos atletas, pressão defensiva e antecipação do goleiro, o modelo permite uma avaliação mais justa e detalhada de atacantes e goleiros. Os resultados evidenciam padrões individuais de desempenho e oferecem análises interpretativas únicas, como distinguir finalizações genuinamente perigosas daquelas que aparentam perigo devido ao mau posicionamento do goleiro.*

1. Introdução

No futebol de alto rendimento, compreender o valor real de uma finalização é um desafio central para analistas, treinadores e cientistas de dados. Modelos tradicionais de Expected Goals (xG) estimam a probabilidade de um chute resultar em gol considerando distância, ângulo e tipo de assistência, mas ainda apresentam limitações relevantes. Em especial, quando limitados a dados de eventos, eles tendem a ignorar o contexto tático que envolve a jogada.

Neste trabalho, propomos um modelo de xG que integra dados de eventos com tracking, incorporando variáveis contextuais que capturam melhor a realidade tática. O objetivo é explorar variações do modelo com diferentes seleções de variáveis para analisar aspectos específicos do jogo. Por exemplo, comparando versões que consideram ou não o posicionamento do goleiro, identificamos se uma finalização representa ameaça real ou apenas aparenta perigo devido ao mau posicionamento, análise impossível com xG tradicional.

2. Metodologia

A metodologia adotada neste trabalho foi dividida em cinco etapas principais: compreensão e preparação dos dados, integração e normalização, engenharia de features, modelagem e avaliação de desempenho.

2.1. Compreensão e Preparação dos Dados

O estudo utilizou uma base rica composta por dados de quase 300 partidas completas da Premier League 2023/24, incluindo arquivos de:

- Tracking (posição da bola e dos jogadores em alta frequência temporal),
- Eventos (ações como passes, chutes e faltas),
- Metadata das partidas,
- Escalações e dados dos jogadores.

O processo iniciou-se com o *data understanding*, explorando arquivos, analisando estrutura, granularidade, qualidade e consistência dos dados. A fidedignidade foi verificada por meio da comparação de amostras de jogadas com vídeos reais, permitindo identificar a veracidade de coordenadas e eventos e entender melhor os dados. Devido ao grande volume dos dados de *tracking* (mais de 50 milhões de registros), foram desenvolvidas funções otimizadas para leitura e transformação, reduzindo-os a menos de 30 mil linhas com apenas os frames relevantes às finalizações. Esse pré-processamento incluiu filtragem, seleção de janelas temporais e formatação compatível com aprendizado supervisionado para a modelagem de *Expected Goals*.

2.2. Integração e Normalização dos Dados

Dados de eventos e tracking foram sincronizados baseado nos timestamps. Usando metadados das partidas, foi possível determinar os times mandantes e visitantes e a direção de ataque em cada tempo, permitindo espelhar todas as jogadas para um lado do campo. Por fim, as coordenadas de todas as bases foram normalizadas para dimensão real (campo 105x68 metros), o que foi um diferencial em relação a modelos convencionais para garantir consistência espacial nas features extraídas.

2.3. Engenharia de Features

A etapa de engenharia de features consistiu na criação de um conjunto de variáveis que capturam tanto aspectos objetivos da finalização quanto o contexto tático envolvido. As principais features desenvolvidas incluem:

- **Distância e ângulo em relação ao gol;**
- **Se foi um chute de cabeça ou não;**
- Pé utilizado na finalização;
- Tipo de assistência (passe, cruzamento, rebote etc.);
- **Número de jogadores oponentes e aliados entre a bola e o gol;**
- Presença de defensores a até 1 metro (pressão forte) e a até 2 metros (pressão leve) da bola;
- **Distância entre o ponto do chute e o goleiro;**
- Deslocamento do goleiro em relação ao centro do gol;
- **Altura da bola no momento do chute;**

Features em negrito foram selecionadas para o modelo final após testes iterativos de contribuição e remoção de variáveis que causavam overfitting ou viés ao modelo, priorizando interpretabilidade e robustez.

2.4. Modelagem

Como baseline, utilizamos um modelo de xG com bom desempenho utilizando apenas features disponíveis em dados de eventos, treinado com XGBoost, o método mais amplamente utilizado na literatura para essa tarefa.

Em seguida, passamos à modelagem com dados enriquecidos por tracking. Diversos algoritmos de classificação foram testados (regressão logística, XGBoost, random forest, Explainable Boosting Machine). No fim, o modelo escolhido para a versão final foi o HistGradientBoostingClassifier, por oferecer o melhor desempenho dentre os testes, além de possuir boa interpretabilidade e tempo de treinamento. Hiperparâmetros foram ajustados via validação cruzada, com foco na generalização do modelo.

2.5. Avaliação de Desempenho

A performance do modelo foi avaliada por meio de validação cruzada estratificada. As métricas utilizadas incluíram acurácia, AUC e, principalmente, o Brier Score e a curva de calibração, pois estamos lidando com um problema de classificação probabilística, no qual é essencial não apenas prever corretamente gols e não-gols, mas principalmente atribuir probabilidades bem calibradas a cada finalização. Além do desempenho preditivo, buscamos com o modelo fornecer ferramentas interpretativas.

3. Resultados

3.1. Desempenho dos Modelos

Tabela 1 – Comparação entre algoritmos (event-based + tracking features)

Modelo	Acurácia	AUC	Brier Score	Erro de Calibração
Regressão Logística	88.05%	0.74	0.0941	0.1287
Random Forest	88.18%	0.73	0.0943	0.0983
XGBoost	88.14%	0.72	0.0961	0.0894
HistGradientBoosting	88.16%	0.73	0.0939	0.0516

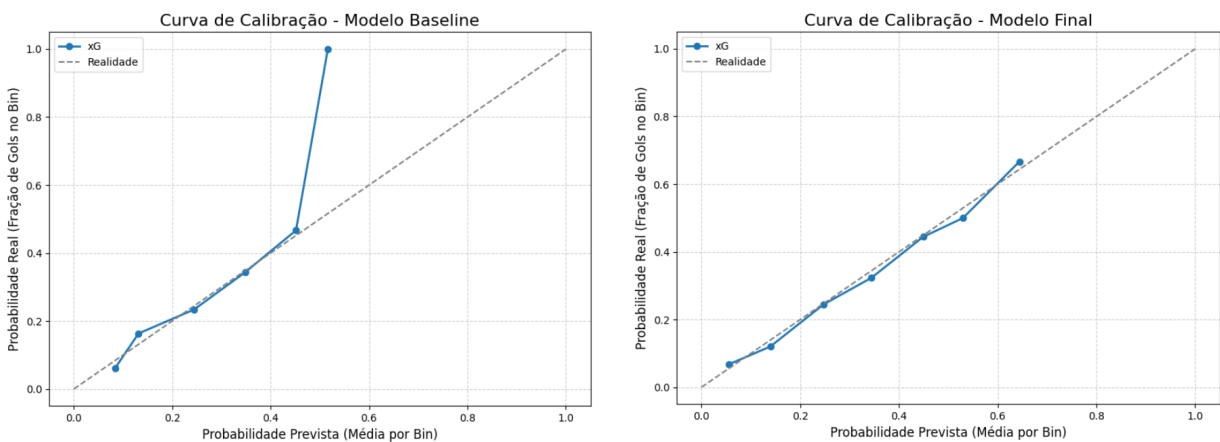
O HistGradientBoosting obteve o melhor desempenho geral e foi selecionado como modelo final.

3.2. Comparação com o Modelo Baseline (Event-Only)

Tabela 2 – Comparação entre baseline e modelo final

Modelo	Acurácia	AUC	Brier Score	Erro de Calibração
Baseline	88.29%	0.70	0.0974	0.1232
Modelo final	88.16%	0.73	0.0939	0.0516

Figuras 1 e 2: Curvas de calibração de modelos baseline e final



Os resultados indicam que a inclusão de informações espaciais detalhadas sobre o posicionamento dos jogadores melhora substancialmente a previsão probabilística da finalização.

3.3. Atacantes

Após aplicar nosso modelo sob todo o dataset, calculamos algumas métricas de análise para os atacantes, como xG total (chances de gol somadas ao longo do campeonato), xG Médio (chance média das oportunidades que recebe), gols adicionados (quantos gols o jogador fez a mais que o esperado) e taxa de conversão (quanto um jogador consegue aproveitar de uma chance).

Figura 3: Os 10 atacantes que mais tiveram chances de gol

	nickname	total_chutes	gols_marcados	xg_total	xg_medio	gols_adicionados	taxa_conversao
0	Ollie Watkins	93	19	17.240977	0.185387	1.759023	1.102026
1	Erling Haaland	86	17	15.815001	0.183895	1.184999	1.074929
2	Dominic Solanke	82	15	14.943263	0.182235	0.056737	1.003797
3	Darwin Nunez	93	12	13.474193	0.144884	-1.474193	0.890591
4	Richarlison	64	13	11.616434	0.181507	1.383566	1.119104
5	Nicolas Jackson	59	12	11.262904	0.190897	0.737096	1.065445
6	Heung-min Son	68	18	11.047171	0.162458	6.952829	1.629376
7	Mohamed Salah	69	13	10.475003	0.151812	2.524997	1.241050
8	Bukayo Saka	77	10	9.589521	0.124539	0.410479	1.042805
9	Luis Díaz	70	9	9.437033	0.134815	-0.437033	0.953690

Figura 4: Atacantes com mais e menos gols adicionados

	nickname	total_chutes	gols_marcados	xg_total	xg_medio	gols_adicionados	taxa_conversao
0	Heung-min Son	68	18	11.047171	0.162458	6.952829	1.629376
1	Jarrod Bowen	67	14	8.549817	0.127609	5.450183	1.637462
2	Chris Wood	27	10	5.228390	0.193644	4.771610	1.912635
3	Hee-chan Hwang	36	10	6.130569	0.170294	3.869431	1.631170
4	Alexander Isak	42	11	7.221618	0.171943	3.778382	1.523204
...	...	...	...	...	...	...	...
441	Frank Onyeka	22	0	2.111338	0.095970	-2.111338	0.000000
442	Dominic Calvert-Lewin	55	5	7.131747	0.129668	-2.131747	0.701090
443	Ryan Christie	30	0	2.171376	0.072379	-2.171376	0.000000
444	Morgan Gibbs-White	52	2	4.305227	0.082793	-2.305227	0.464552
445	Bruno Fernandes	68	2	5.066118	0.074502	-3.066118	0.394780

Figura 5: Atacantes com maiores e menores taxas de conversão (sem filtros)

	nickname	total_chutes	gols_marcados	xg_total	xg_medio	gols_adicionados	taxa_conversao
0	Hannibal Mejbri	1	1	0.043888	0.043888	0.956112	22.785511
1	Micky van de Ven	4	1	0.120023	0.030006	0.879977	8.331719
2	Ola Aina	6	1	0.148334	0.024722	0.851666	6.741550
3	Emerson Royal	5	1	0.201603	0.040321	0.798397	4.960241
4	Carney Chukwuemeka	2	1	0.232558	0.116279	0.767442	4.300004

Figura 6: Atacantes com maiores taxas de conversão (com ao menos 30 chutes)

	nickname	total_chutes	gols_marcados	xg_total	xg_medio	gols_adicionados	taxa_conversao
50	Diogo Jota	37	9	5.333573	0.144151	3.666427	1.687424
52	Simon Adingra	31	6	3.587861	0.115737	2.412139	1.672305
54	Declan Rice	40	6	3.643254	0.091081	2.356746	1.646879
55	Jarrod Bowen	67	14	8.549817	0.127609	5.450183	1.637462
56	Hee-chan Hwang	36	10	6.130569	0.170294	3.869431	1.631170

3.4. Goleiros

Além das estatísticas de atacantes, também calculamos algumas métricas de análise para os goleiros, como xG total (chances de levar gol somadas ao longo do campeonato), xG Médio (perigo médio das finalizações que recebe), gols adicionados (gols salvos em relação ao total esperado) e taxa de defesa (quanto um goleiro é capaz de defender uma situação de perigo). Além disso, aplicamos o modelo sem o posicionamento do goleiro para avaliar como isso influenciou nos resultados.

Figura 7: Goleiros com mais e menos gols salvos

	nickname	total_chutes	gols_tomados	xg_total	xg_medio	gols_salvos	taxa_defesa
0	Andre Onana	461	36	47.572187	0.103193	11.572187	1.321450
1	Bernd Leno	411	40	51.184684	0.124537	11.184684	1.279617
2	Alphonse Areola	411	35	43.631386	0.106159	8.631386	1.246611
3	Alisson	250	21	29.134135	0.116537	8.134135	1.387340
4	Guglielmo Vicario	397	44	50.628905	0.127529	6.628905	1.150657
32	Emiliano Martínez	322	50	46.210791	0.143512	-3.789209	0.924216
34	James Trafford	435	60	55.269320	0.127056	-4.730680	0.921155
35	Dean Henderson	114	20	14.474029	0.126965	-5.525971	0.723701
36	Mark Flekken	413	59	52.176290	0.126335	-6.823710	0.884344
37	Wes Foderingham	435	64	55.678078	0.127996	-8.321922	0.869970

Figura 8: Goleiros com maiores e menores taxas de defesas (com ao menos 50 chutes contra)

	nickname	total_chutes	gols_tomados	xg_total	xg_medio	gols_salvos	taxa_defesa
1	Alisson	250	21	29.134135	0.116537	8.134135	1.387340
2	Aaron Ramsdale	55	5	6.611277	0.120205	1.611277	1.322255
3	Andre Onana	461	36	47.572187	0.103193	11.572187	1.321450
4	Bernd Leno	411	40	51.184684	0.124537	11.184684	1.279617
5	Alphonse Areola	411	35	43.631386	0.106159	8.631386	1.246611
30	Matz Sels	81	11	9.436859	0.116504	-1.563141	0.857896
31	Lukasz Fabianski	60	9	7.551927	0.125865	-1.448073	0.839103
33	Odysseas Vlachodimos	62	11	8.613875	0.138933	-2.386125	0.783080
34	Ivo Grbic	73	11	8.369380	0.114649	-2.630620	0.760853
35	Dean Henderson	114	20	14.474029	0.126965	-5.525971	0.723701

Figura 9: Quantos gols os goleiros evitaram ou tomaram apenas com o posicionamento

	nickname_x	gols_tomados	gols_evitados	taxa_posicionamento
0	Andre Onana	36	3.856084	10.711346
1	Alphonse Areola	35	2.993620	8.553199
2	Nick Pope	14	1.820701	13.005010
3	Matt Turner	28	1.130669	4.038105
4	Guglielmo Vicario	44	1.088173	2.473120
33	Alisson	21	-1.590694	-7.574733
34	Thomas Kaminski	56	-2.213440	-3.952571
35	Emiliano Martínez	50	-2.421872	-4.843744
36	Wes Foderingham	64	-3.279833	-5.124740
37	Bernd Leno	40	-4.388021	-10.970052

4. Conclusão

O modelo desenvolvido integrou dados de eventos e tracking para análise mais rica das finalizações no futebol. Embora os ganhos sobre o baseline tenham sido modestos, a incorporação de variáveis contextuais contribuiu positivamente para calibração das probabilidades e geração de insights relevantes.

O principal valor está na capacidade de gerar interpretações mais ricas, evidenciando situações que escapam à análise superficial, como finalizações aparentemente perigosas que foram menos eficazes devido ao bom posicionamento defensivo. Este trabalho contribui para o debate sobre dados contextuais no futebol, servindo como ponto de partida para aprimoramentos futuros e ferramentas de apoio à decisão em contextos competitivos.

5. Referências

- [1] ENSUM, J.; POLLARD, R.; TAYLOR, S. *Applications of logistic regression to shots at goal in association football: calculation of shot probabilities, quantification of factors and player/team*. Journal of Sports Sciences, 2004.
- [2] CAVUS, M.; BIECEK, P. *Explainable expected goal models for performance analysis in football analytics*. arXiv preprint arXiv:2206.07212, 2022.
- [3] ANZER, G.; BAUER, P. *A goal scoring probability model for shots based on synchronized positional and event data in football*, 2021.
- [4] POLLARD, R.; ENSUM, J.; TAYLOR, S. *Estimating the probability of a shot resulting in a goal: The effects of distance, angle and space*. International Journal of Soccer and Science, 2004.
- [5] TAHIR, S. B. U. *A comparative and statistical analysis of xG to predict football shot outcomes in event data*. In: *Statistics for Innovation I*, 2025.
- [6] BASSEK, M. et al. *An integrated dataset of spatiotemporal and event data in elite soccer*. Scientific Data, v. 12, p. 195, 2025.
- [7] UMAMI, T. et al. *Toward interpretable expected goals modeling using Bayesian methods*. Frontiers in Sports and Active Living, 2025.
- [8] VAN HAAREN, J. *Analyzing performance and playing style using ball event data*, 2019.
- [9] ASSESSING the Impact of Tracking Data on Passing Metrics in Football, 2023.

Entendendo Posicionismo e Relacionismo através de Métricas de Centralidade

1

1. Introdução

A análise de desempenho no futebol tem se aprimorado pela utilização de novas ferramentas computacionais, como a análise de redes complexas e a análise de dados de eventos em larga escala [Uchoa Maia Rodrigues et al. 2019]. No contexto do esporte, essas técnicas permitem uma percepção mais refinada dos padrões de jogo, ajudando a identificar interações de dependências entre diferentes atletas e setores do campo [Arriaza-Ardiles et al. 2018].

Este trabalho explora como essas técnicas podem ser aplicadas para investigar duas filosofias táticas amplamente empregadas e discutidas: o posicionismo e o relacionismo. O posicionismo é caracterizado por uma definição muito clara das zonas que devem ser ocupadas pelos jogadores em campo, buscando uma estrutura mais estável e previsível. Por outro lado, o relacionismo prevê uma maior liberdade de posicionamento para os atletas, com o intuito de criar conexões e padrões de jogo mais dinâmicos através da movimentação.

Com o intuito de realizar distinções entre esses modelos de jogo, a metodologia é baseada na análise de redes de passes [Immler et al. 2021]. As análises terão como foco treinadores renomados no futebol mundial, sendo alguns deles: Pep Guardiola (Bayern de Munique; 2015/2016) e Lionel Scaloni (Argentina; 2022).

O objetivo é demonstrar que o emprego de métricas de centralidade pode fornecer evidências quantitativas sobre essas duas filosofias de jogo. Assim, busca-se estabelecer uma conexão entre esses dois conceitos táticos (posicionismo e relacionismo) e suas manifestações mensuráveis nos padrões de jogo das equipes.

2. Metodologia

Os dados utilizados para as análises foram acessados através da base da StatsBomb¹. Essa fonte fornece dados de eventos que descrevem de forma detalhada a maior parte dos acontecimentos de um jogo de futebol. Para este estudo, o tipo de evento mais relevante é o passe.

Para analisar a estrutura de interações de cada equipe, foram construídas redes de passes, representadas como grafos. Um dos indicadores mais importantes em uma rede de passes é o valor da centralidade de um vértice. A medida de centralidade pode ser interpretada como a relevância de um determinado vértice em função de uma característica selecionada para a rede. Existem várias medidas de centralidade, como: *closeness*, *betweenness* e *coeficiente de clustering* [Barabási and Pósfai 2016]. Cada uma dessas medidas possui uma interpretação diferente no contexto das redes de passes no futebol.

¹<https://statsbomb.com>

A análise de centralidade neste trabalho não foi realizada com foco em jogadores individuais, mas sim em zonas específicas do campo. Para isso, o campo foi dividido de acordo com uma abordagem tática baseada em zonas, frequentemente utilizada na análise do futebol. A Figura 1 ilustra como é feita essa divisão no campo de futebol.

Neste trabalho, o cálculo dos valores das métricas de centralidade foram realizadas com base nessa divisão em zonas, com o intuito de quantificar a relevância de cada uma dessas regiões de um campo de futebol sob diferentes características para cada um dos treinadores avaliados. Além disso, desassociar a análise dos atletas para focar em zonas do campo é uma alternativa para isolar o sistema tático da influência das características e qualidades individuais dos jogadores, permitindo que o foco seja o estilo de jogo proposto pelo técnico.



Figura 1. Divisão tática do campo em 20 zonas.

3. Resultados

Para o trabalho proposto, medidas de centralidade em grafos foram utilizadas para obter evidências quantificáveis a respeito das filosofias de jogo citadas. Com o intuito de direcionar a análise, esta seção de interpretação das métricas terá como foco os treinadores que mais representam cada uma das filosofias. Assim, o Bayern de Munique (15/16) de Pep Guardiola será o expoente do posicionismo, enquanto a Argentina (2022) de Lionel Scaloni será o símbolo do futebol relacional.

Treinador	Betweenness (Média ± DP)	Closeness (Média ± DP)	Clustering Coeff. (Média ± DP)	Densidade
Josep Guardiola	0.0190 ± 0.0156	0.7639 ± 0.1129	0.7914 ± 0.0715	0.7289
Jürgen Klopp	0.0213 ± 0.0162	0.7407 ± 0.1098	0.7601 ± 0.0628	0.7000
Lionel Scaloni	0.0288 ± 0.0222	0.6746 ± 0.1004	0.7160 ± 0.0936	0.5895
Xabi Alonso	0.0203 ± 0.0158	0.7497 ± 0.1070	0.7931 ± 0.0817	0.7132
Zinedine Zidane	0.0219 ± 0.0174	0.7368 ± 0.1112	0.7708 ± 0.0630	0.695
Claudio Ranieri	0.0221 ± 0.0193	0.7358 ± 0.1125	0.7424 ± 0.0614	0.689

Tabela 1. Comparativo de Métricas de Rede de Passes para Diferentes Treinadores

A Tabela 1 mostra os treinadores analisados e as respectivas métricas de centralidade escolhidas neste estudo. A métrica *betweenness* quantifica a importância de uma zona pela quantidade de vezes que ela está no caminho mais curto de passes entre outras duas zonas. *Closeness* é uma métrica que mede o quão próximas as zonas estão uma das outras. Uma zona com alto *closeness* possui uma curta distância média para todas as outras zonas. O coeficiente de *clustering* indica a capacidade das zonas formarem triângulos entre si. Por fim, a densidade é um indicador de coesão entre as zonas. Quando a densidade é baixa, maior é a probabilidade da rede ser dividida em grupos independentes, ao invés de apenas uma rede interconectada.

A partir da análise das métricas de centralidade calculadas, é possível tirar conclusões sobre o estilo de jogo dos times e relacioná-las com a filosofia posicional ou relacional. O Bayern de Munique, treinado por Josep Guardiola, por exemplo, apresenta um baixo valor de *betweenness*. O valor dessa métrica indica que a equipe alemã ocupa as suas zonas de forma homogênea (sem zonas muito mais relevantes que outras para a capacidade de associação da equipe), uma característica simbólica do posicionismo. Em contrapartida, o treinador Lionel Scaloni possui um *betweenness* muito maior. Isso é um indício de que certas zonas são muito mais ocupadas e envolvidas na troca de passes do que outras, qualificando o estilo de jogo relacional.

Seguindo a análise das métricas e sua relação com as zonas, um *closeness* alto indica que, com poucos passes, é possível transportar a bola para outras zonas, ou seja, a rede está intensamente ligada homogeneamente pelas zonas. Continuando com o exemplo de Guardiola, é possível observar que sua equipe possui um alto *closeness*, pois as zonas são altamente interligadas e os passes distribuídos homogeneamente no campo, uma característica típica do posicionismo. Por outro lado, Lionel Scaloni possui um valor baixo de *closeness*, pois os jogadores frequentemente se aglomeram em certas regiões do campo, gerando o isolamento das zonas mais distantes da posição da bola e, consequentemente, uma menor capacidade de transportar a bola rapidamente para essas zonas.

Outra métrica avaliada foi o coeficiente de *clustering*. O seu baixo valor para a Argentina de Scaloni pode ser explicado pela alta proximidade entre os jogadores. Por causa dessa disposição espacial, é muito provável que ocorram sequências de passes que envolvam apenas uma ou duas zonas, o que não favorece a formação dos triângulos que caracterizam essa métrica. Para a equipe de Guardiola, o valor mais alto do coeficiente é consistente com a ocupação mais homogênea das zonas, com os jogadores mantendo uma distância maior entre si. Essa configuração possibilita a formação de triângulos entre zonas com mais frequência.

A equipe de Pep Guardiola também possui uma densidade mais alta, já que um jogador que ocupa uma zona possui opções de passe em diversas zonas do campo, tendo como consequência a criação de fortes conexões entre as zonas. No caso da Argentina relacionista de Scaloni, o valor de densidade é menor porque os jogadores são incentivados a gerar superioridade numérica nas zonas mais próximas à posição do portador da bola. Isso faz com que uma zona tenha menos possibilidades de conexões diretas, especialmente em comparação à filosofia posicional.

Apesar do foco em apenas dois treinadores para um melhor direcionamento da seção de interpretação das métricas, foi possível notar que a maior parte dos treinado-

res europeus (dentro do escopo dos dados disponíveis) possuem métricas de centralidade mais próximas às apresentadas pelo treinador Josep Guardiola. Essa constatação revela a grande influência desse técnico para o desenvolvimento tático do futebol mundial. Por outro lado, com a crescente adoção das ideias de jogo relacional nos últimos anos, especialmente no futebol sul-americano, é esperado que (analisando dados mais recentes e de mais ligas) sejam encontrados mais treinadores de elite com métricas próximas ao Lionel Scaloni.

4. Conclusão

Este trabalho demonstrou que a aplicação de métricas de centralidade em redes de passes zonais é uma abordagem eficaz para identificar e quantificar as diferenças entre as filosofias de jogo do posicionismo e do relacionismo, estabelecendo uma conexão clara entre os princípios táticos e suas manifestações nos dados.

Porém, o estudo encontrou limitações devido à escassez de dados abertos de partidas recentes e de ligas não europeias. Essa restrição é particularmente desafiadora, uma vez que os maiores expoentes da filosofia relacional, como Fernando Diniz no Fluminense e Renato Gaúcho no Grêmio, consolidaram seus trabalhos na última década, fora da cobertura de dados do futebol europeu.

Para trabalhos futuros, pretende-se complementar a metodologia apresentada com novas ideias, como a análise de ciclos de posse de bola. Essa ampliação pode tornar o estudo mais robusto, tanto pela adição de novas dimensões espaciais como pela introdução de dimensões temporais.

Referências

- Arriaza-Ardiles, E., Martín-González, J., Zuniga, M., Sánchez-Flores, J., de Saa, Y., and García-Manso, J. (2018). Applying graphs and complex networks to football metric interpretation. *Human Movement Science*, 57:236 – 243.
- Barabási, A.-L. and Pósfai, M. (2016). *Network science*. Cambridge University Press.
- Immler, S., Rappelsberger, P., Baca, A., and Exel, J. (2021). Guardiola, klopp, and po-chettino: the purveyors of what? the use of passing network analysis to identify and compare coaching styles in professional football. *Frontiers in Sports and Active Living*, 3:725554.
- Uchoa Maia Rodrigues, D. C., Moura, F. A., Cunha, S. A., and da S. Torres, R. (2019). Graph visual rhythms in temporal network analyses. *Graphical Models*, 103:101021.

Otimização Logística no Futebol Brasileiro: Uma Análise do Impacto das Viagens no Desempenho dos Times e Proposta de Calendário Ajustado

Paulo Henrique Sendas Resende¹, André Luiz Braga Jacinto³,
Daniel De Oliveira Capanema¹, Roney Lira de Sales Santos²

¹ Departamento de Ciência da Computação,
Pontifícia Universidade Católica de Minas Gerais,
Belo Horizonte – MG – Brasil

² Instituto de Computação, Universidade Federal da Bahia, Salvador – BA – Brasil

³ Universidade Federal de Minas Gerais, Belo Horizonte – MG – Brasil

paulo.resende@sga.pucminas.br, daniel.capanema@sga.pucminas.br

andreluiz.braga@hotmail.com, roneysantos@ufba.br

Abstract. *The Brazilian national football calendar forces clubs to traverse continental distances, imposing cumulative travel fatigue on athletes. Based on GPS and heart rate data from Time A's first-team squad in 2018, we analyzed match-day metrics—total distance covered, high-intensity running distance, TRIMP, and ratings of perceived exertion (RPE)—correlating each away match with the actual travel distance. Trips were categorized as short (<500), medium (500–1000 km), or long (>1000 km). During long trips, average TRIMP and RPE were slightly higher than other categories. To investigate whether the calendar itself could be optimized, we modeled the problem as a Mirrored Traveling Tournament Problem and applied a Simulated Annealing algorithm to minimize total travel distance while penalizing sequences of more than two consecutive home or away matches. Starting from a polygonal calendar, the algorithm generated a feasible 20-team fixture in 4 seconds, reducing total travel from 709,054 km to 543,780 km (23.3%). Long trips decreased from 79 to 59, while short trips increased by 50, particularly benefiting remote clubs (Sport: 17,762 km; Ceará: 15,472 km).*

By aligning physiological evidence with combinatorial optimization, this study demonstrates that logistical calendar restructuring can mitigate athlete fatigue and promote greater competitive equity in Brazilian football.

Resumo. *O calendário nacional do futebol brasileiro obriga os clubes a percorrer distâncias continentais, impondo fadiga de viagem acumulada aos atletas. Com base nos dados de GPS e frequência cardíaca do elenco principal do Time A em 2018, analisamos, nos dias de jogo, distância total, distância em alta intensidade, TRIMP e percepção subjetiva de esforço (PSE), relacionando cada partida fora de casa ao percurso efetivamente percorrido. As viagens foram categorizadas como curtas (<500) médias (500–1000 km) ou longas (>1000 km). Nos deslocamentos longos observou-se TRIMP médio e PSE ligeiramente superiores às demais categorias. Para investigar se o próprio calendário poderia ser*

otimizado, formulamos o problema como um Mirrored Traveling Tournament Problem e aplicamos um Simulated Annealing que minimiza a quilometragem total, penalizando sequências de mais de duas partidas consecutivas em casa ou fora. Partindo de um calendário poligonal, o algoritmo gerou, em cerca de 4 s, uma tabela viável para 20 equipes, reduzindo a soma dos deslocamentos de 709 054 km para 543 780 km (-23,3 %). O número de viagens longas caiu de 79 para 59 e o de curtas aumentou em 50, beneficiando especialmente clubes mais distantes (Sport: -17 762 km; Ceará: -15 472 km). Ao alinhar evidências fisiológicas com otimização combinatória, o estudo demonstra que a reorganização logística do calendário pode atenuar o desgaste dos atletas e promover maior equidade competitiva no futebol brasileiro.

1. Introdução

Desde que o Campeonato Brasileiro adotou o sistema de pontos corridos, em 2003, nenhuma equipe situada fora da Região Sudeste ergueu a taça. Curiosamente, essa hegemonia não se repete em competições eliminatórias, disputadas pelos mesmos clubes porém com menos jogos: Internacional e Grêmio, do Rio Grande do Sul, conquistaram a Copa Libertadores em 2006, 2010 e 2017, enquanto Grêmio, Sport (PE) e Athletico Paranaense venceram a Copa do Brasil nesse mesmo período (Transfermarkt 2025a, Transfermarkt 2025c, Transfermarkt 2025b). Esse contraste sugere que o calendário extenso — com longas viagens inter-regionais distribuídas de forma desigual — pode impor um ônus competitivo adicional às equipes periféricas no torneio nacional por pontos corridos.

Viagens sucessivas impõem aos atletas um tipo específico de desgaste, denominado fadiga por deslocamento. Quando envolvem mudanças de fuso horário, esse efeito é agravado pelo fenômeno conhecido como *jet lag*. A literatura reconhece que tais condições ampliam a fadiga, aumentam o risco de lesões e comprometem o desempenho físico dos jogadores (Silva et al. 2021). Pesquisas em monitoramento de carga no futebol indicam que métricas internas — baseadas na frequência cardíaca e no estresse cardiovascular — são mais eficazes para prever a percepção subjetiva de esforço do que variáveis externas, como a distância percorrida (Silva et al. 2024).

Entre as principais métricas internas figuram o TRIMP (*Training Impulse*), que combina intensidade e duração dos estímulos via frequência cardíaca, e a Percepção Subjetiva de Esforço (PSE), uma medida qualitativa que integra fatores fisiológicos, psicológicos e sensoriais. Num país continental como o Brasil, onde grandes deslocamentos são rotina, compreender o impacto das viagens sobre TRIMP e PSE é essencial para projetar calendários mais sustentáveis.

A organização do cronograma pode ser modelada como um Problema do Torneio Viajante Espelhado (mTTP), visando minimizar a distância total percorrida sob restrições de mando e espelhamento de turnos. Como até mesmo versões relaxadas são NP-difíceis (Easton and Nemhauser 2001), métodos exatos (programação inteira, *branch-and-bound*) só resolvem instâncias muito pequenas; para um campeonato de 20 clubes, recorremos a meta-heurísticas.

Este estudo busca alcançar dois objetivos: (i) avaliar se a extensão das viagens de uma equipe — classificadas em curtas (< 500 km), médias (500–1000 km) e lon-

gas (> 1000 km) — se relaciona a variações em TRIMP e PSE nos dias de jogo; e (ii) investigar se um calendário reorganizado por *Simulated Annealing* (SA), respeitando no máximo duas partidas consecutivas no mesmo mando, pode mitigar tais efeitos. Para atender a esses objetivos, foi disponibilizada a base de dados de uma equipe profissional de futebol, doravante denominada “**Time A**” por questões de confidencialidade, referente à temporada de 2018.

Nos tópicos seguintes apresentam-se metodologia, resultados e implicações para a competitividade do futebol brasileiro.

2. Referencial Teórico: Impacto das Viagens no Desempenho Físico em Jogos Fora de Casa

No futebol profissional, fatores situacionais, como jogar em casa ou fora, têm reconhecido impacto nas demandas físicas e no desempenho dos atletas (Sampaio et al. 2019). Jogos fora de casa frequentemente envolvem deslocamentos e adaptações a ambientes desconhecidos, influenciando significativamente o rendimento físico durante as partidas (Draper et al. 2024). Este fator é especialmente relevante no contexto deste projeto, que analisa a influência das viagens nos jogadores do Time A durante a temporada de 2018. Pesquisas recentes têm buscado quantificar diretamente esse impacto, utilizando métricas fisiológicas semelhantes às coletadas nesta pesquisa, como dados de GPS e estatísticas de jogo.

Draper et al. (2024) destacam que equipes que acumulam grandes distâncias em viagens durante a temporada tendem a apresentar maiores volumes de corrida em alta intensidade durante jogos fora de casa. No entanto, essa maior intensidade não necessariamente melhora o desempenho competitivo; pelo contrário, observa-se uma leve redução nos pontos obtidos e aumento no número de gols sofridos pelos times visitantes. Esses resultados corroboram a hipótese do presente estudo, sugerindo que a fadiga gerada por viagens prolongadas pode obrigar os jogadores a aumentarem o esforço físico em campo, porém com eficiência reduzida (Draper et al. 2024).

Além disso, a modalidade e duração da viagem imediatamente antes da partida influenciam diretamente o desempenho dos atletas. Hands et al. (2023) mostram que, no futebol australiano (A-League), jogos fora de casa resultaram em um padrão diferente de esforços físicos em comparação aos jogos em casa, com maior volume de corridas em baixa intensidade fora de casa. Viagens terrestres realizadas no dia do jogo foram associadas a uma redução nas ações de alta intensidade, provavelmente devido à rigidez muscular e à fadiga acumulada durante o trajeto. Por outro lado, viagens aéreas mais longas feitas com maior antecedência mostraram manter níveis semelhantes de intensidade aos jogos realizados em casa (Kruyt et al. 2023). Esses achados reforçam a relevância do presente estudo em analisar precisamente como diferentes distâncias percorridas e métodos de viagem influenciaram os jogadores do Time A, permitindo assim sugerir otimizações logísticas concretas para o calendário competitivo.

Para propor intervenções logísticas concretas, este projeto utiliza como base o *Traveling Tournament Problem* (TTP), formalizado por Easton, Nemhauser e Trick (2001). Este modelo abstrai desafios práticos na criação de calendários esportivos que minimizem

distâncias totais percorridas, respeitando restrições de mandos consecutivos e frequência de jogos (Easton and Nemhauser 2001). Todavia, por ser NP-difícil, o TTP demanda soluções engenhosas em instâncias maiores — como o Campeonato Brasileiro, com 20 equipes. Embora métodos exatos (como programação inteira e técnicas de branch-and-bound) ofereçam garantia de otimalidade (Bertsimas et al. 2011), seu sucesso é limitado pela complexidade computacional, tornando-os inviáveis em escalas reais.

Diante desse desafio, adotamos o *Simulated Annealing* (SA) como núcleo metodológico. Ancorado na termodinâmica do recozimento de metais (Kirkpatrick et al. 1983), o SA opera com uma função de custo que quantifica distâncias totais e violações de restrições. Sua aplicação ao TTP é particularmente eficaz, como demonstrado por Anagnostopoulos et al. 2006, que alcançou reduções de até 25% nas distâncias de viagem em instâncias complexas do problema, superando abordagens convencionais. O mecanismo central é a regra de Metropolis — aceitação probabilística de soluções vizinhas via $\exp(-\Delta E/T)$ que permite escapar de ótimos locais.

Dessa forma, este referencial teórico fundamenta diretamente o objetivo deste projeto, que visa avaliar o impacto real das viagens nas métricas fisiológicas dos atletas do Time A e verificar como um calendário otimizado, utilizando *Simulated Annealing*, pode reduzir fadiga acumulada, melhorar a recuperação e potencialmente aumentar o desempenho competitivo ao longo da temporada.

3. Metodologia

Este estudo caracteriza-se como uma pesquisa de natureza quantitativa e descritivo-comparativa, tendo como objetivo examinar a relação entre viagens e desempenho físico em partidas de futebol profissional. Para tanto, serão utilizados dois conjuntos principais de dados: (a) métricas de GPS coletadas dos atletas ao longo de toda a temporada do Time A na temporada de 2018; e (b) informações do calendário de partidas e resultados obtidas no *Football Reference*. A seguir, detalham-se as etapas de coleta, limpeza, integração e análise de dados.

3.1. Coleta e Organização dos Dados

- **Dados de GPS:** Foram extraídas informações referentes a volume de corrida (distância total percorrida), distância em alta velocidade, número de *sprints*, acelerações e desacelerações de cada atleta, além de outras variáveis fisiológicas disponibilizadas pelo sistema de rastreamento. Esses dados foram coletados durante as sessões de treinamento e durante os jogos realizados pela equipe ao longo da temporada, tendo sido disponibilizados pelo fisiologista do time.
- **Calendário e Resultados das Partidas:** O calendário oficial, contendo datas, horários, adversários e local das partidas (casa ou fora), foi obtido no site *Football Reference*. Adicionalmente, foram coletados os resultados dos jogos (vitória, empate ou derrota), bem como outros indicadores de desempenho (saldo de gols, posse de bola etc., caso disponíveis), que serviram de base para a correlação com as variáveis de GPS.
- **Distâncias de Viagem:** Para cada partida fora de casa, calculou-se a distância geodésica entre a cidade onde a equipe se encontrava imediatamente antes do jogo e a cidade do estádio adversário. Quando a sequência do calendário prevê dois ou

mais jogos seguidos como visitante, a origem da medição passa a ser a cidade do jogo anterior, e não a sede do clube; o retorno à cidade-sede só é considerado quando o time voltar a atuar como mandante. Após esse cálculo ponto a ponto, as distâncias de cada deslocamento foram classificadas em faixas de 300–500 km, 500–700 km, 700–1000 km e acima de 1000 km, conforme o delineamento da pesquisa.

3.2. Procedimentos de Limpeza e Tratamento dos Dados

- **Verificação de Consistência:** Inicialmente, proceder-se-á à conferência de consistência e completude das variáveis, removendo duplicatas e registros eventualmente corrompidos. Como os dados de GPS podem conter leituras errôneas (por perda de sinal ou erro de registro), serão estabelecidos critérios para exclusão de valores fora de faixas plausíveis (ex.: velocidades incompatíveis com o deslocamento humano).
- **Integração com o Calendário de Jogos:** Após a limpeza, será construída uma base consolidada com as variáveis de GPS referentes ao período que antecede os jogos (sessões pré-partida), ao dia da partida e ao período subsequente (para análise de recuperação).
- **Filtramos o conjunto de dados,** apenas jogadores titulares e excluimos goleiros. Em seguida calculamos a distância de viagem usando a matriz de distâncias 2018, mapeando o nome do oponente para a coluna correspondente na matriz. Por fim categorizamos as viagens em Curta (< 500), Média (500–1000 km) e Longa (> 1000 km).

3.3. Definição de Variáveis e Indicadores

O conjunto de indicadores adotado divide-se em duas dimensões complementares. A *carga externa* será mensurada pelas métricas obtidas via GPS — distância total percorrida, distância em alta intensidade e número de *sprints* —, que descrevem o volume e a intensidade mecânica das ações em campo. Já a *carga interna* abrangerá, de um lado, o **PSE** (Percepção Subjetiva de Esforço), que traduz a sensação de fadiga reportada pelo atleta após cada sessão, e, de outro, o **TRIMP** (Training Impulse), índice fisiológico que combina frequência cardíaca e duração do exercício para estimar o estresse cardiovascular acumulado.

3.4. Análise de Dados

- **Análises Descritivas:** Inicialmente, foram calculadas estatísticas descritivas (média, desvio-padrão, valores mínimos e máximos) das variáveis de GPS para cada condição de jogo (casa vs. fora), bem como para cada faixa de distância percorrida.
- **Comparações Entre Grupos:** Para analisar a influência das viagens, foram realizadas comparações das métricas de desempenho físico em função das faixas de distância. Definiu-se a categoria de viagem para jogos fora segundo a distância entre Belo Horizonte e a cidade adversária: viagens curtas ($< 500km$), médias (500–1000 km) e longas ($> 1000km$)

3.5. Justificativa da Escolha do Simulated Annealing

A escolha do algoritmo *Simulated Annealing* (SA) neste estudo justifica-se pela sua eficiência, flexibilidade e adequação prática ao problema do Torneio Viajante Espelhado (mTTP) com 20 equipes. Dada a natureza combinatória e NP-difícil do problema, métodos exatos como Programação Inteira e *Branch-and-Bound* tornam-se inviáveis para instâncias de tamanho real, devido ao crescimento exponencial do tempo de processamento.

Esta técnica - uma busca estocástica inspirada no recozimento de metais, em que o material é aquecido a um estado de alta energia e então resfriado lentamente até atingir configuração estável (Kirkpatrick et al. 1983) — ajusta-se de modo particularmente elegante ao problema de otimizar o calendário de um campeonato, formalizado como o *Mirrored Traveling Tournament Problem* (mTTP). Nesse contexto, cada estado S representa uma tabela completa de jogos sujeita a restrições de mandos consecutivos, janelas de turno-retorno e limites de distância; a “temperatura” T controla a aceitação de calendários provisoriamente piores que aumentam o custo total de viagens ou violam levemente as regras, permitindo ao algoritmo escapar de arranjos locais e explorar combinações de rodadas mais promissoras. À medida que T diminui, o SA concentra-se em soluções que minimizam a soma das distâncias percorridas por todas as equipes, produzindo cronogramas factíveis que se aproximam do ótimo conhecido em segundos para instâncias reais da Série A com 20 clubes (Urrutia and Ribeiro 2006). Dessa forma, o recozimento simulado transforma a metáfora física de resfriamento controlado em um método eficaz para alinhar logística esportiva e desempenho competitivo dentro de calendários viáveis.

Entre as principais vantagens do SA, destacam-se:

- **Velocidade:** encontra boas soluções em poucos segundos, mesmo para instâncias grandes ($n = 20$ ou mais) (Urrutia and Ribeiro 2006, Anagnostopoulos et al. 2006);
- **Simplicidade e Flexibilidade:** o algoritmo é de fácil implementação (cerca de 200 linhas em Python) e permite ajustes nos parâmetros e restrições;
- **Qualidade das Soluções:** obtém resultados próximos ao ótimo conhecido, com desvios típicos de apenas 1–3% (Easton and Nemhauser 2001);
- **Capacidade de escapar de ótimos locais:** o critério probabilístico de aceitação favorece a exploração do espaço de soluções, reduzindo a chance de estagnação (Kirkpatrick et al. 1983).

Entre as limitações reconhecidas do SA, estão:

- **Ausência de garantia de otimalidade:** como abordagem heurística, não assegura a melhor solução possível;
- **Dependência de parâmetros:** requer calibragem adequada da temperatura inicial, taxa de resfriamento e número de iterações;
- **Variabilidade entre execuções:** resultados flutuam porque o processo é estocástico, exigindo múltiplas rodadas ou sementes para avaliar robustez, como sugerem (Anagnostopoulos et al. 2006, Urrutia and Ribeiro 2006).

Apesar dessas limitações, o Simulated Annealing demonstrou-se adequado ao propósito deste estudo, oferecendo um excelente equilíbrio entre qualidade da solução e

tempo computacional. Enquanto métodos exatos permanecem relevantes para instâncias pequenas ou para fins teóricos, em aplicações reais como a geração de tabelas do Campeonato Brasileiro, o uso de SA proporciona respostas rápidas, factíveis e com impacto logístico mensurável.

3.6. Geração do calendário

- **Geração do calendário inicial:** cria-se um primeiro turno *round-robin* simples (metodologia poligonal), em que cada equipe joga uma vez contra todas as outras (aleatoriamente definido quem será mandante em cada confronto). Em seguida, o segundo turno é obtido espelhando-se o primeiro (invertendo mandante e visitante), garantindo que cada par de equipes se encontre em casa e fora.
- **Função custo:** define-se $C(S) = D(S) + \lambda \cdot V(S)$ para cada calendário candidato S , em que $D(S)$ é a distância total percorrida por todas as equipes (em km, calculada a partir da matriz de distâncias entre sedes), $V(S)$ é o número de violações da restrição de no máximo dois jogos seguidos em casa ou fora, e λ é um fator de penalização grande que desencoraja soluções que infrinjam essa restrição.
- **Movimentações locais:** a cada iteração do algoritmo gera-se uma solução vizinha S' a partir da solução atual S por meio de uma modificação aleatória simples: (i) troca-se o mando (casa/fora) de uma partida aleatória no primeiro turno; ou (ii) permutam-se duas equipes em todo o primeiro turno, trocando assim seus adversários em todas as rodadas. Após cada movimento, reconstrói-se o calendário completo (atualizando também o segundo turno espelhado) e calcula-se o custo $C(S')$.
- **Critério de aceitação:** se $C(S') < C(S)$, a nova solução S' é automaticamente aceita como estado corrente. Caso contrário, S' pode ainda ser aceita com probabilidade $\exp[-C(S') - C(S)T]$, em que T é a temperatura atual. A cada iteração, a temperatura é atualizada por $T \leftarrow \alpha T$ (esquema de resfriamento exponencial, com $0 < \alpha < 1$).

Algorithm 1: Simulated Annealing para o mTTP

Input: $T_0, \alpha, T_{\min}, N =$
Output: Melhor calendário viável encontrado

```
1  $S \leftarrow RoundRobin(n)$                                      // calendário do 1º turno
2  $S \leftarrow Espelhar(S)$                                      // gera o 2º turno
3  $melhor \leftarrow S$ 
4  $T \leftarrow T_0$ 
5 while  $it < N$  and  $T > T_{\min}$  do
6    $S' \leftarrow Vizinho(S)$                                      // troca mandante/visitante ou identidades
7    $custo(S') \leftarrow Dist(S') + \lambda \cdot Violações(S')$ 
8   if  $custo(S') < custo(S)$  or  $rand() < \exp(-(custo(S') - custo(S))/T)$ 
9     then
10       $S \leftarrow S'$ 
11      if  $Violações(S) = 0$  and  $Dist(S) < Dist(melhor)$  then
12         $melhor \leftarrow S$ 
12    $T \leftarrow \alpha \cdot T$ 
13 return  $melhor$ 
```

4. Resultados

A análise dos dados de desempenho nos dias de jogo revelou diferenças sutis nas métricas fisiológicas de acordo com a distância das viagens. A Tabela 1 apresenta as médias ($\pm$ desvio-padrão) de distância total percorrida, distância em alta intensidade, PSE e TRIMP para cada categoria de viagem (curta, média e longa) do time analisado.

Tabela 1. Médias ($\pm$ desvio-padrão) das variáveis fisiológicas em dias de jogo segundo a categoria de viagem (G1).

Média $\pm$ desvio-padrão em dias de jogo (G1)					
Categoria	N	Dist. total (m)	Dist. alta (m)	PSE	TRIMP
Curta	50	8 720 $\pm$ 1 854	423 $\pm$ 205	8.04 $\pm$ 1.19	334 $\pm$ 102
Média	30	9 032 $\pm$ 2 047	491 $\pm$ 205	7.88 $\pm$ 1.34	337 $\pm$ 89
Longa	40	9 112 $\pm$ 2 022	409 $\pm$ 265	8.18 $\pm$ 1.38	353 $\pm$ 81

Observa-se que viagens longas acarretaram, em média, uma carga fisiológica ligeiramente maior: os jogadores percorreram cerca de 9 112 m por jogo nessas partidas, valor um pouco acima das categorias curta e média ($\sim$ 8720–9032m), assim como apresentaram TRIMP médio mais elevado ($\sim$ 353 versus $\sim$ 335 nas demais). A percepção subjetiva de esforço (PSE) também foi maior nas viagens longas (média 8,18) em comparação às médias (7,88), embora tenha se mantido similar às curtas (8,04). Por outro lado, a distância em alta intensidade não seguiu o mesmo padrão: surpreendentemente, o valor médio dessa métrica foi mais alto nas viagens médias (491 m) do que nas curtas ou longas, ainda que com grande variabilidade entre os jogos.

Importante destacar que nenhuma das diferenças acima se mostrou estatisticamente significativa quando testadas por ANOVA unidirecional ($p > 0.05$ em todos os casos). O tamanho de amostra limitado e as variações de desempenho entre partidas dificultam a confirmação de um efeito claro apenas com os dados disponíveis. Ainda assim, observam-se tendências coerentes com a expectativa de que viagens mais longas elevem o cansaço e a carga de jogo dos atletas. No conjunto analisado, partidas precedidas de viagens longas resultaram em métricas de carga (distância percorrida, TRIMP e PSE) ligeiramente superiores às de viagens curtas ou médias. Essas diferenças, embora modestas, sugerem que atuar após grandes deslocamentos tende a deixar os jogadores mais exaustos em dia de jogo.

Resultados de estudos prévios corroboram essas observações. Por exemplo, Draper et al. 2024 constataram que equipes que acumulam maiores distâncias de viagem ao longo da temporada tendem a apresentar volumes mais altos de corrida em alta intensidade nos jogos fora de casa; contudo, esse esforço adicional não se traduz em melhor desempenho competitivo – ao contrário, observou-se uma leve queda na pontuação obtida e um aumento no número de gols sofridos pelos times visitantes que mais viajaram. Tais achados apoiam a hipótese do presente estudo, de que a fadiga gerada por viagens prolongadas pode forçar os atletas a elevarem o esforço físico em campo, porém com eficiência reduzida. Além disso, Hands et al. 2023 demonstraram, no contexto do futebol australiano (A-League), que partidas fora de casa apresentam um perfil de esforços físicos distinto dos jogos em casa, com maior volume de corridas de baixa intensidade nos visitantes. Os autores verificaram que viagens terrestres realizadas no dia do jogo estão associadas a uma redução nas ações de alta intensidade durante a partida – possivelmente devido à rigidez muscular e fadiga acumulada no trajeto – enquanto deslocamentos aéreos mais longos efetuados com maior antecedência permitiram manter níveis de intensidade semelhantes aos dos jogos em casa. Esses achados reforçam a relevância de se analisar em detalhe como diferentes distâncias e modos de viagem influenciam o rendimento físico dos jogadores, fornecendo base científica para propor otimizações logísticas no calendário competitivo.

Do ponto de vista logístico, nossos resultados evidenciam que é possível mitigar significativamente a sobrecarga de viagens por meio de um reordenamento do calendário de jogos. Modelamos o problema de agendamento como uma instância do *Mirrored Traveling Tournament Problem* (TTP), que abstrai os desafios de construir tabelas esportivas minimizando distâncias totais percorridas sob restrições de mandos consecutivos e intervalos regulares entre partidas.

A natureza estocástica do SA implica que diferentes execuções podem resultar em soluções distintas, com calendários que favorecem, em graus variados, certos clubes. Por isso, executamos o algoritmo com 20.000 sementes diferentes. A tabela ?? corresponde à solução que minimizou a soma total das distâncias percorridas por todos os times ao longo do campeonato. Caso outros critérios fossem priorizados – como a equidade entre os deslocamentos ou a redução máxima para um time específico – outras sementes poderiam ter sido escolhidas, produzindo configurações alternativas com impactos distintos sobre os clubes.

A aplicação do SA mostrou-se eficaz e eficiente. Em aproximadamente 4 s de processamento (15.000 iterações em um computador equipado com um Ryzen 5 5500X,

16 GB de RAM e uma placa gráfica NVIDIA GeForce RTX 2060), obteve-se uma tabela de jogos válida para as 20 equipes, que respeita a restrição de no máximo duas partidas seguidas em casa ou fora. Houve uma redução substancial de distâncias: a quilometragem total percorrida por todos os times ao longo do campeonato caiu de 709 054 km (calendário original da CBF 2018) para 543 780 km no calendário otimizado – uma economia de $\sim 23,3\%$. A Tabela ?? apresenta a distância total percorrida por cada clube antes e depois da otimização, bem como a redução absoluta obtida para cada um.

Tabela 2. Distâncias percorridas por equipe nos calendários CBF e otimizado, com percentual de redução.

Time	CBF 2018 (km)	Otimizado (km)	Redução (km)	% Redução
Sport	67 724.8	49 962.5	17 762.3	26.23%
Ceará	76 598.1	61 126.3	15 471.8	20.20%
Internacional	44 093.4	29 420.2	14 673.2	33.28%
Grêmio	43 320.3	32 104.8	11 215.6	25.89%
Bahia	45 983.1	35 464.8	10 518.3	22.87%
Vitória	46 367.6	36 542.1	9 825.5	21.18%
Flamengo	27 092.6	18 097.5	8 995.1	33.19%
Chapecoense	39 745.1	30 833.6	8 911.6	22.42%
Botafogo	27 207.8	18 599.0	8 608.8	31.63%
Paraná	29 679.4	21 106.6	8 572.8	28.88%
Vasco	25 217.9	17 413.8	7 804.2	30.95%
Atlético-PR	29 198.4	22 099.5	7 098.9	24.30%
Corinthians	26 915.4	20 169.8	6 745.6	25.06%
Fluminense	24 349.2	18 114.8	6 234.4	25.59%
São Paulo	26 422.9	20 565.9	5 857.0	22.16%
Cruzeiro	26 275.0	20 926.4	5 348.6	20.36%
América-MG	28 153.5	22 899.3	5 254.1	18.66%
Santos	26 234.4	21 309.4	4 925.0	18.77%
Atlético-MG	25 645.9	24 698.5	947.4	3.69%
Palmeiras	22 829.6	22 324.9	504.7	2.21%
Total	709 054.3	543 779.5	165 274.8	23.30%

Percebe-se que as maiores reduções absolutas beneficiaram principalmente as equipes geograficamente mais distantes dos principais centros. Clubes como Sport (PE), Ceará (CE), Internacional (RS), Grêmio (RS), Bahia (BA) e Vitória (BA) deixaram de percorrer cerca de 9,8 a 17,8 mil km cada um no novo calendário. Em contrapartida, clubes do Sudeste, como Palmeiras (SP) e Atlético-MG (MG), apresentaram diminuições bem menores (da ordem de poucas centenas de quilômetros), uma vez que seus deslocamentos originais já eram relativamente curtos. Esse resultado reforça a motivação inicial do estudo: times de regiões periféricas têm mais a ganhar com um calendário otimizado do que aqueles situados em regiões centrais.

Adicionalmente, examinou-se a distribuição das distâncias de viagem por categoria no calendário antes e depois da otimização. Definiram-se viagens curtas como aquelas de até 500 km de distância, médias entre 501 e 999 km, e longas acima de 1000 km. A Tabela 3 resume o número de viagens de cada categoria no calendário oficial da CBF

2018 em comparação ao calendário otimizado pelo SA.

Tabela 3. Número de viagens por categoria antes e depois da otimização.

Categoria	Oficial (CBF 2018)	Otimizado	Diferença
Curta (≤ 500 km)	233	278	+45
Média (501–999 km)	145	122	–23
Longa (> 1000 km)	286	219	–67

No calendário original ocorreram 286 viagens longas; no calendário otimizado esse número caiu para 219, eliminando 67 deslocamentos superiores a 1 000 km. Em contrapartida, o total de viagens curtas aumentou de 233 para 323 (+90). Isso indica que muitas viagens inicialmente médias foram reprojetadas como trajetos mais curtos, e diversas viagens que seriam longas passaram a se enquadrar como médias ou curtas no novo cronograma. Assim, a distribuição das distâncias concentrou-se em deslocamentos mais curtos, reduzindo trajetos extenuantes e, potencialmente, a fadiga associada.

Apesar do êxito da heurística, o *Simulated Annealing* apresenta limitações inerentes: (i) não proporciona garantia de optimalidade; (ii) depende de parâmetros como temperatura inicial, taxa de resfriamento e número de iterações; e (iii) produz resultado ligeiramente diferente a cada execução, em razão de sua natureza estocástica. Ainda assim, o SA mostrou-se adequado aos propósitos deste estudo, oferecendo ótimo equilíbrio entre qualidade da solução e tempo computacional. Métodos exatos permanecem úteis para instâncias menores ou análises teóricas, mas em aplicações práticas – como a geração da tabela do Brasileirão – heurísticas como o SA fornecem respostas rápidas, viáveis e com impacto logístico mensurável.

Em suma, nossos resultados demonstram que a reestruturação logística do calendário de jogos consegue diminuir o total percorrido pelas equipes ao longo da temporada. Ao alinhar os dados fisiológicos à otimização combinatória via SA, este estudo evidencia que ajustes inteligentes no calendário têm o potencial de mitigar a fadiga dos jogadores e promover um maior equilíbrio competitivo no futebol brasileiro de elite.

5. Conclusão

Este trabalho avaliou o uso do *Simulated Annealing* (SA) para resolver o Problema do Torneio Viajante Espelhado (mTTP) aplicado ao calendário da Série A de 2018. Detalhamos todo o ciclo do algoritmo — geração poligonal da solução inicial, função de custo com penalização de violações, vizinhanças e esquema de recozimento — e mostramos que seu custo por iteração é polinomial em n^2 , o que explica a execução em poucos segundos para instâncias reais de 20 clubes (Anagnostopoulos et al. 2006). Em contraste, métodos exatos baseados em Programação Inteira ou *branch-and-bound* mantêm apelo teórico, mas tornam-se impraticáveis além de 8–10 equipes devido à explosão combinatória (Bhattacharyya 2009, Urrutia and Ribeiro 2006). Mesmo sem garantia formal de ótimo, o SA atingiu calendários muito próximos do melhor conhecido, justificando-se como abordagem pragmática.

Do ponto de vista logístico, o cronograma otimizado reduziu a quilometragem total das equipes de 709 054 para 543 780 km ($\sim 23,3\%$). Ainda que o clube-caso analisado não figure entre os que mais viajam, sua carga despencou de 28 153,5 para 22 899,3 km,

economia de $\sim 5\,254\text{ km}$ ($\approx 18,6\%$). Em esportes de alto rendimento, onde frações de por cento podem decidir uma temporada, cada quilômetro poupado traduz-se em horas extras de recuperação, menor estresse fisiológico e maior probabilidade de rendimento máximo em campo. Esses ganhos marginais, somados ao longo de 38 rodadas, podem valer pontos preciosos.

Além disso, a redistribuição das viagens atenua desigualdades regionais: clubes geograficamente periféricos deixam de arcar com longos deslocamentos semanais, enquanto as equipes do eixo Sudeste — historicamente favorecidas pela proximidade entre si — veem sua vantagem logística reduzir-se. Um calendário mais equânime, portanto, não só melhora a saúde dos atletas como também pavimenta o caminho para maior equilíbrio competitivo; quem sabe, num futuro próximo, possamos testemunhar o primeiro campeão da era dos pontos corridos proveniente fora do Sudeste.

Em síntese, a combinação de um algoritmo meta-heurístico eficiente com a análise das demandas fisiológicas dos jogadores mostra que otimizar o calendário não é mero exercício acadêmico: é uma intervenção concreta, de baixo custo computacional e alto impacto esportivo. O *Simulated Annealing* oferece soluções de qualidade quase ótima em tempo exíguo, enquanto métodos exatos permanecem úteis como referência para instâncias pequenas ou validação teórica. Para as dimensões do Brasileirão, porém, a abordagem heurística provou-se a ferramenta mais efetiva — capaz de mitigar a fadiga, elevar o espetáculo e tornar a disputa pelo título mais justa para todos os torcedores.

6. Trabalhos futuros

Embora este estudo tenha demonstrado o potencial da otimização logística para contrubuir com a redução do desgaste físico dos atletas, reconhecemos limitações na base de dados utilizada. Nos próximos passos desta linha de pesquisa, pretende-se ampliar substancialmente o escopo da base de dados — hoje restrita a uma única temporada de um clube — de modo a incorporar séries históricas de várias equipes e múltiplos anos competitivos. Esse acréscimo de observações permitirá estimar com precisão estatística mais robusta além da relação entre métricas internas de carga (TRIMP, PSE) e o perfil de deslocamentos. Contemplando não apenas a distância percorrida, mas também variáveis situacionais como modo de transporte, intervalos de recuperação, altitude do destino e densidade do calendário. Com uma amostra mais rica, será possível treinar modelos preditivos híbridos, combinando regressão multivariada e aprendizado de máquina para quantificar o impacto marginal de cada quilômetro viajado sobre o desgaste fisiológico projetado para diferentes cenários de logística.

Uma vez calibrados, esses modelos servirão de núcleo a um simulador de desempenho físico acoplado ao algoritmo de otimização do calendário. A ideia é que, a cada iteração do *Simulated Annealing* (ou de meta-heurísticas complementares), o simulador estime não só a quilometragem total, mas também a fadiga acumulada prevista para cada atleta, permitindo que a função-custo inclua penalidades dinâmicas baseadas em risco fisiológico. Tal integração abre caminho para agendas de investigação em que calendários otimizados não se limitem à minimização de distâncias, mas busquem minimizar um índice composto de desgaste, potencializando ganhos de performance e reduzindo lesões. Estudos futuros deverão, ainda, validar as projeções do simulador comparando-as com dados de campo subsequentes, refinando iterativamente o modelo e aproximando-o

de um sistema de apoio à decisão operável em tempo real por clubes e federações.

Referências

aaaa.

Anagnostopoulos, A., Michel, L., Hentenryck, P. V., and Vergados, Y. (2006). A simulated annealing approach to the traveling tournament problem. *Journal of Scheduling*, 9(2):177–193.

Bertsimas, D., Farias, V. F., and Trichakis, N. (2011). The price of fairness. *Operations Research*, 59(1):17–31.

Bhattacharyya, R. (2009). A note on complexity of traveling tournament problem. Optimization Online Preprint 2480. Submetido em 29 de novembro de 2009; última atualização em 20 de dezembro de 2009. Acesso em 28 abr. 2025.

Draper, M., Smith, J., and Clarke, R. (2024). The effect of travel on the performance of elite soccer teams: A multi-club analysis. *Journal of Sports Sciences*, 42(3):315–324.

Easton, K. and Nemhauser, G. (2001). The traveling tournament problem description and benchmarks. *Lecture Notes in Computer Science*, 2239:580–584.

Hands, R., Cormack, S., and Coutts, A. (2023). Influence of travel modality on external match load in professional soccer. *International Journal of Sports Physiology and Performance*, 18(4):316–324.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.

Kruyt, F., Freedman, J., Hall, M. J., Coutts, A. J., and Kempton, T. (2023). The influence of time between travel and competition on external workload characteristics among professional soccer players. *Journal of Strength and Conditioning Research*, 37(9):1802–1809.

Sampaio, J., Lago, C., and Gonçalves, B. (2019). Fatores situacionais no futebol de alto nível: Jogar em casa vs. fora e implicações para o desempenho físico. *International Journal of Performance Analysis in Sport*, 19(3):219–223.

Silva, P. R., Machado, F. A., de Almeida, M. F., de Souza, C. D., Simim, M. A. M., Mendes, E. V., Nakamura, F. Y., and da Mota, G. R. (2021). The effect of match-related fatigue on short-passing ability in young soccer players. *Sports Medicine - Open*, 7(1):47.

Silva, P. R., Machado, F. A. d. S., de Almeida, M. F., Simim, M. A. M., Mendes, E. V., Nakamura, F. Y., and da Mota, G. R. (2024). Effects of match-related fatigue on short-passing accuracy in young soccer players: A full-season study. *Journal of Sports Science and Medicine*, 23:744–752.

Transfermarkt (2025a). Campeonato brasileiro série a 2025. <https://www.transfermarkt.com.br/campeonato-brasileiro-serie-a/startseite/wettbewerb/BRA1>. Acesso em 26 mai. 2025.

Transfermarkt (2025b). Copa do brasil — todos os vencedores. <https://www.transfermarkt.com.br/copa-do-brasil/erfolge/pokalwettbewerb/BRC>. Acesso em 26 mai. 2025.

Transfermarkt (2025c). Copa libertadores — todos os vencedores. [⟨https://www.transfermarkt.com.br/copa-libertadores/erfolge/pokalwettbewerb/CLI⟩](https://www.transfermarkt.com.br/copa-libertadores/erfolge/pokalwettbewerb/CLI). Acesso em 26 mai. 2025.

Urrutia, S. and Ribeiro, C. C. (2006). Maximizing breaks and bounding solutions to the mirrored traveling tournament problem. *Discrete Applied Mathematics*, 154(13):1932–1938.

Agrupamento de Jogadores de Futebol Similares do Campeonato Brasileiro de 2023 Utilizando Algoritmos de Aprendizado de Máquina Não-Supervisionados

1

1. Introdução

Com o aumento da complexidade do futebol moderno e a disponibilidade crescente de dados, torna-se essencial empregar análises sofisticadas para compreender os perfis dos jogadores. Nesse contexto, técnicas de aprendizado de máquina consolidam-se como ferramentas eficazes para segmentar atletas segundo características semelhantes. Entre essas, destaca-se a clusterização, que permite agrupar jogadores com funções e estilos de jogo compatíveis, auxiliando os departamentos técnicos na formação de elencos, no planejamento tático e na avaliação de desempenho (2).

Para reforçar a aplicação da ciência de dados no âmbito do futebol profissional, este trabalho realiza uma análise comparativa entre quatro técnicas de clusterização: K-Means, Gaussian Mixture Model (GMM), Clustering Espectral e Hierárquico. Utiliza-se uma base de dados abrangente obtida do Football Manager 2023 (FM2023) (1), contendo atributos físicos, técnicos e comportamentais dos jogadores. Busca-se examinar como cada algoritmo estrutura os atletas em grupos funcionalmente coerentes, avaliando sua competência em refletir padrões relevantes para o alto rendimento esportivo.

2. Metodologia

Este estudo utilizou uma amostra de 1.141 jogadores da Série A do Campeonato Brasileiro, extraída do FM2023 (1), selecionada a partir de um conjunto original de mais de 170 mil atletas para garantir maior homogeneidade dos dados.

O pré-processamento incluiu a remoção de registros inválidos, a normalização da idade, a padronização de variáveis numéricas e ordinais, e a codificação *one-hot* para variáveis categóricas. Em seguida, aplicou-se a Análise de Componentes Principais (PCA) para redução de dimensionalidade, limitando os dados a três componentes principais. A escolha do número de grupos variou conforme o algoritmo: método do cotovelo para K-Means, critérios BIC/AIC para GMM, análise da matriz laplaciana com 12 vizinhos para o clustering espectral e dendograma para o hierárquico.

Todos os algoritmos foram executados com parâmetros padronizados, convergindo para oito agrupamentos, conforme descrito na Tabela 1.

Tabela 1. Configurações dos algoritmos de clusterização utilizadas

Algoritmo	Parâmetros
K-Means	8 clusters <code>n_init = 10</code> <code>random_state = 42</code>
Gaussian Mixture Model (GMM)	8 componentes <code>covariância full</code> <code>random_state = 42</code>
Spectral Clustering	8 clusters afinidade por vizinhos mais próximos (<code>n_neighbors = 12</code>) <code>random_state = 42</code>
Hierárquico	8 clusters distância euclidiana ligação ward

Por fim, os rótulos gerados por cada algoritmo foram incorporados a um novo *DataFrame*, juntamente com variáveis adicionais como nome, posição, função, estilo de jogo, idade e quantidade de gols. Para facilitar a visualização dos agrupamentos, os dados foram representados em gráficos tridimensionais com o uso da biblioteca `Matplotlib`.

3. Resultados

Esta seção apresenta uma análise comparativa dos agrupamentos produzidos pelos quatro modelos: K-Means, GMM, clusterização espectral e hierárquica. A avaliação considera três aspectos principais: média de idade, média de gols e coerência funcional dos perfis identificados. As Figuras 1, 2, 3 e 4 ilustram a distribuição dos jogadores segundo cada algoritmo, oferecendo suporte visual à comparação entre os modelos.

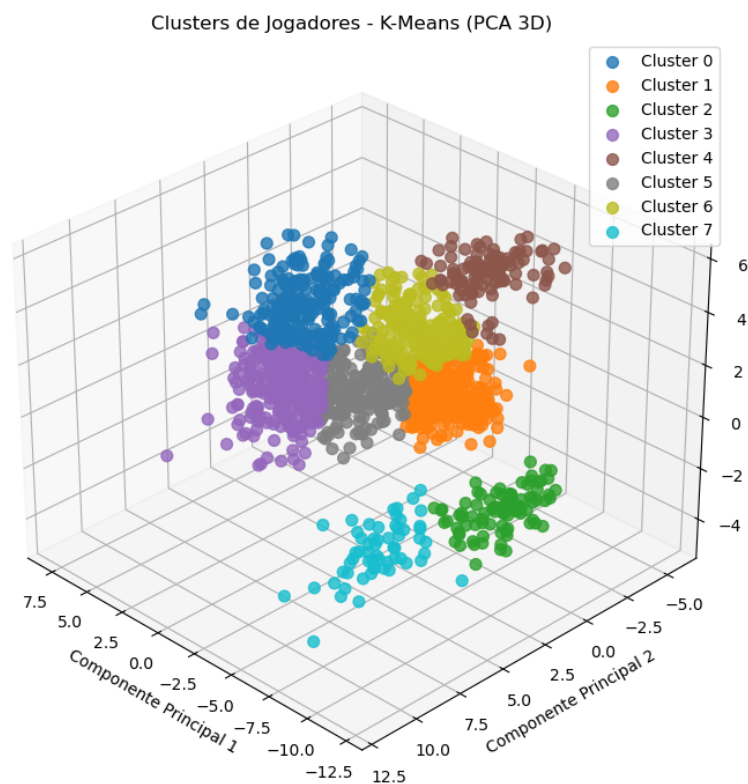


Figura 1. Agrupamento com K-Means

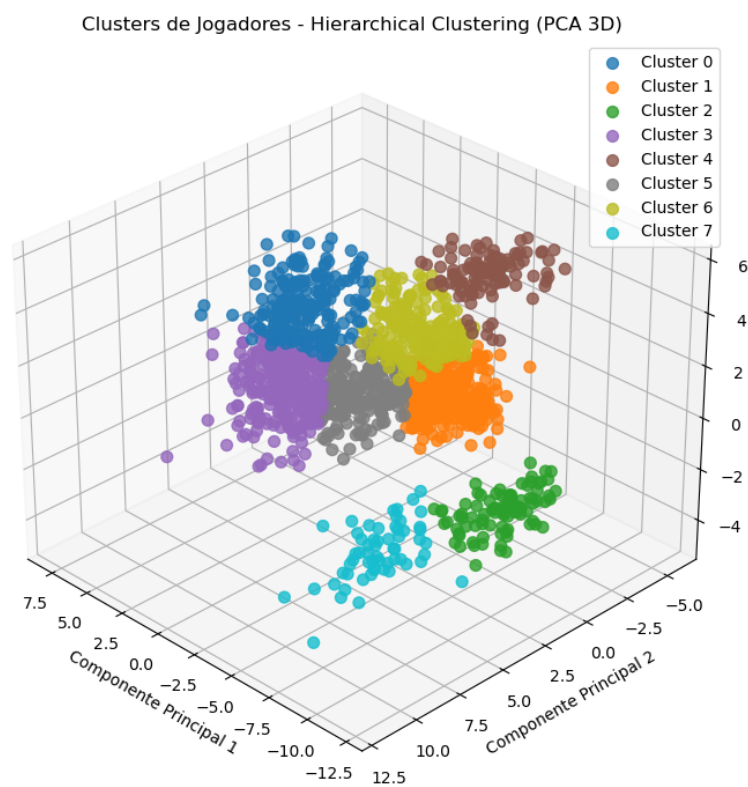


Figura 2. Agrupamento com Clusterização Hierárquica

Clusters de Jogadores - Spectral Clustering (PCA 3D)

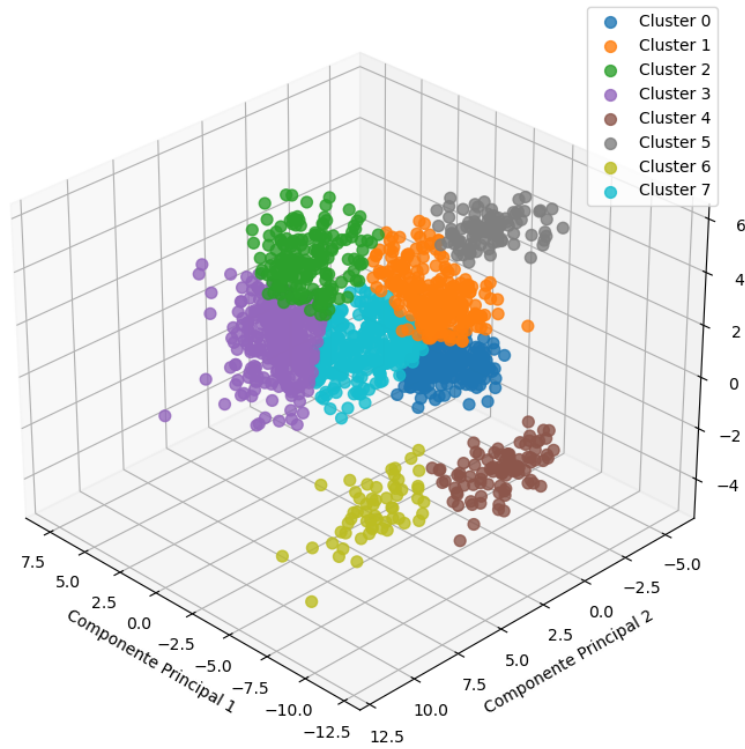


Figura 3. Agrupamento com Clusterização Espectral

Clusters de Jogadores - Gaussian Mixture Model (GMM) (PCA 3D)

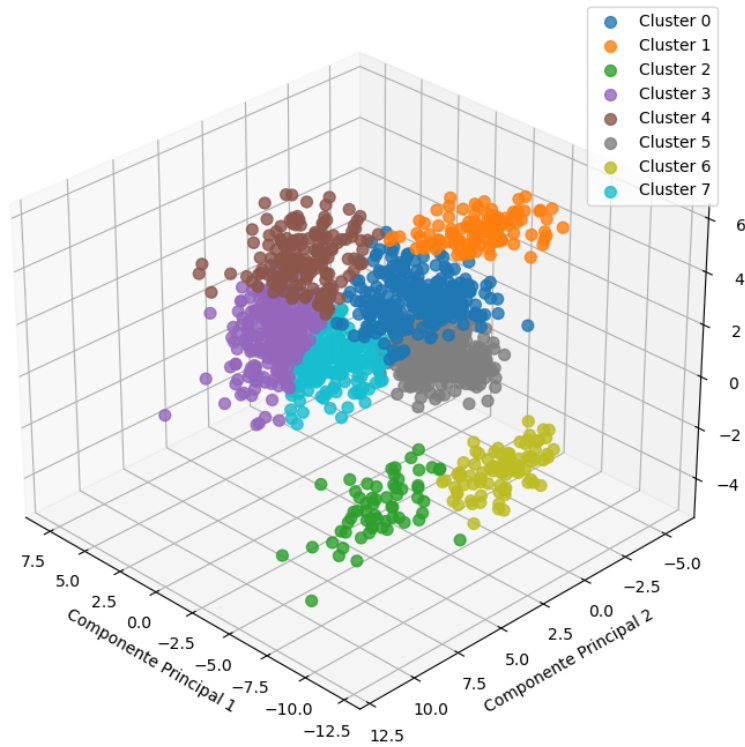


Figura 4. Agrupamento com *Gaussian Mixture Model*

3.1. Goleiros

Os modelos analisados identificaram dois grupos funcionalmente distintos de goleiros. No K-Means, o cluster 2 (90 jogadores, 19,19 anos, 0,00 gols) reuniu 55 goleiros tradicionais e 35 líberos, enquanto o cluster 7 (59 jogadores, 28,61 anos) inverteu essa proporção (13 tradicionais e 46 líberos), sugerindo que goleiros mais jovens tendem a desempenhar funções mais conservadoras. O GMM apresentou padrão semelhante: cluster 6 (54 tradicionais, 33 líberos, 19,05 anos) e cluster 2 (14 tradicionais, 48 líberos, 28,35 anos). O clustering espectral e o hierárquico reforçaram esse padrão etário-estilístico, com seus respectivos grupos de goleiros mais velhos (clusters 6 e 4) concentrando maior número de líberos (46 e 35). A consistência entre os modelos demonstra a relação entre idade e estilo de jogo, indicando maturidade como fator determinante para atuações mais ofensivas na posição.

3.2. Laterais e Meias Defensivos

O *cluster* 6 do K-Means (201 jogadores, 0,17 gols, 19,03 anos) destacou-se pela presença predominante de laterais defensivos (67) e zagueiros centrais (30), indicando um perfil essencialmente defensivo. O GMM (*cluster* 0), com média de idade de 19,10 anos, ampliou esse escopo ao incluir 83 laterais defensivos, 27 meias centrais e 19 volantes de distribuição, sugerindo maior versatilidade na transição entre defesa e meio-campo. O *cluster* 1 do modelo espectral (18,74 anos) equilibrou 78 laterais com 33 zagueiros centrais e 32 meias centrais, evidenciando um perfil mais híbrido. Por fim, o modelo hierárquico (*cluster* 7), com 18,30 anos de média, foi o mais restrito, reunindo 59 laterais defensivos e apenas dois zagueiros centrais. A análise revela que GMM e espectral capturaram melhor a sobreposição entre defesa e meio-campo, enquanto K-Means e hierárquico resultaram em perfis mais polarizados.

3.3. Centroavantes Técnicos e Meias Ofensivos Jovens

O *cluster* 1 do K-Means (261 jogadores, 0,14 gols, 18,22 anos) combinou 51 centroavantes de área ("camisas 9") e 45 meias centrais, refletindo um perfil equilibrado entre finalização e construção. O GMM (*cluster* 5), com média de idade de 18,30 anos, reforçou o viés ofensivo ao reunir 54 camisas 9, 42 meias centrais e 57 pontas invertidos. O modelo espectral (*cluster* 0), com 18,14 anos, apresentou composição semelhante (38 camisas 9, 23 meias centrais), mas com ênfase maior em pontas invertidos (42). O hierárquico (*cluster* 5), com média de 18,28 anos, foi o mais balanceado: 49 camisas 9, 36 meias centrais e 54 pontas invertidos. A consistência entre os modelos demonstra a existência de um grupo jovem com vocação ofensiva mista, variando apenas na proporção entre funções de finalização e de criação.

3.4. Pontas e Jogadores de Beirada

O *cluster* 5 do K-Means (216 jogadores, 3,05 gols, 20,36 anos) apresentou maioria de pontas invertidos (67) e pontas tradicionais (31), formando um grupo claramente ofensivo. O GMM (*cluster* 7), com média de idade de 22,17 anos, ampliou esse perfil, reunindo 90 pontas invertidos e 33 pontas tradicionais, o que explica sua média mais elevada de gols (9,30). O modelo espectral (*cluster* 7), com a mesma média etária de 20,36 anos, mostrou configuração semelhante (73 pontas invertidos, 40 tradicionais), mas incluiu 19 meias interiores, adicionando diversidade tática ao grupo. O hierárquico (*cluster* 1) foi o mais

extremo, reunindo 134 pontas invertidos — o maior valor absoluto entre todos os modelos, com média de idade de 25,05 anos e expressiva média de gols (21,66). Os resultados indicam uma forte correlação entre a presença de pontas invertidos e a produtividade ofensiva dos agrupamentos.

3.5. Zagueiros Centrais Clássicos Novos

Todos os modelos apresentaram *clusters* notavelmente puros para zagueiros. O K-Means (*cluster* 4) concentrou 94 Zagueiros Centrais entre seus 101 jogadores (0,04 gols), com média de idade de 18,60 anos. O GMM (*cluster* 1) foi ainda mais específico — 105 Zagueiros Centrais em 112 jogadores, com média de 18,72 anos. O Espectral (*cluster* 5), com 89 Zagueiros Centrais, apresentou 18,50 anos de média etária, enquanto o Hierárquico (*cluster* 3), com 106 Zagueiros Centrais, registrou 18,67 anos. A impressionante consistência entre modelos, tanto em posição (D (C) em mais de 90 por cento dos casos) quanto em função (Zagueiro Central predominante), demonstra que este é o perfil mais bem definido e isolável no espaço de características.

3.6. Zagueiros e Volantes Defensivos

O *cluster* 0 do K-Means (198 jogadores, 7,79 gols, 26,29 anos) reuniu 53 volantes de marcação e 45 zagueiros centrais, configurando um grupo de atuação defensiva com alguma participação ofensiva. O GMM (*cluster* 4), com 26,49 anos de média, manteve proporção similar (35 volantes, 46 zagueiros), com ligeira elevação nos gols. O modelo espectral (*cluster* 2), com 25,97 anos, apresentou o perfil mais defensivo (43 volantes, 47 zagueiros), com média de gols mais baixa (6,21). Já o hierárquico (*cluster* 0), com média etária de 27,15 anos, incluiu também 27 alas defensivos e 63 volantes, elevando a produção ofensiva para 13,96 gols. A análise mostra que a presença de alas e a proporção entre volantes e zagueiros têm impacto direto no potencial ofensivo do agrupamento.

3.7. Atacantes e Armadores Ofensivos

O *cluster* 3 do K-Means (332 jogadores, 29,54 gols, 27,40 anos) destacou-se como o mais produtivo ofensivamente, com 91 pontas invertidos e 31 alas defensivos. O GMM (*cluster* 3), com média de idade de 27,73 anos, apresentou composição semelhante (53 pontas invertidos, 37 alas) e incluiu 25 volantes de marcação, resultando em leve queda na média de gols (28,71). O modelo espectral (*cluster* 3), com 27,56 anos, manteve a tendência (87 pontas invertidos, 37 alas), com alta produtividade (29,25 gols). Já o hierárquico (*cluster* 1), embora tenha reunido 134 pontas invertidos, apresentou maior heterogeneidade funcional (43 pontas tradicionais, 25 alas), o que reduziu sua média de gols para 21,66, mesmo com média etária elevada (25,05 anos). Os resultados indicam que a concentração de pontas invertidos, quando não diluída por outras funções, é o principal fator para maximizar a produção ofensiva dos agrupamentos.

3.8. Correlação entre Idade e Gols

A análise dos agrupamentos revela uma correlação direta entre idade média e produtividade ofensiva. *Clusters* compostos por jogadores mais experientes, como K-Means 3 (27,40 anos), GMM 3 (27,73 anos) e Espectral 3 (27,56 anos), apresentam médias de gols acima de 28. Em contrapartida, agrupamentos predominantemente jovens, como K-Means 1 (18,22 anos) e Espectral 0 (18,14 anos), não ultrapassam 0,25 gols. Esse padrão

é compatível com a lógica competitiva: tempo de atuação e experiência favorecem maior impacto ofensivo. Mesmo sendo técnicas não supervisionadas, os modelos conseguiram capturar essa tendência estatística natural com consistência.

4. Conclusão

Os quatro modelos de clusterização aplicados demonstraram capacidade de segmentar jogadores de forma funcional, cada um com ênfases distintas. O K-Means destacou-se pela formação de grupos mais homogêneos, especialmente entre goleiros e zagueiros, além de apresentar o *cluster* mais ofensivo (29,54 gols, 27,40 anos). O GMM mostrou flexibilidade na identificação de perfis híbridos e também alto desempenho ofensivo (*cluster* 3: 28,71 gols, 27,73 anos). O modelo espectral combinou coerência posicional com nuances táticas, mantendo alta produtividade no *cluster* 3 (29,25 gols, 27,56 anos). Já o hierárquico, embora com maior heterogeneidade interna, agrupou bem jogadores ofensivos em larga escala, como no *cluster* 1 (21,66 gols, 25,05 anos).

Todos os modelos capturaram com precisão a correlação entre idade e produção ofensiva, evidenciando maturidade como fator determinante para desempenho. Enquanto K-Means e espectral se mostraram mais eficazes na segmentação clara de perfis, GMM e hierárquico foram superiores na identificação de funções mistas. Os resultados reforçam o potencial do aprendizado não supervisionado como ferramenta estratégica para análise tática e tomada de decisão no futebol profissional.

Referências

- [1] ULUTAŞ, Furkan. *Football Manager 22 Complete Player Dataset*. Kaggle. Disponível em: <https://www.kaggle.com/datasets/furkanuluta/football-manager-22-complete-player-dataset>. Acesso em: 2 ago. 2024.
- [2] DECROOS, Tom; DAVID, Stijn; VAN HAAREN, Jan; DAVIS, Jesse. Actions speak louder than goals: Valuing player actions in soccer. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019. p. 1851–1861.
- [3] MITCHELL, T. M. *Machine Learning*. New York: McGraw-Hill, 1997.
- [4] LUDERMIR, Teresa Bernarda. Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências. *Estudos Avançados*, São Paulo, v. 35, n. 101, p. 85–94, jan. 2021. Disponível em: <https://doi.org/10.1590/s0103-4014.2021.35101.007>. Acesso em: 3 ago. 2025.

Análise das Características da Posse de Bola e sua Relação com os Gols Esperados

1

1. Introdução

A posse de bola tem se tornado, ao longo dos últimos anos, alvo de diversos estudos que buscam compreender sua influência sobre o resultado e a organização tática das equipes [Wang et al. 2022, Casal et al. 2019, Chacoma and Kuperman 2020]. Segundo o CIES Football Observatory [Observatory 2023], times campeões em grandes ligas mantiveram, em média, 57% de posse durante suas temporadas vencedoras. No entanto, uma revisão sistemática de 75 estudos [Wang et al. 2022] concluiu que o tempo de posse, por si só, não determina o resultado da partida.

Nesse contexto, surge a necessidade de analisar não apenas a duração de posse, mas sua efetividade na geração de oportunidades de gol. A métrica de Gols Esperados (xG) permite quantificar essa efetividade com base nas características das finalizações. Assim, este estudo busca investigar quais atributos de um ciclo de posse estão associados a maiores valores de xG.

Além desta introdução, o artigo está estruturado em três seções. A Seção 2 detalha os métodos utilizados no estudo. A Seção 3 descreve os resultados obtidos. Finalmente, a Seção 4 discute as principais conclusões e sugere direções para pesquisas futuras.

2. Metodologia

O teste de Kruskal-Wallis foi utilizado para avaliar se existiam diferenças estatisticamente significativas entre grupos de dados não paramétricos, especificamente para variáveis categóricas que influenciam o ciclo de passe. Estas categorias, denominadas 'Zero', 'Baixo' e 'Alto', foram definidas com base no valor de 'total_xg' de cada ciclo de passe: 'Zero', quando o xG do ciclo era igual a 0; 'Baixo', para valores maiores que 0 e menores ou iguais a 0.5; e 'Alto', para valores acima de 0.5. Para comparar essas três categorias, foi aplicado o teste de Kruskal-Wallis, que é robusto a desvios da normalidade e apropriado para comparar três ou mais grupos independentes, abordagem similar à utilizada por [Stafylidis et al. 2024a] para comparar indicadores de desempenho em diferentes resultados de partidas. Após a identificação de diferenças significativas, testes pós-hoc foram utilizados para verificar quais grupos diferiam entre si.

Dois modelos, Regressão Linear e Random Forest Regressor, foram aplicados para prever o desempenho dos ciclos de passe. A Regressão Linear quantificou a influência das variáveis, Essa abordagem segue uma linha semelhante à adotada por [Stafylidis et al. 2024b], que utilizou modelos estatísticos para identificar fatores preditivos relevantes no futebol, embora, no presente estudo, o foco esteja na previsão contínua do valor de xG por posse, e não em desfechos binários como o resultado da partida. Enquanto o Random Forest Regressor foi utilizado para identificar a importância de cada uma delas, por ser um método robusto a relações complexas. A performance dos modelos foi avaliada usando métricas como R^2 , MAE, RMSE e MSE.

2.1. Dados

Os dados utilizados neste trabalho são fornecidos gratuitamente pela StatsBomb, disponíveis em seu repositório no GitHub¹. Estes dados contêm informações detalhadas sobre eventos ocorridos em partidas de futebol. Para a construção da base de dados, foram extraídos todas as posses de bola das equipes de cada partida da La Liga. Cada posse foi considerada a partir da recuperação da bola até sua perda, finalização ou interrupção do jogo. Foi adotado um critério específico: apenas posses que resultaram em finalização foram consideradas para o cálculo do total_xg associado a cada ciclo ofensivo.

As variáveis utilizadas neste estudo foram derivadas a partir dos eventos registrados em cada ciclo de posse de bola. O `n_pass` e o número total de passes realizados durante o ciclo de posse, o `n_carries` e o número de conduções de bola durante o ciclo. A `dist_bola` e a soma das distâncias percorrida pela bola durante o ciclo em metros. O `tempo_ciclo` e o tempo total do ciclo em segundos. A partir dessa medida, foi calculada a `v_méd`, definida como a razão entre a distância percorrida e o tempo da posse, expressando a velocidade média da movimentação da bola. O direcional dos passes foi avaliado por meio de três proporções: `pass_lat`, que indica a proporção de passes realizados lateralmente (com ângulo aproximadamente perpendicular à direção do ataque); `pass_dir_def`, que representa a fração de passes realizados em direção contrária ao gol adversário; e `pass_dir_atq`, que quantifica os passes executados em direção ao ataque.

A variável `pressure` quantifica o número de vezes que um jogador adversário aplicou pressão em um jogador com a bola (seja recebendo, conduzindo ou soltando a bola) durante o ciclo. Por fim, `inicio_ciclo` representa a posição inicial da posse, normalizada para indicar se o ciclo começou no campo de defesa (0) ou no de ataque (1). Uma descrição resumida de cada variável pode ser encontrada na Tabela 1.

Table 1. Descrição das variáveis extraídas dos ciclos de posse

Variável	Descrição
<code>n_pass</code>	Número de passes realizados
<code>n_carries</code>	Número de conduções de bola
<code>dist_bola</code>	Distância total percorrida pela bola (m)
<code>tempo_ciclo</code>	Duração total da posse (s)
<code>v_méd</code>	Velocidade média da bola no ciclo (m/s)
<code>pass_lat</code>	Proporção de passes laterais
<code>pass_dir_def</code>	Proporção de passes para trás (direção defesa)
<code>pass_dir_atq</code>	Proporção de passes para frente (direção ataque)
<code>pressure</code>	Número de adversários aplicando pressão
<code>inicio_ciclo</code>	Posição inicial da posse (normalizada de 0 a 1)

3. Resultados

Os resultados do teste de Kruskal-Wallis revelaram diferenças significativas para a maioria das variáveis analisadas indicando que a categorização dos ciclos de passe (Alto, Baixo, Zero) tem um impacto distinto sobre elas, como mostra a Tabela 2. A análise

¹<https://github.com/statsbomb/open-data>

estatística utilizando o teste de Kruskal-Wallis revelou diferenças significativas entre categorias de posse de bola com xG classificados como 'Alto', 'Baixo' e 'Zero'. Essas diferenças foram observadas em algumas variáveis do ciclo ofensivo, incluindo número de passes, número de conduções, distância percorrida pela bola, tempo do ciclo, velocidade média da posse, percentuais de passes laterais, e direção dos passes (em direção à defesa e ao ataque), além da posição inicial do ciclo. A maioria das variáveis apresentaram p-valores inferiores a 0.05.

Table 2. Resultados do teste de Kruskal-Wallis e comparações pós-hoc de Dunn para as variáveis dos ciclos de posse de bola, considerando as categorias de xG (Alto, Baixo, Zero).

Variável	Alto (M±DP)	Baixo (M±DP)	Zero (M±DP)	H	Comp.
n_pass	6.21 ± 5.97	5.42 ± 5.40	4.94 ± 4.68	17.24	† * ‡
n_carries	5.06 ± 5.16	4.59 ± 4.98	3.68 ± 4.25	273.669	* ‡
dist_bola	147.36 ± 124.12	131.89 ± 120.84	117.58 ± 107.52	112.620	† * ‡
tempo_ciclo	19.11 ± 18.40	17.93 ± 17.59	15.76 ± 16.36	142.163	* ‡
v_méd	9.36 ± 5.58	9.08 ± 4.99	9.44 ± 14.61	45.124	† ‡
pass_lat	0.58 ± 0.28	0.60 ± 0.31	0.43 ± 0.32	1842.942	* ‡
pass_dir_def	0.18 ± 0.21	0.19 ± 0.25	0.27 ± 0.31	434.933	* ‡
pass_dir_atq	0.23 ± 0.25	0.19 ± 0.25	0.27 ± 0.31	474.104	‡
pressure	0.50 ± 0.46	0.55 ± 0.51	0.54 ± 0.47	3.637	—
inicio_ciclo	0.48 ± 0.50	0.54 ± 0.50	0.31 ± 0.46	1614.154	* ‡

Nota. M ± DP = média ± desvio padrão; H = estatística H do teste de Kruskal-Wallis; Comparações: símbolos indicam diferenças significativas pós-hoc ($p < 0.05$): * Alto ≠ Zero, † Alto ≠ Baixo, ‡ Baixo ≠ Zero.

Ciclos de posses classificadas como 'Alto' e 'Baixo' apresentaram maior número de passes e conduções em comparação com a categoria 'Zero', sugerindo que ciclos ofensivos mais produtivos envolvem maior movimentação da bola e dos jogadores. As variáveis de distância percorrida pela bola e tempo do ciclo também foram significativamente maiores em posses de maior produção ofensiva, indicando que esses ciclos tendem a ser mais longos e a cobrir maiores distâncias. Contudo, a velocidade média da posse não se mostrou um fator decisivo isolado para a geração de xG elevado, apresentando diferença significativa apenas entre os grupos 'Baixo' e 'Zero'.

O percentual de passes laterais foi menor nas posses 'Zero', enquanto a direção dos passes (para defesa ou ataque) e a posição de início do ciclo foram fortemente associadas ao potencial ofensivo. A localização espacial de início da posse demonstrou ser um fator crucial para o potencial ofensivo do ciclo. Por outro lado, a proporção de tempo da posse no terço defensivo do campo não apresentou variação significativa entre os grupos, indicando que essa variável não está sistematicamente ligada à produtividade ofensiva do ciclo.

A avaliação dos modelos de regressão linear e Random Forest Regressor revelou um desempenho insatisfatório em ambos os casos. O modelo de regressão linear apresentou um R^2 de 0.0179, indicando uma capacidade muito limitada de explicar a variabilidade da variável dependente, com erros quantificados por RMSE de 0.0504, MSE de 0.0025 e MAE de 0.0186. Por sua vez, o Random Forest Regressor obteve um R^2 negativo de -0.0234, sugerindo que seu desempenho foi inferior ao de um modelo base que prevê ape-

nas a média, com métricas de erro ligeiramente superiores (RMSE de 0.0514, MSE de 0.0026 e MAE de 0.0194).

4. Conclusões

A análise estatística das posses de bola revelou que características como número de passes, conduções, distância percorrida e tempo do ciclo são significativamente distintas entre posses de alto, baixo e zero xG, com maior produtividade ofensiva associada a ciclos mais longos e com maior movimentação da bola. A posição inicial do ciclo também se mostrou crucial para o potencial ofensivo. Contudo, a velocidade média da posse e a proporção de tempo no terço defensivo não foram fatores determinantes.

Apesar dessas associações estatísticas, os modelos de Regressão Linear e Random Forest Regressor apresentaram desempenho preditivo muito limitado, com valores de R^2 próximos ou inferiores a zero. Esses resultados indicam que, embora existam diferenças significativas entre os grupos de posse, as variáveis analisadas não foram suficientes para prever com precisão o valor contínuo de xG em cada ciclo ofensivo.

Para aprimorar a compreensão e a capacidade preditiva, trabalhos futuros poderiam explorar a inclusão de variáveis adicionais. A investigação de diferentes abordagens para a categorização do xG e a validação dos modelos em outros campeonatos ou com conjuntos de dados maiores também seriam contribuições valiosas.

References

- Casal, C. A., Maneiro, R., Ardá, A., Marí, F. J., and Losada, J. L. (2019). Possession in football: More than a quantitative aspect – a mixed method study. *Frontiers in Psychology*, 10:501.
- Chacoma, A. and Kuperman, M. (2020). Modeling ball possession dynamics in the game of football. *arXiv preprint arXiv:2005.04020*.
- Observatory, C. F. (2023). How important is ball possession in football? Accessed: 2025-08-02.
- Stafylidis, A., Mandroukas, A., Michailidis, Y., and Metaxas, T. I. (2024a). Decoding success: Predictive analysis of uefa euro 2024 to uncover key factors influencing soccer match outcomes. *Applied Sciences*, 14(17).
- Stafylidis, A., Mandroukas, A., Michailidis, Y., Vardakis, L., Metaxas, I., Kyranoudis, A. E., and Metaxas, T. I. (2024b). Key performance indicators predictive of success in soccer: A comprehensive analysis of the greek soccer league. *Journal of Functional Morphology and Kinesiology*, 9(2).
- Wang, Y., Zhang, S., and Zhou, C. (2022). A systematic review about the performance indicators related to ball possession. *PLOS ONE*, 17(3):e0265540.

Predição de Sucesso para Situações de Pressão no Futebol

Luiza de Melo Gomes¹, Caio Souza Grossi¹, Júlia Souza Moura¹,
Kaique de Oliveira e Silva¹, Victor Vieira Brito Amaral Pessoa¹

¹ Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte – MG – Brasil

{luizademelo caio.grossi}@dcc.ufmg.br

Abstract. *This article presents a study on the prediction of success in pressure situations in professional football, using data from the 2023/2024 Premier League season. Tracking and event data were integrated to extract contextual and spatial information from each play. Among the main features used, we highlight the number of defenders near the ball carrier, the density of nearby players, the speed of the play, as well as measures such as distance to goal. Based on this information, a machine learning model using the XGBoost algorithm was trained to automatically classify whether a pressure was successful. The model achieved an accuracy close to 80%, demonstrating the potential of data-driven techniques to analyze tactical behavior in football.*

Resumo. *Este artigo apresenta um estudo sobre a predição de sucesso em situações de pressão no futebol profissional utilizando dados da temporada 2023/2024 da Premier League. Os dados de tracking e de eventos foram integrados para extrair informações contextuais e espaciais de cada jogada. Dentre as principais variáveis utilizadas, é possível destacar o número de defensores próximos ao portador da bola, a densidade de jogadores ao redor, a velocidade da jogada, além de medidas como distância ao gol. A partir dessas informações, um modelo de aprendizado de máquina baseado no algoritmo XGBoost foi treinado, com o objetivo de classificar automaticamente se uma pressão foi bem-sucedida. O modelo alcançou uma acurácia próxima de 80%, demonstrando um possível potencial de técnicas baseadas em dados para analisar comportamentos táticos no futebol.*

1. Introdução

A evolução das tecnologias de rastreamento possibilitou o mapeamento preciso das posições de jogadores e da bola no futebol profissional. Com registros em alta frequência (10-25Hz), geram-se séries temporais detalhadas complementadas por dados de eventos como passes e finalizações. Apesar destes avanços, a pressão defensiva permanece pouco explorada quantitativamente, embora seja uma tática fundamental no futebol contemporâneo.

O presente trabalho propõe uma definição operacional para o conceito de pressão bem-sucedida, caracterizando-a como ações defensivas coordenadas que resultam em perdas de posse ou intercepções. Para quantificar sua intensidade, incorporamos variáveis derivadas das posições relativas entre jogadores e bola em cada instante analisado, desenvolvendo um modelo preditivo capaz de classificar automaticamente o sucesso de situações de pressão.

2. Metodologia

O comportamento de pressão de um time pode ser descrito como a combinação de múltiplas instâncias de interações defensivas no campo. Em uma situação de defesa, um time busca restringir as opções do adversário por meio da aproximação coordenada ao portador da bola ou a jogadores próximos, caracterizando uma tentativa de pressão. A efetividade dessa ação depende de fatores como posicionamento, densidade de jogadores na região da bola, distância ao gol e velocidade da jogada.

Neste artigo, propõe-se uma análise quantitativa dos fatores associados ao sucesso em situações de pressão, com o desenvolvimento de um modelo preditivo baseado em aprendizado de máquina. O modelo foi construído a partir de variáveis extraídas dos dados de tracking e de eventos da temporada 2023/2024 da Premier League. As principais features utilizadas estão descritas na Tabela 1. Como algoritmo de aprendizado, foi adotado o XGBoost (eXtreme Gradient Boosting) [Chen et al. 2015], escolhido por sua capacidade de lidar com dados tabulares e por incorporar mecanismos de regularização e controle de sobreajuste.

Tabela 1. Features utilizadas para treinar o modelo de predição de pressão

ID	Feature	Descrição
1	Distância esperada do passe	Distância do passe, ou a média de passes se o evento não for passe
2	Densidade de jogadores	Número de jogadores em um raio de 10 metros da bola
3	Defensores próximos	Número de jogadores adversários em um raio de 5 metros
4	Distância mais próxima de um adversário	Distância mais próxima da bola a um adversário
5	Número de defensores à frente	Número de jogadores adversários à frente do eixo y onde a bola se encontra
6	Distância do centro	Distância da bola ao centro do campo
7	Velocidade da bola	Velocidade da bola naquele frame
8	Velocidade dos defensores	Velocidade média dos adversários
9	Coordenada x da bola	Posição do eixo x da bola
10	Coordenada y da bola	Posição do eixo y da bola
11	Distância da linha lateral	Distância da bola das linhas laterais
12	Distância do gol	Distância da bola do gol
13	Compressão vertical	Distância da bola ao adversário mais avançado no eixo x
14	Ângulo do gol	Ângulo da bola ao centro do gol
15	Direção angular da bola	Direção do movimento da bola entre dois frames consecutivos

2.1. Integração de dados de tracking com dados de eventos

A integração entre os dados de tracking e de eventos foi realizada com base na variável *frameNum*, comum aos dois conjuntos. Esse número identifica o instante do jogo em que determinado evento ocorreu e permite localizar as posições dos jogadores no campo

no exato momento do evento. A partir dessa correspondência, foi possível enriquecer cada evento com informações espaciais extraídas do tracking, como a localização dos jogadores adversários e companheiros, além da posição estimada da bola.

2.2. Limpeza de dados

Na etapa de limpeza, foram removidas instâncias com valores ausentes ou inconsistentes em variáveis relevantes para o modelo. Também foi realizada a filtragem dos tipos de eventos utilizados. Para este trabalho, foram considerados apenas eventos classificados como passes, chutes ou duelos, pois são os mais diretamente associados a momentos de pressão no contexto tático do futebol. Além disso, foram excluídos eventos em que não havia informações suficientes para calcular as variáveis de interesse.

2.3. Funcionamento do modelo

O objetivo do modelo desenvolvido é prever se uma tentativa de pressão será bem-sucedida, a partir de variáveis contextuais que caracterizam o momento da jogada. O processo de construção do modelo está representado na Figura 1. Após a etapa de limpeza dos dados, foi realizada a extração de features espaciais e contextuais, seguida pela normalização das variáveis numéricas. Em seguida, aplicou-se o algoritmo SMOTE [Chawla et al. 2002] para balancear o conjunto de dados, que originalmente apresentava uma proporção desigual entre eventos com e sem sucesso na pressão.

A etapa final consistiu no treinamento do modelo XGBoost com os dados processados. O desempenho foi avaliado com base em métricas padrão de classificação binária, como acurácia, precisão, *recall*, *f1-score*, e análise de curvas ROC e de precisão-revocação.

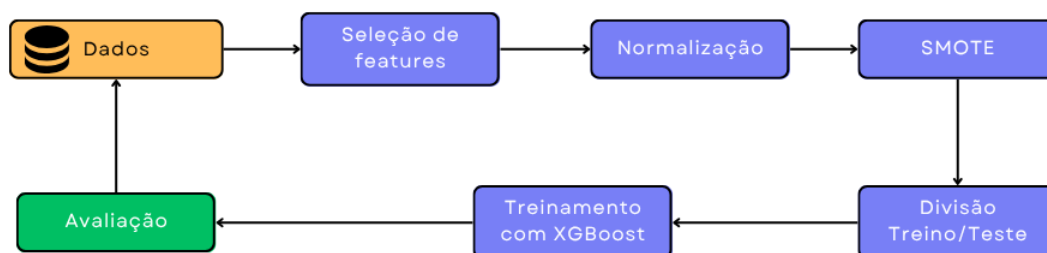


Figura 1. Pipeline do modelo

3. Resultados

O modelo alcançou uma acurácia geral de 80%, com valores equilibrados de *precision*, *recall* e *f1-score* entre as duas classes avaliadas: pressão bem-sucedida e pressão mal-sucedida (Tabela 2).

A matriz de confusão (Figura 2) evidencia a capacidade do modelo de realizar classificações corretas com frequência semelhante para ambas as classes. Foram corretamente identificadas 92.253 instâncias de pressão bem-sucedida (verdadeiros positivos) e 92.556 instâncias de pressão mal-sucedida (verdadeiros negativos). Em contrapartida,

Tabela 2. Resultados de avaliação do modelo de predição de pressão

Classe	Precision	Recall	F1-score	Suporte
Falso	0,80	0,80	0,80	116263
Verdadeiro	0,80	0,80	0,80	115935
Média (macro)	0,80	0,80	0,80	232198
Média ponderada	0,80	0,80	0,80	232198
Acurácia geral: 0,80				

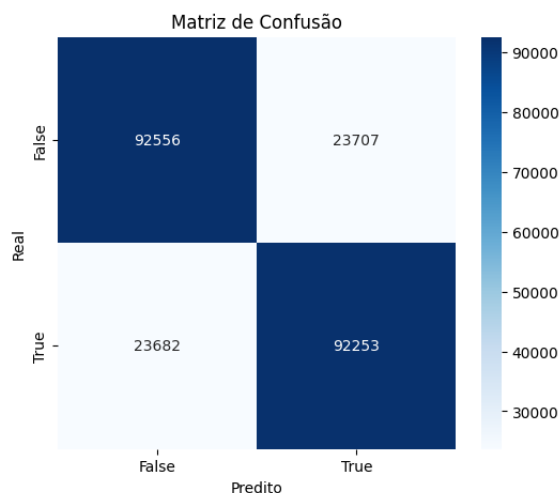


Figura 2. Matrix de confusão

ocorreram 23.707 falsos positivos e 23.682 falsos negativos. Essa distribuição sugere que o modelo não apresenta viés significativo para nenhuma das classes.

A curva ROC apresenta área sob a curva (AUC) de 0,87, indicando boa capacidade discriminativa do modelo. A curva de Precisão/Revocação demonstra estabilidade ao longo do espectro de limiares de decisão, sugerindo robustez do classificador mesmo em cenários com classes desbalanceadas.

A análise de importância das features (Figura 3) revelou que variáveis como distância esperada do passe, número de defensores próximos e densidade de jogadores estão entre as que mais contribuíram para a predição de sucesso da pressão.

4. Conclusão e trabalhos futuros

Este artigo demonstrou a viabilidade de um modelo de aprendizado supervisionado para prever o sucesso de situações de pressão no futebol profissional, com acurácia de 80%. Os resultados indicam que fatores espaciais como densidade de jogadores e número de defensores próximos são determinantes para a efetividade da pressão defensiva.

Duas limitações principais foram identificadas: (1) possibilidade de *overfitting*, devido à ausência de validação com dados de outras temporadas ou competições; e (2) alto custo computacional da extração de features, com tempo estimado de aproximadamente uma hora por partida.

Trabalhos futuros podem explorar otimizações como paralelização na extração de

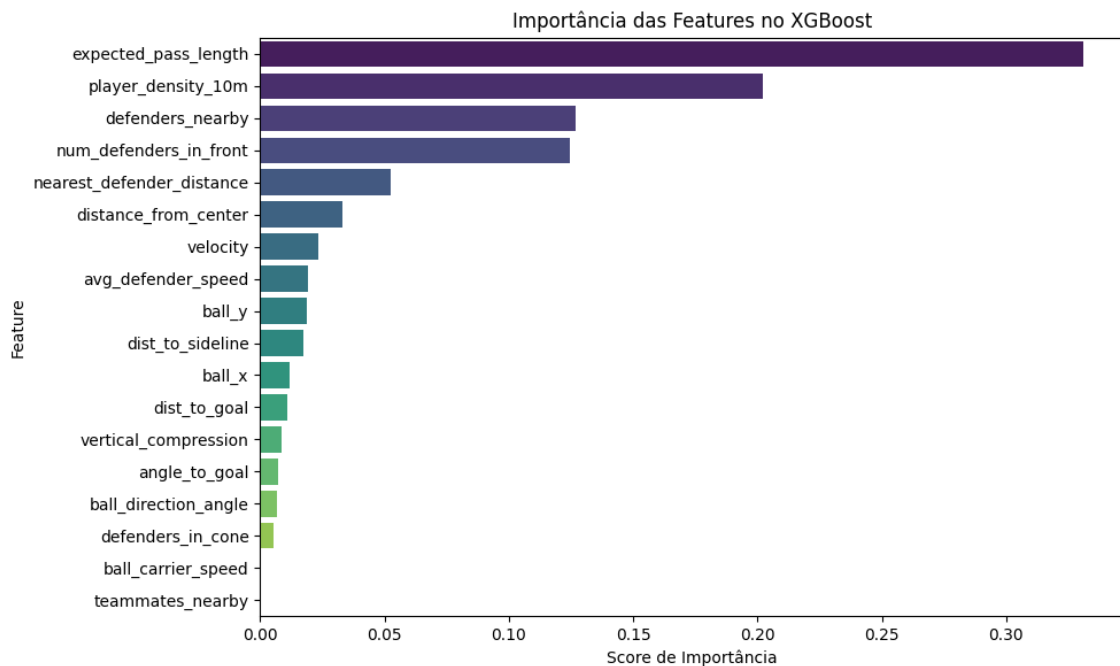


Figura 3. Features mais importantes

features, técnicas de amostragem inteligente ou abordagens baseadas em redes neurais recorrentes que incorporem a dinâmica temporal do jogo. Adicionalmente, uma validação qualitativa com especialistas pode contribuir para refinar o modelo e confirmar sua utilidade prática em contextos de análise tática.

Referências

- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.
- Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

Variáveis Táticas Espaciais Associadas à Vitória no Futebol

Lenizio M. Oliveira¹, Gibson M. Praça¹, André G. P. Andrade¹

¹Escola de Educação Física, Fisioterapia e Terapia Ocupacional –
Universidade Federal de Minas Gerais (UFMG)

leniziomorais@gmail.com

1. Introdução

No futebol de alto rendimento, o acesso a uma quantidade crescente de dados gerados em partidas tem impulsionado a integração de técnicas de aprendizado de máquina (ML) na análise esportiva. Essa transformação modifica a maneira como o desempenho das equipes é analisado e previsto, oferecendo novas possibilidades para extrair padrões mais profundos sobre as dinâmicas do jogo e lidar com a abundância de informações disponíveis. Tradicionalmente, muitos modelos preditivos utilizam métricas baseadas em eventos, como finalizações, passes e desarmes, por fornecerem medidas diretas das ações das equipes. Contudo, esse tipo de dado, conforme Praça et al. (2024a), tende a negligenciar o contexto espacial e temporal em que as ações ocorrem.

Em contraste, os indicadores derivados de dados posicionais capturam a distribuição espacial dos jogadores e o movimento da bola.[Memmert et al., 2017; Memmert and Rein, 2018]. Isso permite uma compreensão mais detalhada das estratégias coletivas. Embora mais complexos e menos disponíveis, os dados posicionais possibilitam a identificação de padrões relacionados à imprevisibilidade e à adaptabilidade tática, elementos considerados cruciais no futebol moderno, de acordo com Memmert et al. (2017). Apesar do crescente interesse em métricas baseadas em eventos e dados posicionais, a literatura ainda carece de estudos que investigam o poder preditivo desses tipos de indicadores derivados desses dados. Diante desse cenário, o presente estudo tem como objetivo identificar quais variáveis táticas, extraídas de dados posicionais, apresentam maior impacto na predição de resultados de partidas.

2. Métodos

A amostra consistia de 26 variáveis táticas previamente extraídas a partir de dados posicionais de 273 jogos da temporada 2021 - 2022 da Bundesliga. As variáveis utilizadas no estudo foram coletadas tanto no primeiro quanto no segundo tempo das partidas e incluem: "team width"(nas fases ofensiva e defensiva), "team length"(nas fases ofensiva e defensiva), "teams length per width ratio"(nas fases ofensiva e defensiva), "team stretching index (nas fases ofensiva e defensiva), "in-phase"(eixo x e y), "anti-phase"(eixo x e y) e "near the ball space control".

Para a construção do modelo preditivo, foi utilizado o algoritmo XGBoost, com divisão dos dados em 85% para treinamento e 15% para teste. A busca pelos melhores hiperparâmetros foi realizada por meio de Grid Search com validação cruzada de 5 folds. Após a seleção dos parâmetros ideais, o modelo foi treinado com todos os dados de treinamento e, em seguida, avaliado no conjunto de teste. As métricas utilizadas para avaliação de desempenho foram acurácia e F1-score, uma vez que as classes estavam

balanceadas. Além disso, foi aplicada a análise SHAP (Shapley Additive Explanations) [Lundberg and Lee 2017] no modelo final, com o objetivo de identificar as variáveis com maior contribuição para as previsões realizadas. Partidas que terminaram em empate foram excluídas da amostra, a fim de focar apenas nos resultados decisivos (vitórias ou derrotas), tornando o problema uma tarefa de classificação binária.

3. Resultados

O modelo XGBoost alcançou desempenho satisfatório na predição de resultados de partidas, com acurácia de 67,74% e F1-score ponderado de 67,74% no conjunto de teste. Durante a busca por hiperparâmetros via Grid Search, o modelo apresentou acurácia média de 66,47% ($\pm 5,59\%$) e F1-score médio de 66,26% ($\pm 5,57\%$) na validação cruzada, demonstrando consistência entre os conjuntos de validação e teste. A análise SHAP identificou as dez variáveis com maior impacto preditivo no modelo (Figura 1). Entre estas, as duas variáveis mais importantes foram "Near-the-Ball Control 2nd Half" e "Width 1st Half Attack", apresentando valores SHAP médios de aproximadamente 0,35 e 0,34, respectivamente.

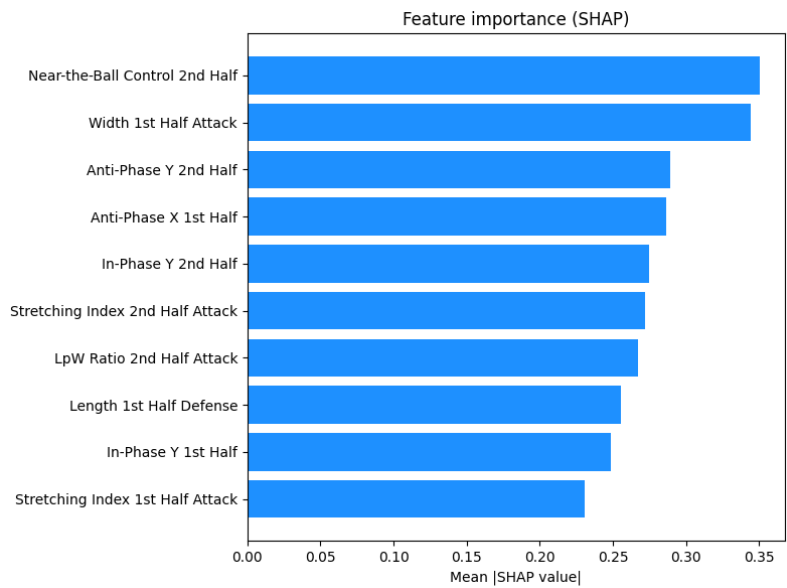


Figura 1. Top 10 variáveis por valor SHAP médio.

A variável "Near-the-Ball Control 2nd Half" demonstrou o maior impacto preditivo no modelo. A análise dos valores SHAP individuais (Figura 2) indica que quando esta métrica apresenta valores elevados, há uma contribuição positiva significativa para a predição de vitória da equipe, com valores SHAP frequentemente superiores a 0,5. Isso sugere que equipes com maior controle espacial próximo à bola durante o segundo tempo possuem maior probabilidade de vencer a partida. A segunda variável mais relevante, "Width 1st Half Attack", também apresentou padrão similar (Figura 2). Valores elevados desta métrica correspondem a contribuições SHAP positivas para a predição de vitória, frequentemente atingindo valores próximos a 1,0. Este resultado corrobora estudos anteriores que demonstram a importância da amplitude ofensiva para o sucesso no futebol. Conforme evidenciado por Praça et al. (2024b), equipes vencedoras tendem a expandir

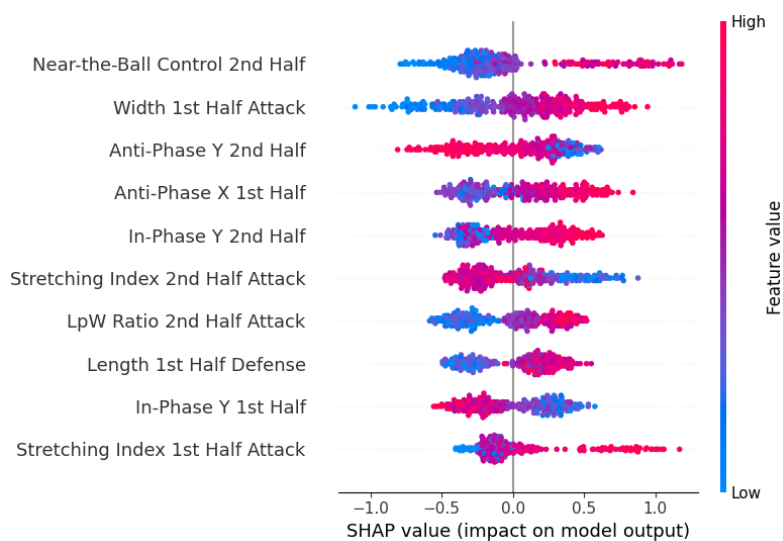


Figura 2. Top 10 variáveis – SHAP summary plot.

a ocupação do campo durante a fase ofensiva no primeiro tempo, estratégia fundamental que favorece a construção de jogadas de ataque e está associada ao maior tempo de posse de bola. Entre as demais variáveis de maior impacto, destacam-se "Anti-Phase Y 2nd Half", "Anti-Phase X 1st Half" e "In-Phase Y 2nd Half", todas apresentando valores SHAP médios entre 0,25 e 0,30 (Figura 1).

4. Conclusão

O modelo XGBoost alcançou acurácia de 67,74% na predição de resultados e demonstrou que as variáveis táticas derivadas de dados posicionais apresentam potencial significativo para a predição de resultados no futebol de elite. A análise SHAP revelou que o controle espacial próximo à bola no segundo tempo ("Near-the-Ball Control 2nd Half") e a amplitude ofensiva no primeiro tempo ("Width 1st Half Attack") constituem os fatores táticos mais determinantes para o sucesso das equipes. É importante reconhecer que os resultados baseiam-se exclusivamente em dados da Bundesliga (temporada 2021-2022), o que pode limitar a generalização dos achados para outras ligas e contextos futebolísticos. Além disso, o tamanho relativamente pequeno da amostra (273 jogos) pode restringir a robustez estatística das conclusões. Estudos futuros devem incluir múltiplas ligas e temporadas para validar a aplicabilidade universal destes indicadores táticos.

Referências

- Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*.
- Memmert, D., Lemmink, K. A. P. M., and Sampaio, J. (2017). Current approaches to tactical performance analyses in soccer using position data. *Sports Medicine*, 47:1–10.
- Memmert, D. and Rein, R. (2018). Match analysis, big data and tactics: Current trends in elite soccer. *Deutsche Zeitschrift für Sportmedizin*, 69(3):65–72.

- Praça, G., Brandão, L. A., Moura, F. A., Bedo, B. L., Aquino, R., and Andrade, A. G. (2024a). Quantifying within-matches tactical behaviors using position data and notational analysis in soccer: The effect of goal scoring. *Kinesiology*, 56.
- Praça, G., Brandão, L. A., Silva, G., Oliveira, P., Costa, V., and Pinheiro, G. (2024b). Positioning to win: Key position-based performance indicators in elite soccer. *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*, 0(0).

Impacto do Calendário no Desempenho das equipes da Série A do Campeonato Brasileiro

Leonardo Pessoa Miranda, João Vitor Bicalho da Silva,

João Pedro Reis Teclo de Miranda e Gabriel Chaves Ferreira

Orientador: Prof. Adriano César Machado Pereira

August 4, 2025

Abstract

O calendário do futebol brasileiro é reconhecidamente um dos mais congestionados do mundo. Entre abril e dezembro, clubes da Série A precisam disputar até 70 partidas oficiais, entre Campeonato Brasileiro, Copa do Brasil, fases de mata-mata da Libertadores ou Sul-Americana. O curto intervalo entre jogos reduz as janelas de recuperação fisiológica, e expõe os atletas a viagens de longa duração e múltiplos fusos horários.

Este trabalho tenta descobrir qual o provável impacto do calendário no desempenho dos times e atletas participantes do campeonato brasileiro de 2023. Para isso, o estudo envolveu análises quantitativas das viagens de todos os 20 clubes, métricas de desempenho dos jogadores e resultado das partidas.

Os resultados deste projeto oferecem uma base para a discussão sobre os impactos do calendário no desempenho dos times do campeonato brasileiro.

1 Introdução

Os times que estão na Série A do Campeonato Brasileiro, enfrentam uma dificuldade diferente de outras ligas ao redor do mundo: a densidade de seu calendário. Entre abril e dezembro os clubes enfrentam uma sequência extenuante de jogos que pode chegar a mais de 70 partidas oficiais no ano. Somando compromissos nacionais e internacionais, como Copa do Brasil, Libertadores e Sul-Americana. Em meio a essa maratona, os atletas lidam com curtos períodos de descanso, treinos reduzidos e deslocamentos frequentes, muitas vezes envolvendo viagens longas e trocas de fusos horários.

Essa sobrecarga levanta uma questão crítica: até que ponto o calendário apertado influencia negativamente o desempenho dos times e de seus jogadores? Embora essa preocupação esteja presente em debates esportivos e na mídia especializada, estudos com evidências quantitativas sobre os efeitos do calendário nas métricas físicas dos jogadores e nos resultados em campo ainda são limitados no contexto brasileiro.

2 Problema e Desafios Metodológicos

Estudar o impacto do calendário no desempenho físico in-game dos times da Série A envolve alguns desafios importantes, tanto na coleta quanto na análise dos dados. A seguir, descrevemos os principais pontos enfrentados neste trabalho.

2.1 Dificuldade em isolar o efeito do calendário

O desempenho físico dos jogadores durante uma partida pode ser afetado por vários fatores além do intervalo entre jogos, como a estratégia adotada pela equipe, o nível do adversário, o clima, a altitude, e até a motivação no jogo. Por isso, é difícil afirmar com certeza que uma queda em métricas como velocidade média ou distância percorrida foi causada apenas pelo calendário apertado.

2.2 Limitações dos dados disponíveis

Apesar de conterem informações relevantes sobre desempenho físico (como km percorridos, velocidade média e máxima), os dados utilizados neste estudo não incluem variáveis fisiológicas mais específicas, como batimento cardíaco, fadiga muscular ou tempo de recuperação. Além disso, nem todas as partidas possuem os mesmos dados disponíveis, o que pode afetar a consistência das análises.

2.3 Diferenças estruturais entre os clubes

Os clubes da Série A possuem realidades logísticas e financeiras diferentes. Enquanto alguns times têm acesso a voos fretados e estruturas de recuperação avançadas, outros enfrentam viagens longas em voos comerciais, com menos tempo de descanso. Isso faz com que o impacto do calendário não seja o mesmo para todos, exigindo cuidado na hora de comparar diferentes equipes.

2.4 Interpretação das métricas físicas

Métricas como distância total ou velocidade média não indicam, por si só, se o jogador estava desgastado fisicamente. Um time pode correr menos por uma decisão tática, e não por cansaço. Da mesma forma, altos valores podem acontecer mesmo em situações de fadiga. Por isso, é preciso interpretar essas métricas dentro do contexto do jogo.

2.5 Correlação não é causalidade

Mesmo que os dados mostrem uma relação entre jogos em sequência e queda de desempenho, isso não significa que um causou o outro. Outros fatores podem estar por trás da correlação, e sem controle experimental é difícil provar que o calendário foi a causa direta da mudança no desempenho.

2.6 Poucas situações extremas

Casos mais extremos de sobrecarga, como três jogos em menos de sete dias, são menos frequentes e, por isso, difíceis de analisar com profundidade estatística. Isso limita a comparação entre diferentes graus de exigência do calendário.

3 Objetivo

O objetivo central deste estudo reside na tentativa de mensurar os impactos reais do calendário sobre o rendimento físico e competitivo dos clubes da Série A. Ao integrar dados de desempenho físico dos atletas (como distância percorrida, velocidade média, velocidade máxima, etc) com o histórico de partidas do Brasileirão 2023, buscamos identificar padrões e correlações que possam embasar discussões sobre a organização do futebol nacional.

Os resultados deste projeto oferecem uma base para reflexões mais amplas sobre a necessidade de ajustes no calendário, contribuindo para uma gestão esportiva mais eficiente e centrada no bem-estar dos atletas.

4 Análise quantitativa das viagens

Para entender o impacto logístico do calendário no desempenho das equipes, foi realizada uma análise quantitativa das distâncias percorridas por cada time da Série A ao longo da temporada de 2023. O objetivo dessa etapa foi mensurar o esforço de deslocamento exigido pelo calendário.

4.1 Construção dos dados de viagem

A primeira etapa envolveu a obtenção do calendário completo de jogos de todas as competições de cada uma das 20 equipes da Série A do Campeonato Brasileiro de 2023. Em seguida, foi necessário fazer o scrapping das coordenadas de todos os estádios onde os jogos foram realizados. Com isso, conseguimos calcular a distância aproximada que o time percorreu a partir do seu estádio até o local do jogo.

Por fim, os dados foram cruzados para gerar o total de quilômetros viajados por cada equipe ao longo da temporada. Esse cálculo considerou todas as competições disputadas no ano (incluindo Brasileirão, Estaduais, Copa do Brasil e torneios continentais) somando as distâncias de todas as viagens realizadas.

Os dados referentes aos jogos e competições foram obtidos através do site <https://www.ogol.com.br/>.

4.2 Interpretação do gráfico

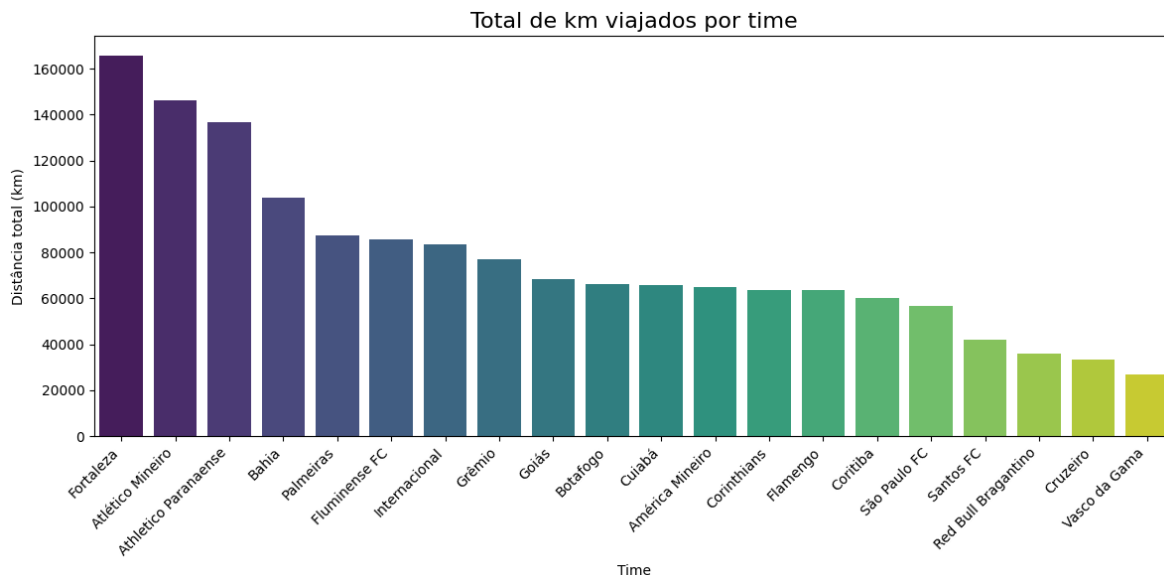


Figure 1: Total de quilômetros viajados por time

O gráfico acima mostra o total de quilômetros viajados por cada clube ao longo da temporada de 2023. Observa-se uma grande desigualdade entre os times:

O Fortaleza foi o clube que mais viajou, acumulando mais de 160.000 km no ano. Atlético Mineiro e Athletico Paranaense aparecem logo em seguida, com aproximadamente 145.000 km e 137.000 km, respectivamente. Por outro lado, clubes como Vasco da Gama, Cruzeiro e Red Bull Bragantino percorreram distâncias bem menores, ficando abaixo dos 40.000 km.

Esse resultado evidencia como o local de origem dos clubes e sua participação em competições paralelas influenciam diretamente a carga de viagens enfrentada por cada equipe. Essa desigualdade pode gerar diferenças relevantes no desgaste físico ao longo da temporada.

5 Análise de Correlação entre Estresse de Viagem e Desempenho Físico

Nesta seção, buscamos avaliar se há alguma relação estatística entre métricas de estresse logístico e indicadores de desempenho físico dos atletas durante as partidas. A análise foi realizada com base nas seguintes variáveis:

- Variáveis dependentes (desempenho físico): distância total percorrida (`total_dist_m`), distância em alta intensidade (`high_int_m`) e velocidade média (`mean_speed`).
- Variáveis preditoras (estresse de viagem): quilômetros viajados até o jogo (`travel_km`), distância total viajada nos últimos 7 dias (`travel_7d_km`) e 28 dias (`travel_28d_km`), além dos dias de descanso desde o último jogo (`rest_days`).

Foram aplicadas duas abordagens de correlação:

- **Correlação de Pearson:** avalia a relação linear entre duas variáveis contínuas.
- **Correlação de Spearman:** avalia a relação monotônica, sendo menos sensível a outliers e a distribuições não lineares.

5.1 Resultados

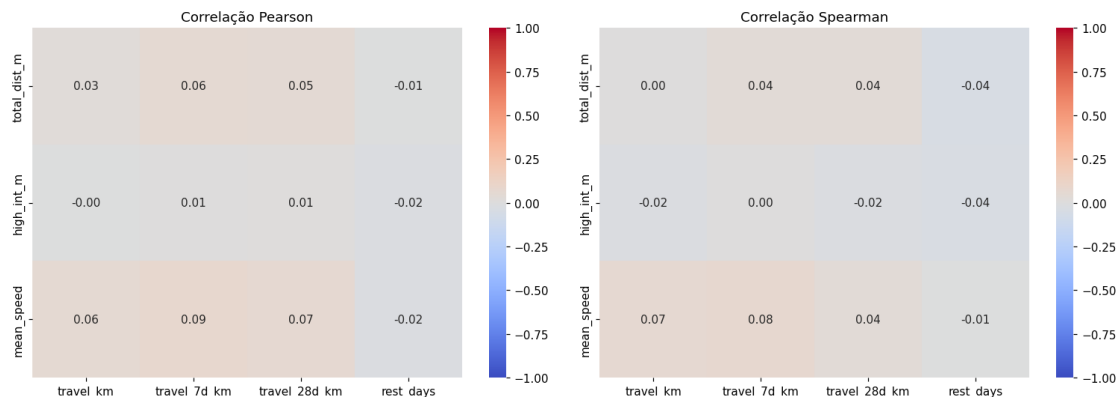


Figure 2: Matriz de correlação entre estresse de viagem e métricas de desempenho físico (Pearson à esquerda, Spearman à direita).

Os valores de correlação apresentados nas matrizes acima indicam associações fracas ou inexistentes entre as variáveis analisadas. Em ambas as abordagens, os coeficientes variaram entre -0.04 e 0.09, o que sugere baixa influência direta das métricas de estresse logístico no desempenho físico registrado.

A maior correlação observada foi entre a variável `mean_speed` e a distância viajada nos últimos 7 dias (`travel_7d_km`), com coeficiente de 0.09 na correlação de Pearson e 0.08 na de Spearman.

5.2 Interpretação dos Resultados

Apesar das expectativas de que maiores distâncias percorridas ou menor tempo de descanso impactariam negativamente o desempenho físico, os dados analisados não sustentam essa hipótese de forma robusta. Os baixos coeficientes de correlação indicam que os efeitos do calendário podem ser mais sutis ou mediados por outros fatores, como estratégias táticas, nível do adversário, ou estrutura logística dos clubes.

6 Análise de Correlação Agregada por Mês

Com o objetivo de suavizar variações individuais por partida e captar tendências sazonais ao longo da temporada, realizamos uma análise agregada por mês, considerando a média das métricas físicas e a soma das métricas de estresse logístico.

6.1 Metodologia

O dataset foi agrupado por `team_id`, `ano` e `mês`, e as seguintes métricas foram calculadas para cada grupo:

- `dist_mo`: média da distância total percorrida por jogo no mês.
- `hi_mo`: média da distância em alta intensidade por jogo no mês.
- `stress_mo`: média da métrica de estresse acumulado antes de cada jogo no mês.
- `travel_mo`: soma total dos quilômetros viajados no mês.
- `games_mo`: número de jogos disputados no mês.

Com esses dados, normalizamos as métricas físicas (dist_norm, hi_norm) dividindo pelas partidas mensais, de modo a obter uma média por jogo no mês. A correlação de Spearman foi então aplicada entre:

- **Variáveis dependentes:** dist_norm e hi_norm
- **Variáveis explicativas:** travel_mo e stress_mo

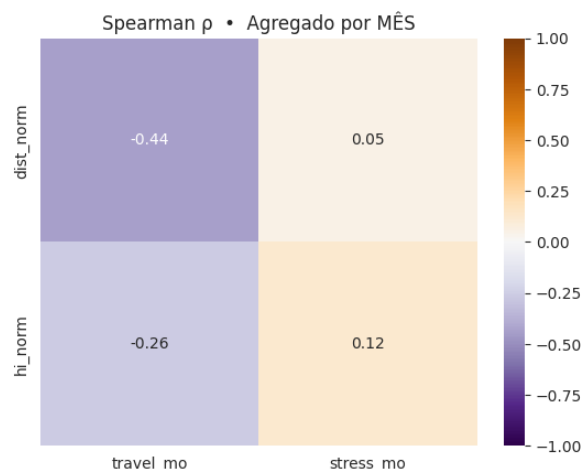


Figure 3: Correlação de Spearman entre métricas de estresse mensal e desempenho físico normalizado por jogo.

6.2 Resultados e Discussão

Os resultados indicam uma correlação negativa moderada entre a variável travel_mo e a distância total percorrida por jogo (dist_norm), com coeficiente de $\rho = -0.44$. Isso sugere que em meses com maior carga de deslocamentos os atletas tendem a apresentar uma leve redução na distância total percorrida em campo.

Da mesma forma, observamos uma correlação negativa entre travel_mo e hi_norm ($\rho = -0.26$) sugerindo que o volume de viagens também pode impactar negativamente os deslocamentos de alta intensidade por jogo.

Por outro lado, a métrica stress_mo, que representa o estresse médio acumulado antes dos jogos, apresentou correlação fraca e positiva com ambas as métricas físicas (0.05 e 0.12, respectivamente), sem evidência de associação clara.

Esses resultados reforçam a hipótese de que o acúmulo logístico de viagens ao longo do mês pode impactar negativamente o desempenho físico médio por partida, ainda que o efeito seja moderado. A análise mensal permite evidenciar tendências que poderiam ser mascaradas por variabilidades pontuais nos dados jogo a jogo.

7 Indicador de Sobrecarga Esportiva (ISE) e Oscilação Física

Nesta etapa do estudo, propomos um índice sintético para representar a sobrecarga física e logística enfrentada por cada equipe ao longo da temporada. Além disso, avaliamos a regularidade do desempenho físico em termos de variação intra-equipe.

7.1 ISE – Indicador de Sobrecarga Esportiva

A Figura 4 (esquerda) mostra a média do ISE ao longo da temporada de 2023 para cada clube da Série A. O clube com maior média foi o Fortaleza, seguido por Bahia e Cuiabá. Por outro lado, equipes como América Mineiro, Vasco da Gama e São Paulo FC apresentaram valores menores de ISE. Um ISE alto indica que essas equipes estiveram mais frequentemente expostas a situações de sobrecarga.

7.2 Oscilação Física

Na parte direita da Figura 4, apresentamos a oscilação física medida como o desvio padrão da métrica ΔHI -dist. Esta métrica captura a variação na intensidade dos jogos disputados por cada clube e reflete quão estável foi o esforço físico ao longo da temporada.

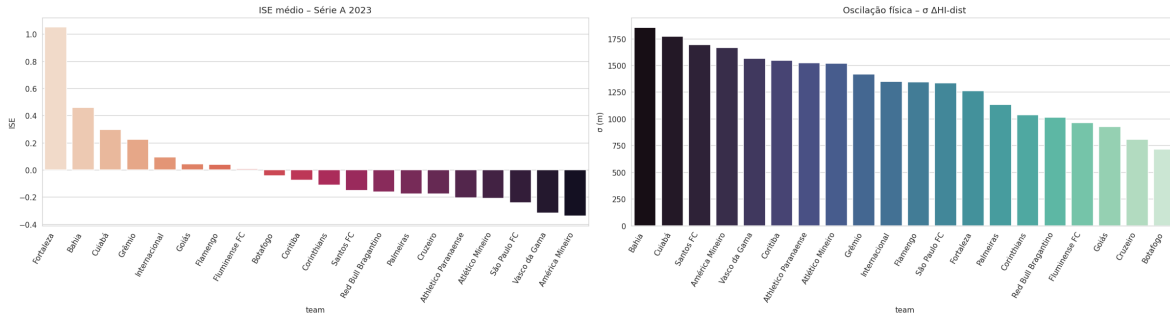


Figure 4: À esquerda: ISE médio por clube. À direita: Oscilação física (σ da diferença entre distâncias total e em alta intensidade por jogo).

7.3 Discussão

Os dados apontam que o Fortaleza, além de liderar em volume de viagens, também apresenta o maior ISE médio, reforçando o cenário de sobrecarga enfrentado pela equipe. A análise combinada do ISE com a oscilação física também pode oferecer outro insight: clubes com alto ISE e alta oscilação podem ter enfrentado desafios tanto em volume quanto em consistência de desempenho físico, o que pode afetar estratégias de recuperação e rotação de elenco.

8 Análise de Clusters: ISE médio $\times$ Oscilação Física

Com o intuito de identificar padrões entre as equipes da Série A em 2023 quanto à exposição ao estresse competitivo e à variabilidade de desempenho físico, foi realizada uma análise de clusters baseada em duas dimensões principais:

- ISE médio: indicador de sobrecarga esportiva, calculado a partir do volume de viagens e dos dias de descanso antes das partidas;
- Oscilação HI-distância: medida de variabilidade na intensidade física dos jogos, (nesse caso calculamos como o desvio padrão da diferença entre distância total e ΔHI -distância).

8.1 Resultados

A Figura 5 mostra os cinco clusters formados:

- Cluster isolado: com o maior ISE médio da liga, o Fortaleza representa um caso extremo de sobrecarga, embora sua oscilação física seja moderada.
- Cluster de alta variabilidade (Bahia, Cuiabá): equipes com grande oscilação física e valores médios de ISE elevados, indicando alta exposição e instabilidade de desempenho.
- Cluster de alta oscilação com menor sobrecarga (Athletico-PR, Coritiba): clubes com baixa média de ISE, mas grande variação na intensidade física dos jogos.
- Cluster central (maioria dos clubes): grupo compacto com médias e variações físicas intermediárias, indicando perfis mais equilibrados.
- Cluster de menor estresse e menor oscilação (Botafogo, Cruzeiro, Goiás): times com ISE médio negativo e baixa variabilidade, refletindo uma temporada mais estável e menos desgastante.

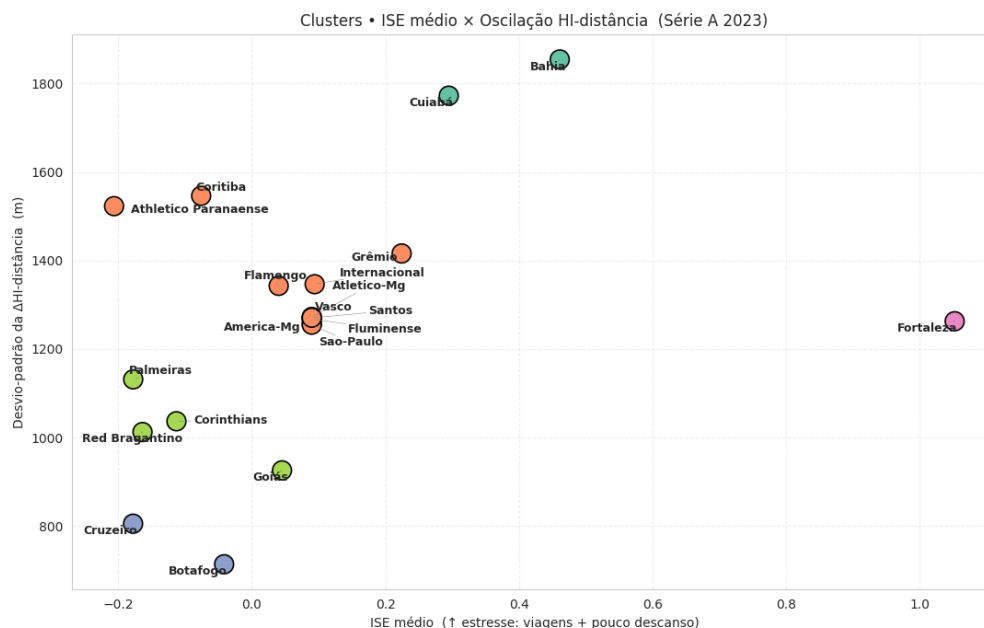


Figure 5: Clusterização dos clubes da Série A 2023 com base no ISE médio e na oscilação de deltaHI-distância.

8.2 Discussão

A análise de clusters revela que as equipes da Série A 2023 enfrentaram realidades logísticas e fisiológicas bastante distintas. Clubes como Fortaleza, Bahia e Cuiabá operaram em cenários extremos de estresse, enquanto outros vivenciaram uma temporada mais estável. Esses padrões podem refletir fatores como localização geográfica, estrutura logística, e desempenho em múltiplas competições.

9 Correlação de Distância entre Estresse Logístico e Desempenho Físico

A fim de capturar possíveis relações não lineares entre variáveis de estresse de viagem e métricas de desempenho físico, aplicamos a correlação de distância (distance correlation), uma medida estatística capaz de detectar associações mesmo quando a relação não é estritamente linear ou monotônica.

9.1 Metodologia

A correlação de distância foi calculada entre as seguintes variáveis:

- Métricas físicas (dependentes):
 - total_dist_m: distância total percorrida (m)
 - high_int_m: distância em alta intensidade (m)
 - mean_speed: velocidade média (m/s)
- Indicadores de estresse logístico (preditoras):
 - travel_km: distância percorrida até o jogo (km)
 - travel_7d_km: soma das distâncias dos últimos 7 dias (km)
 - travel_28d_km: soma das distâncias dos últimos 28 dias (km)
 - rest_days: número de dias de descanso desde o último jogo

A Figura 6 apresenta os coeficientes de correlação de distância entre cada par de variáveis.

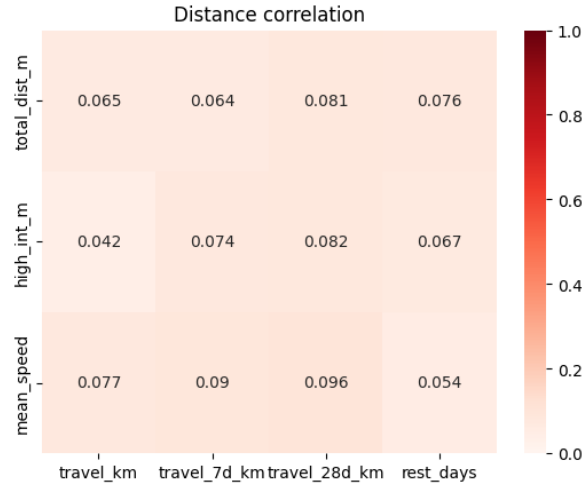


Figure 6: Correlação de distância entre métricas de estresse de viagem e desempenho físico.

9.2 Resultados

Os coeficientes observados foram relativamente baixos em todos os pares avaliados, variando entre 0.042 e 0.096. Os maiores valores foram registrados nas correlações entre:

- travel_28d_km e mean_speed: 0.096
- travel_7d_km e mean_speed: 0.090
- travel_28d_km e high_int_m: 0.082

Esses valores, embora ainda fracos, são ligeiramente superiores aos obtidos por métodos tradicionais como Pearson e Spearman, sugerindo que pode haver relações não lineares ou dependências fracas entre o acúmulo logístico de viagens e o desempenho físico dos atletas.

9.3 Discussão

A correlação de distância confirma os achados anteriores de que não há uma relação forte entre estresse de viagem e desempenho físico direto, mas levanta a possibilidade de que esses efeitos possam se manifestar de maneira mais sutil ou não linear — exigindo modelagens mais sofisticadas, como árvores de decisão ou modelos baseados em aprendizado de máquina para melhor capturar tais padrões.

10 Modelagem Preditiva com Random Forest

Com o objetivo de avaliar o potencial preditivo das variáveis de estresse logístico sobre o desempenho físico dos atletas, foi treinado um modelo de Random Forest Regressor para prever três métricas físicas: distância total percorrida (total_dist_m), distância em alta intensidade (high_int_m) e velocidade média (mean_speed).

10.1 Metodologia

O conjunto de preditores incluiu as seguintes variáveis de estresse:

- rest_days: dias de descanso desde o último jogo
- travel_km: distância percorrida até o jogo (km)
- travel_7d_km: soma das distâncias percorridas nos últimos 7 dias
- travel_28d_km: soma das distâncias percorridas nos últimos 28 dias

A base foi dividida em 80

Os melhores hiperparâmetros encontrados foram:

- `n_estimators = 300`
- `max_depth = 5`
- `min_samples_split = 10`
- `min_samples_leaf = 2`
- `max_features = 'log2'`

10.2 Importância das Variáveis

A importância relativa de cada preditor no modelo foi calculada com base na média da redução de impureza dos nós. Os valores obtidos foram:

- `travel_28d.km`: 34.1
- `rest_days`: 25.6
- `travel_7d.km`: 24.4
- `travel.km`: 15.9

Esses resultados indicam que o acúmulo de viagens nos últimos 28 dias foi o fator mais relevante para o modelo, seguido pelos dias de descanso.

10.3 Desempenho do Modelo

Apesar do processo de ajuste fino dos hiperparâmetros, o desempenho do modelo final foi muito limitado, com $R^2 = 0.0007$ no conjunto de teste. Isso significa que o modelo foi praticamente incapaz de explicar a variabilidade nas métricas físicas com base apenas nas variáveis de estresse logístico utilizadas.

10.4 Discussão

A baixa capacidade preditiva do modelo sugere que fatores adicionais — como decisões táticas, condições climáticas, perfil dos jogadores e estratégia de rotação do elenco — provavelmente desempenham um papel mais relevante no desempenho físico durante as partidas. Ainda assim, a análise de importância das variáveis reforça a relevância de métricas acumuladas como `travel_28d.km` e `rest_days` como potenciais indicadoras de sobrecarga ao longo da temporada.

11 Modelagem Preditiva com XGBoost

Além do modelo de Random Forest, também foi testada uma abordagem baseada em XGBoost — um algoritmo de gradient boosting amplamente utilizado por sua robustez e capacidade de modelar interações não lineares entre variáveis.

11.1 Metodologia

O modelo foi ajustado para prever conjuntamente as três métricas físicas de desempenho dos atletas:

- `total_dist_m`
- `high_int_m`
- `mean_speed`

Como variáveis explicativas, foram utilizadas:

- rest_days
- travel_km
- travel_7d_km
- travel_28d_km

O ajuste dos hiperparâmetros foi feito com `RandomizedSearchCV`, utilizando validação cruzada com 5 folds e 50 combinações aleatórias, maximizando o R^2 . Os melhores hiperparâmetros encontrados foram:

- n_estimators = 500
- learning_rate = 0.05
- max_depth = 3
- subsample = 0.8
- colsample_bytree = 0.7
- gamma = 0

11.2 Importância das Variáveis

A importância das variáveis explicativas, com base no ganho acumulado de cada divisão na árvore, foi distribuída da seguinte forma:

- travel_28d_km: 28.6
- rest_days: 26.5
- travel_7d_km: 22.5
- travel_km: 22.4

Assim como no modelo Random Forest, a variável mais relevante foi o acúmulo de viagens nos últimos 28 dias, seguida pelos dias de descanso.

11.3 Desempenho do Modelo

O modelo final apresentou um desempenho ainda inferior ao Random Forest, com um valor de $R^2 = -0.0824$ no conjunto de teste. Um R^2 negativo indica que o modelo teve desempenho pior do que uma simples média constante — ou seja, não conseguiu capturar nenhuma estrutura útil nos dados para prever as métricas físicas.

11.4 Discussão

O fraco desempenho do XGBoost reforça os achados anteriores: as métricas de estresse logístico consideradas, isoladamente, não explicam suficientemente a variação no desempenho físico dos atletas. Isso sugere a necessidade de inclusão de variáveis contextuais adicionais — como escalas, clima, altitude, e carga de treino — em análises futuras. Ainda assim, a análise de importância das variáveis continua indicando consistência nos efeitos do acúmulo de viagens e do tempo de descanso.

12 Conclusão

Buscamos investigar os impactos do calendário apertado do futebol brasileiro no desempenho físico e competitivo dos clubes da Série A em 2023. Por meio de análises quantitativas envolvendo distância percorrida, intensidade dos deslocamentos e indicadores de desempenho físico dos atletas, não foi possível identificar padrões relevantes ou levantar hipóteses fundamentadas sobre a influência da logística no desempenho das equipes.

Apesar das limitações dos dados e da complexidade de isolar o efeito do calendário, o Indicador de Sobrecarga Esportiva (ISE) permitiu uma visão da carga enfrentada por cada clube e a análise de oscilação física forneceu pistas sobre a regularidade e consistência do desempenho ao longo da temporada.

A análise de clusters evidenciou que os times não enfrentam a temporada em condições iguais: clubes como Fortaleza, Bahia e Cuiabá operaram sob condições logísticas e fisiológicas muito mais exigentes do que outros, o que reforça a necessidade de considerar essas desigualdades nas discussões sobre competitividade e justiça esportiva.

Como contribuição final, este trabalho oferece dados quantitativos para o debate sobre a reformulação do calendário do futebol nacional. Um planejamento mais racional e sustentável pode não apenas preservar a integridade física dos atletas, como também elevar o nível técnico das competições, promovendo um futebol mais justo, competitivo e atrativo.

Referências

1. **Fullagar, H. H. K.**, Skorski, S., Duffield, R., Julian, R. Meyer, T. *Sleep and athletic performance: the effects of sleep loss on exercise performance and physiological and cognitive responses to exercise.* **Sports Medicine**, 45(2), 161-186 (2015).
2. **Fowler, P. M.**, Duffield, R., Coutts, A. J. Vaile, J. M. *Effects of long-haul transmeridian travel on subjective jet-lag and wellness in professional soccer players.* **International Journal of Sports Physiology and Performance**, 15(4), 553-560 (2020).
3. **Dupont, G.**, Nedelec, M., McCall, A., McCormack, D., Berthoin, S. Wisløff, U. *Effect of two soccer matches in a week on physical performance and injury rate.* **The American Journal of Sports Medicine**, 38(9), 1752-1758 (2010).
4. **Gabbett, T. J.** *The training-injury prevention paradox: should athletes be training smarter and harder?* **British Journal of Sports Medicine**, 50(5), 273-280 (2016).
5. **Bowen, L.**, Gross, A. S., Gimpel, M. Li, F.-X. *Accumulated workloads and the acute:chronic workload ratio relate to injury risk in elite youth football players.* **British Journal of Sports Medicine**, 54(4), 281-287 (2020).
6. **Lago-Peñas, C.**, Lago, I., Rey, E. Sampaio, J. *How travel distance and match location affect performance of professional soccer players.* **Journal of Sports Sciences**, 36(19), 2213-2218 (2018).
7. **Samuels, C.** *Jet lag and travel fatigue: Management for sport medicine physicians and high-performance support teams.* **Clinics in Sports Medicine**, 38(3), 543-553 (2019).
8. **Pollard, R.**, Silva, C. D. & Medeiros, N. C. *Home advantage in football in Brazil: differences between teams and the effects of distance traveled.* **Revista Brasileira de Futebol – Brazilian Journal of Soccer Science**, 1(1), 3–10 (2008).

Mapping and Characterization of Defensive Pressing in Football

Júlia Paes, Lucas de Oliveira, Lucas Affonso, Bruno Buti

Federal University of Minas Gerais (UFMG), Department of Computer Science

juliapaes@dcc.ufmg.br

Abstract. Defensive pressure is a tactical cornerstone of modern football, yet its quantification remains a challenge. Traditional metrics, often based on player proximity to the ball, fail to capture the spatial dynamics and collective intentionality of defensive actions. This paper proposes a pressing intensity model based on tracking data, calculating the probability of interception for each defender on every attacker and the ball at each game frame. Using data from a Premier League 23/24 match, we applied this model to identify moments of maximum pressure, characterizing them by pitch location, numerical superiority, and success rate derived from event data. While the initial analysis was limited to a single match to establish the methodology, our findings demonstrate that this model is a promising and little-explored approach to objectively dissecting defensive strategies and provides a decent framework for further investigation.

1 Introduction

The tactical evolution of contemporary football has established organized defensive pressure as a key strategic weapon. However, objectively quantifying pressure is an analytical challenge, as traditional event-based metrics (like tackles) provide a limited view and often miss the value of good positioning that forces errors without a registered action. The availability of high-frequency tracking data opens a new frontier for tactical analysis, allowing for the modeling of complex spatial concepts and the intensity of pressure at every moment.

This paper's central contribution is the empirical demonstration that "high-pressure" moments, identified by a already proposed model applied to tracking data, are not just an increase in defensive intensity but can represent a distinct tactical strategy. We analyze whether these events show different patterns of location, success, and dependency on numerical superiority compared to standard defensive actions.

2 Modeling and Methodology

To capture the dynamic nature of defensive pressure, we implemented a model based on the work of Joris Bekkers [Bekkers 2024] and William Spearman [Spearman 2018]. The core of the model is calculating the "Time to Intercept"—the time a defender would take to reach a moving attacker or the ball. This calculation is more realistic than simple distance, as it incorporates the players' current velocity, reaction time (0.4s used as default), and

penalties for changes in the direction of the movement (if it is against the ball's). This time value is then converted into an interception probability ($p_{i,j}$) using a sigmoid function. Finally, the individual probabilities of all defenders on the ball carrier are aggregated to produce a single, frame-by-frame `ball_pressure` metric, which quantifies the total collective pressure.

The analysis was conducted using tracking data from a Manchester City vs. Burnley Premier League match (2023/24). Due to the high computational cost, the 1.3 million row dataset was strategically subsampled, while ensuring all key defensive events (`DUEL`, `TAKE ON`, `CLEARANCE`) were preserved to maintain tactical integrity. Ideally, more tracking events would be included. Finally, defensive success was defined based on the outcome of these events; for instance, a clearance was always a success, while a duel was a success if the attacker failed.

3 Results: A Tale of Two Defending Styles

Our analysis revealed a sharp contrast between general defensive actions and moments of high pressure.

The **general defensive profile** (Table 1, approx. 800 events) was predominantly reactive. About 75% of actions occurred in the team's own defensive third, with the highest success rate (47.6%) also in this zone. This suggests a standard defensive posture of reacting to threats near the goal, where success is also aided by greater player density and numerical superiority as analyzed.

In contrast, the **high-pressure profile** (Table 2, approx. 70 events with the highest `ball_pressure` scores) was drastically different. These actions were proactive, with defensive activity shifting massively to the midfield (64%) and the offensive third (29%). This indicates a deliberate strategy to contest possession in advanced areas rather than reacting to danger. Counter-intuitively, the immediate success rate of these actions was significantly lower, dropping to 25% in the midfield and 0% in the attacking third. Furthermore, besides the correlation between pitch location and success strengthens, increases from 0.27 to 0.33 in the high-pressure sample (Table 3), peak success occurred not with numerical superiority, but in situations of numerical equality, suggesting that coordinated structure is more critical than sheer numbers in these scenarios.

Table 1: Descriptive Analysis and Success Rate by Pitch Third (Full Sample, N≈800).

Pitch Third (Relative to Defense)	Frequency (N)	Frequency (%)	Success Rate (%)
1 (Defensive)	≈600	≈75%	47.62%
2 (Middle)	≈150	≈19%	-
3 (Offensive)	≈50	≈6%	20.00%

Table 2: Descriptive Analysis and Success Rate by Pitch Third (High-Pressure Events, $N \approx 70$).

Pitch Third (Relative to Defense)	Frequency (N)	Frequency (%)	Success Rate (%)
1 (Defensive)	≈ 5	$\approx 7\%$	-
2 (Middle)	≈ 45	$\approx 64\%$	25.00%
3 (Offensive)	≈ 20	$\approx 29\%$	0.00%

Table 3: Comparative Correlation Matrices of Game Variables.

Full Sample ($N \approx 800$)				
	Owning Pl.	Defending Pl.	Pitch Third	Success Rate
Owning Pl.	1.00	0.54	0.24	-0.25
Defending Pl.	0.54	1.00	0.17	-0.13
Pitch Third	0.24	0.17	1.00	0.27
Success Rate	-0.25	-0.13	0.27	1.00
High-Pressure Sample ($N \approx 70$)				
	Owning Pl.	Defending Pl.	Pitch Third	Success Rate
Owning Pl.	1.00	0.54	0.39	-0.22
Defending Pl.	0.54	1.00	-0.15	-0.12
Pitch Third	0.39	-0.15	1.00	0.33
Success Rate	-0.22	-0.12	0.33	1.00

4 Discussion and Conclusion

The results empirically suggest that high pressure is a qualitatively distinct, proactive tactical mode, confirmed by its location in advanced zones. The paradox of its low immediate success rate suggests its primary goal is disruption—forcing poor passes—rather than direct ball recovery, relying more on tactical structure than numerical superiority.

While this study presents a quantitative framework, the model’s high computational cost limited the analysis to a subset of events from a single match, and therefore its findings should be considered preliminary. Moreover, the definition of “defensive success” was deliberately restricted to clear, immediate possession-recovery events (DUEL, TAKE ON, CLEARANCE), a metric which may fail to capture the primary objective of disruptive tactics like the high press. Consequently, the observed low success rate is likely an artifact of the metric rather than a tactical failure.

Future work should expand this analysis by applying the model across multiple games to compare different teams’ defensive profiles. This broader application could be enhanced by incorporating refined success metrics, like forced errors, and other contextual features that would better describe the pressing moments. Moreover, a significant opportunity lies in leveraging the model’s all-to-all pressure calculation. This would enable a granular evaluation of individual players’ defensive styles and profiles, moving

beyond team-level tactics to offer a more complete picture of effectiveness.

References

- [Bekkers 2024] Bekkers, J. (2024). Pressing Intensity: An Intuitive Measure for Pressing in Soccer. *UnravelSports Blog*.
- [Decroos et al. 2019] Decroos, T., et al. (2019). Actions Speak Louder than Goals: Valuing Player Actions in Soccer. *KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.
- [Fernández & Bornn 2018] Fernández, J., & Bornn, L. (2018). Wide Open Spaces: A statistical technique for measuring space creation in professional soccer. *MIT Sloan Sports Analytics Conference*.
- [Friends of Tracking n.d.] Friends of Tracking. (n.d.). *Friends of Tracking GitHub Repository*. GitHub.
- [Higgins et al. 2023] Higgins, L., Galla, T., Prestidge, B., & Wyatt, T. (2023). Measuring the pitch control of professional football players using spatiotemporal tracking data. *Journal of Physics: Complexity*, 4(2).
- [Spearman 2018] Spearman, W. (2018). Beyond Expected Goals. *MIT Sloan Sports Analytics Conference*.
- [Spearman et al. 2017] Spearman, W., Basye, A., Dick, G., Hotovy, R., & Pop, P. (2017). Physics-Based Modeling of Pass Probabilities in Soccer. *MIT Sloan Sports Analytics Conference*.
- [Sumpter 2016] Sumpter, D. (2016). *Soccermatics: Mathematical Adventures in the Beautiful Game*. Bloomsbury Publishing.

1. Problema e Objetivo Geral

O Liverpool Football Club enfrentava o desafio de encontrar um lateral-direito que apresente características técnicas e táticas similares às de Trent Alexander-Arnold (TAA), jogador essencial no modelo de jogo do clube, especialmente na construção ofensiva e distribuição de passes. Considerando fatores como eventual ausência por lesões, desgaste físico ou até possibilidades futuras de transferência, é fundamental antecipar a reposição desse atleta-chave.

O objetivo geral deste trabalho é identificar, por meio de análise de dados, técnicas de clusterização e avaliação de mercado (valuation), jogadores com perfil semelhante a TAA, que possam ser potenciais alvos de contratação para o Liverpool, mantendo o alto padrão de desempenho exigido.

2. Metodologia

A metodologia adotada foi estruturada em quatro etapas principais:

- **2.1. Coleta de Dados:**

Foi realizada a coleta de dados estatísticos de laterais-direitos que atuam nas sete principais ligas europeias (Premier League, La Liga, Serie A, Bundesliga, Ligue 1, Eredivisie e Liga Portugal), utilizando plataformas como FBref, Wyscout e Transfermarkt. As variáveis selecionadas incluíram métricas ofensivas (duelos ofensivos, corridas progressivas, conduções para o último terço, dribles bem sucedidos), defensivas (desarmes, interceptações, duelos ganhos) e de construção (passes curtos, médios, longos, cruzamentos, passes para o terço final, passes para a área do penalti.).

- **2.2. Processamento e Normalização dos Dados:**

Os dados foram normalizados para garantir a comparabilidade entre atletas, levando em conta tanto minutos jogados quanto ajustes de contexto (atletas com no mínimo 600 minutos de temporada + consideração de nível de liga e teto).

- **2.3. Clusterização:**

Aplicou-se o algoritmo de clusterização K-Means, buscando agrupar os laterais em perfis baseados nas características de jogo.

-

- **2.4. Análise de Mercado e Valuation:**

Após identificar os jogadores pertencentes ao mesmo cluster de TAA (perfil mais semelhante), foi feita uma análise de mercado. Essa etapa considerou valor, idade, tempo de contrato, potencial técnico e financeiro e nomes ventilados em janelas anteriores.

3. Resultados

O processo de clusterização identificou **5 clusters principais** entre os laterais-direitos analisados:

- **Cluster 1:** Laterais equilibrados, com atribuições defensivas e ofensivas semelhantes, sendo pouco incisivos no ataque, optando por apoiar na zona intermediária.
- **Cluster 2:** Laterais com alta atribuição defensiva, muitos duelos defensivos vencidos, por vezes, zagueiros de origem que se tornam laterais.
- **Cluster 3:** Laterais construtores, com alto volume de passes progressivos, inversões e participação na fase ofensiva, perfil em que TAA está inserido.
- **Cluster 4:** Alas-Ofensivos, perfil de Frimpong(sucessor de TAA), laterais com alta presença no terço final, entrada na área e finalização de média/curta distância.
- **Cluster 5:** Laterais físicos, alta capacidade de vencer duelos, cobrir o campo e pisar na área, exemplo: Dumfries.

Dentro do **Cluster 3 (construtores)**, foram encontrados alguns jogadores que apresentam métricas próximas às de TAA. Após cruzamento com dados de mercado, destacaram-se três atletas com forte aderência ao perfil buscado:

1. **Pedro Porro(Tottenham Spurs)** – Destaca-se pela capacidade de construção próxima a de Arnold, com ótima qualidade para sair com a bola e encontrar espaços que poucos laterais encontrariam.
2. **Yukinari Sugawara(Southampton)** – Atleta mais “cru” para o contexto que o Liverpool busca, mas também, outro jogador com alta capacidade de construir o jogo a partir da defesa, embora seja menos físico para o jogo do Liverpool(alta pressão).
3. **Conor Bradley(Liverpool)** – Jogador jovem, taticamente adaptado e no mesmo agrupamento que arnold, ainda que esteja na fronteira com os alas, Conor demonstra similaridades com o TAA, principalmente em seu jogo sem bola e a capacidade de atuar internamente.
4. **Jeremie Frimpong(Bayer Leverkusen)** - Embora não estivesse no mesmo cluster que Arnold(e com razão) foi considerado por conta de sua influência no ataque e como uma alternativa caso o LFC buscasse um outro estilo, e não um substituto igual a TAA. Estava no agrupamento de alas-ofensivos, tendo o perfil ofensivo mais dominante entre todos os atletas vistos.

Outros nomes também foram levantados como opções secundárias, considerando uma análise de custo-benefício e adaptação ao modelo do Liverpool(Ex: Givairo Read, Vanderson, Ratiu, Dumfries).

4. Conclusões

A utilização de clusterização se mostrou extremamente eficiente para segmentar os laterais-direitos de acordo com seus perfis técnicos e táticos, permitindo uma análise mais

precisa e objetiva. A metodologia adotada confirmou que o perfil de TAA é raro no mercado, o que implica custos elevados para jogadores que apresentam características muito próximas.

Entretanto, o trabalho conseguiu oferecer alternativas viáveis, equilibrando fatores técnicos, táticos e financeiros. Outro fator importante foi reparar que os dados no futebol não são o fim, e sim um meio, pois o que mais vale para a análise é assistir e concluir por conta, já que o contexto não aparece em gráficos.

Por fim, a aplicação de ciência de dados no processo de scouting demonstra ser uma ferramenta poderosa, não apenas na busca por atletas, mas também na otimização dos processos decisórios dentro dos clubes de futebol. Para além disso, traçar perfis, permitiu uma indicação de nome por estilo, o que traria um mundo de opções para quem toma a decisão.

Sessão de Best Papers

Shared Goal Value: Uma Nova Métrica para Avaliar Contribuições em Gols no Futebol

1

1. Introdução

O futebol tem passado por uma profunda transformação impulsionada pelo uso crescente de análise de dados. Com a profissionalização do esporte, clubes têm investido em ferramentas analíticas para explicar e otimizar o desempenho individual e coletivo, utilizando técnicas avançadas de coleta, processamento e modelagem de dados, incluindo métricas baseadas em aprendizado de máquina [Decroos et al. 2019].

Devido a natureza do futebol, que é um esporte de poucos gols e alta complexidade tática, avaliar o impacto de ações individuais representa um desafio considerável. Métricas simples como número de assistências, finalizações ou passes certos não levam em conta o contexto da jogada, o grau de dificuldade da ação e, de certa forma, não diferenciam o impacto de cada jogador. Por isso, pesquisadores passaram a desenvolver novas métricas que buscam quantificar o valor de ações específicas, como passes [Anzer and Bauer 2022] e assistências [The Football Analyst 2022], com base em seu impacto potencial no resultado da jogada.

Apesar do avanço nas métricas ofensivas, grande parte das análises ainda concentra sua atenção no jogador que finaliza a jogada. Métricas como o xG (expected goals)[Mead et al. 2023] e suas variantes valorizam o posicionamento e a execução do chute, mas acabam não considerando o papel do assistente, que em diversas ocasiões, realiza uma jogada que quebra as linhas adversárias, essencial para a construção do gol. Mesmo métricas como o xA (expected assists) [The Football Analyst 2022] focam na probabilidade de um passe gerar um gol, mas não tratam da divisão de contribuição entre o assistente e o artilheiro no lance em que o gol realmente ocorre. Assim, torna-se relevante propor uma métrica que busque avaliar o impacto de uma assistência no gol marcado, oferecendo uma visão mais equilibrada e justa sobre a contribuição dos jogadores envolvidos na jogada decisiva.

Além desta introdução, o artigo está estruturado em três seções. A Seção 2 detalha os métodos utilizados no estudo. A Seção 3 descreve os resultados obtidos. Finalmente, a Seção 4 discute as principais conclusões e sugere direções para pesquisas futuras.

2. Metodologia

2.1. Dados de Eventos

No futebol, os dados de eventos são dados que representam uma sequência de ações com a bola durante uma partida. No presente trabalho, foram utilizados os dados de evento aberto da temporada 2015-16 da Bundesliga, a primeira divisão do campeonato alemão, oferecidos pelo StatsBomb ¹.

¹https://www.hudl.com/en_gb/products/statsbomb

Para o entendimento da métrica proposta, outras três métricas são importantes: Gols Esperados no Alvo (xGOT), Ameaça Esperada (xT) e valor de Shapley.

A métrica Gols Esperados no Alvo, conhecida como xGOT, mede a probabilidade de um chute que foi no alvo resultar em gol, com base na qualidade da chance, que é medida pelo xG e a localização final do chute dentro da área do gol. Ela beneficia os chutes que foram nos cantos do gol e penaliza os chutes nas áreas centrais.

A métrica Ameaça Esperada, conhecida como xT [Singh 2018], é um modelo que busca quantificar o valor de cada ação em uma partida de futebol. Diferente de métricas como o xG que se concentram apenas no momento do chute, o xT avalia como a posse de bola em uma determinada zona do campo aumenta a probabilidade de um gol ocorrer no futuro, seja através de uma finalização direta ou de movimentos subsequentes.

O Valor de Shapley [Shapley 1953] é um conceito da teoria dos jogos que oferece um método matemático para definir a participação ou o ganho total de uma colaboração entre seus participantes. A sua principal vantagem é que ele não considera apenas a contribuição individual de cada um, mas também o valor que eles geram ao trabalharem em conjunto. O cálculo para uma colaboração entre duas pessoas, que será feito no presente trabalho, pode ser realizado usando a equação a seguir:

$$\phi_{At} = \frac{[v(\{At\}) - v(\emptyset)] + [v(\{At, As\}) - v(\{As\})]}{2} \quad (1)$$

$$\phi_{As} = \frac{[v(\{As\}) - v(\emptyset)] + [v(\{At, As\}) - v(\{At\})]}{2} \quad (2)$$

Na qual, pode-se descrever as variáveis da seguinte forma:

- $v(\emptyset)$: O valor gerado se ninguém participar.
- $v(\{At\})$: O valor gerado apenas pelo Atacante.
- $v(\{As\})$: O valor gerado apenas pela Assistente.
- $v(\{At, As\})$: O valor gerado quando o Atacante e o Assistente colaboram.

2.2. Shared Goal Value

A métrica Shared Goal Value (SGV) tem como objetivo calcular a participação do assistente e do artilheiro em uma jogada que resultou em gol. A realização do cálculo do SGV é feita utilizando o valor de Shapley. Para isso é necessário determinar os valores de $v(\{At\})$, $v(\{As\})$ e $v(\{At, As\})$. O $v(\emptyset)$ foi considerado igual a zero. Isso acontece porque, sem a participação do assistente e do artilheiro, a jogada não aconteceria e não existiria gol. As tabelas 1 e 2 demonstram como foram realizados os cálculos de $v(\{At\})$ e $v(\{As\})$. Para calcular $v(\{At, As\})$ basta somar os resultados de $v(\{At\})$ e $v(\{As\})$.

Table 1. Participação Atacante

Ações	Cálculo
Chute	$xGOT - xG$
Chute + Condução	$(xGOT - xG) + xT$

Table 2. Participação Assistente

Ações	$v(\{As\})$
Passe	xT_{Passe}
Passe + Condução	$xT_{Passe} + xT_{Conducao}$

Após o cálculo dos valores de Shapley para o atacante e assistente, os números são normalizados para somarem 1.

3. Resultados

Esta seção apresenta os principais resultados obtidos com a aplicação da métrica Shared Goal Value (SGV) aos gols da temporada 2015-16 da Bundesliga. Primeiramente, são analisadas três jogadas representativas, cada uma ilustrando diferentes padrões de contribuição entre assistente e artilheiro. Em seguida, realiza-se uma análise agregada que relaciona a contribuição total dos jogadores com seus respectivos números de gols e assistências, permitindo avaliar sua eficiência em jogadas decisivas. Por fim, foi realizada uma análise agregada dos jogadores com maior contribuição total em jogadas de gol.

3.1. VfB Stuttgart 1 x 4 Eintracht Frankfurt

Foi selecionado o lance do gol de Daniel Didavi pelo Stuttgart, com assistência de Sereso Dié. Nesse lance houve uma longa condução do assistente até a área que gerou 0.06 de xT e então foi feito um passe dentro da área que aumentou consideravelmente o valor de xT em aproximadamente 0.23. O artilheiro da jogada apenas chutou a bola e, portanto, teve uma baixa diferença entre $xGOT$ e xG de apenas 0.06. Após término de cálculo da métrica, o resultado foi uma contribuição de 0.17 para Didavi e 0.83 para Dié. O lance está exibido na Figura 1.

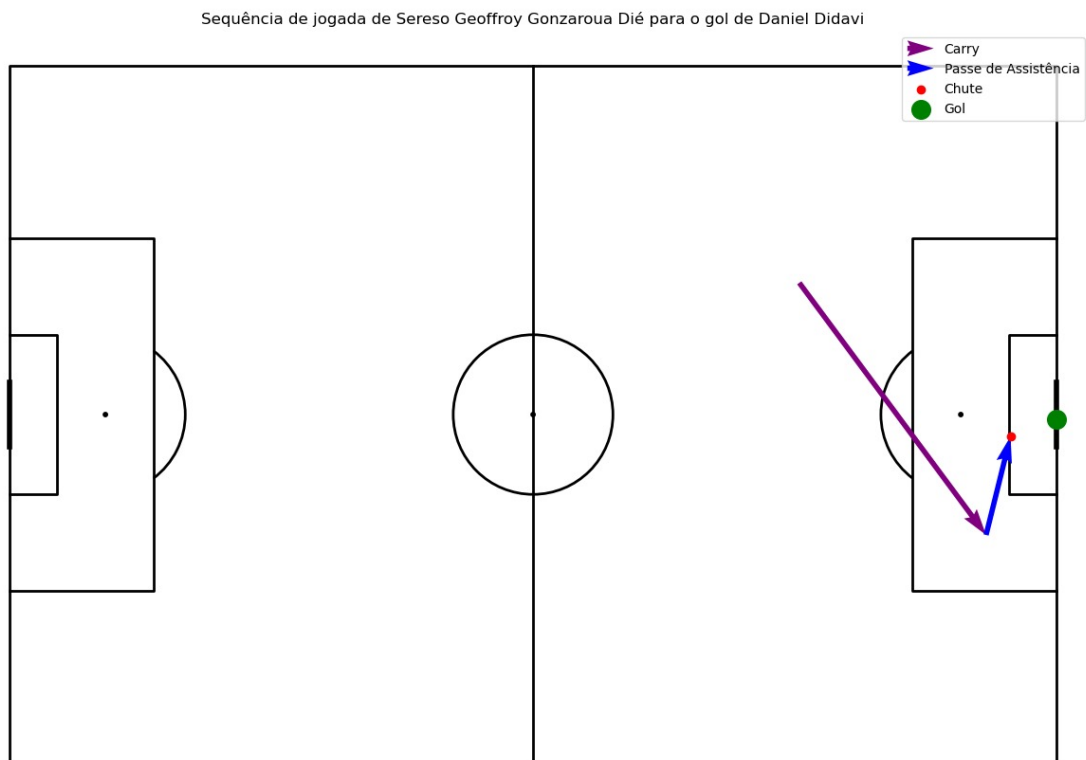


Figure 1. Dié recebeu a bola em frente a área e a conduziu até a área esquerda do gol, dando um passe perigoso na pequena área para a finalização de Didavi.

3.2. Hertha Berlin 2 x 2 Hannover 96

Foi selecionado o lance do gol de Vedad Ibišević pelo Hertha, com assistência de Marvin Plattenhardt. Nesse lance houve uma pequena condução do assistente na lateral que gerou 0.03 de xT, e então foi feito um cruzamento para a área que aumentou bastante o valor de xT em aproximadamente 0.27. O artilheiro da jogada chutou com qualidade, gerando uma diferença entre xGOT e xG de 0.33. Após cálculo da métrica, o resultado foi uma contribuição de 0.50 para cada um dos jogadores. O lance está exibido na Figura 2.

3.3. Hamburger SV 0 x 1 Schalke 04

Foi selecionado o lance do gol de Leroy Sané pelo Schalke, com assistência de Leon Goretzka. Nesse lance houve uma condução e um passe do assistente que geraram 0.09 de xT. O artilheiro Sané carregou a bola dentro da área, aumentando o xT em 0.08 e chutou com qualidade, gerando uma diferença de 0.26 entre xGOT e xG (logo, a contribuição de Sané foi 0.09+0.26). Após cálculo da métrica, o resultado foi uma contribuição de 0.78 para Sané e 0.22 para Goretzka. O lance está exibido na Figura 3.

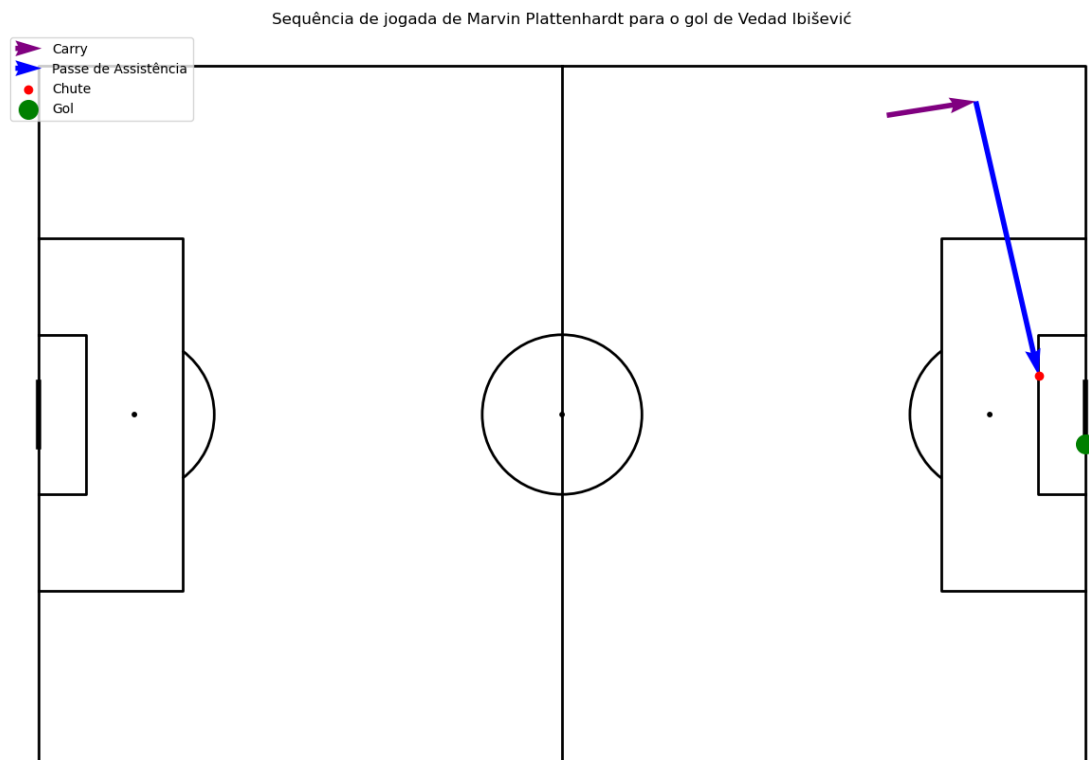


Figure 2. Plattenhardt conduziu a bola na lateral do campo e cruzou para a pequena área, permitindo uma boa finalização de Ibišević.

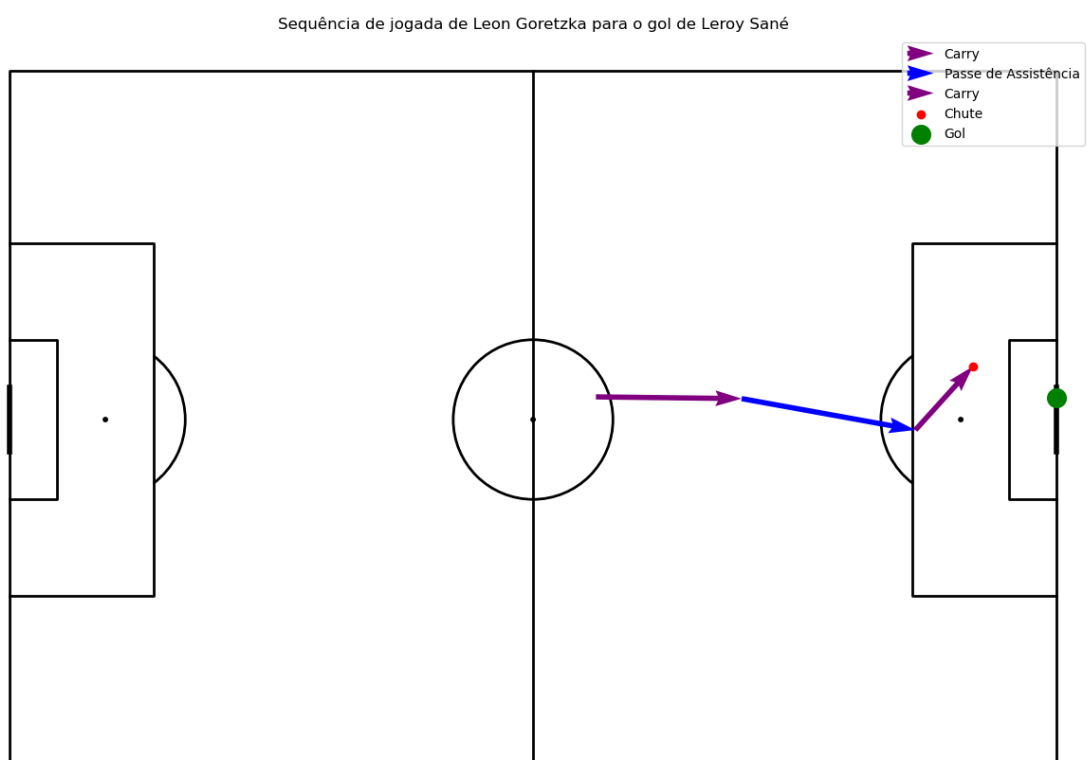


Figure 3. Goretzka conduziu e deu um passe na entrada da área, onde Sané conduziu e finalizou com qualidade.

3.4. Jogadores e Contribuição

As Tabelas 3 e 4 mostram os jogadores com maior contribuição na hora do gol, seja na hora de chutar ou de dar a assistência. A razão (contribuição total/gols) nos mostra a importância dos jogadores nos lances que participaram, permitindo medir a sua eficiência em jogadas de gol.

Table 3. Artilheiros com maior contribuição de gol

Artilheiro	Contribuição Total	Gols	Razão
Pierre-Emerick Aubameyang	12,761842	20	0,638092
Robert Lewandowski	9,103944	18	0,505775
Anthony Modeste	7,332949	12	0,611079
Javier Hernández Balcázar	7,010364	12	0,584197
Salomon Armand Magloire Kalou	6,486909	10	0,648691

Table 4. Assistentes com maior contribuição de gol

Assistente	Contribuição Total	Assistências	Razão
Douglas Costa de Souza	5,429543	9	0,603283
Henrikh Mkhitaryan	4,133648	11	0,375786
Karim Bellarabi	3,864935	10	0,386494
Pierre-Emerick Aubameyang	3,449480	5	0,689896
Hiroshi Kiyotake	3,289390	6	0,548232

4. Conclusão

Os resultados obtidos demonstram que a métrica Shared Goal Value (SGV) é capaz de atribuir de forma consistente e interpretável os valores de participação de assistentes e artilheiros em jogadas de gol. A abordagem se destaca por permitir uma análise mais justa e detalhada do desempenho individual dos jogadores, superando limitações de métricas tradicionais que tratam essas contribuições de forma isolada ou simplificada.

Embora os resultados sejam promissores, há espaço para aprimoramentos. Trabalhos futuros devem explorar conjuntos de dados mais amplos e diversificados, bem como incorporar novas variáveis contextuais e métricas complementares. Essas extensões podem refinar ainda mais a capacidade preditiva e explicativa do SGV, tornando-o uma ferramenta valiosa para análise tática, scouting e tomada de decisão no futebol profissional.

References

- Anzer, G. and Bauer, P. (2022). Expected passes. *Data Mining and Knowledge Discovery*, 36(1):295–317.
- Decroos, T., Bransen, L., Van Haaren, J., and Davis, J. (2019). Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 1851–1861. ACM.
- Mead, J., O'Hare, A., and McMenemy, P. (2023). Expected goals in football: Improving model performance and demonstrating value. *PLOS ONE*, 18(4):1–29.
- Shapley, L. S. (1953). A value for n-person games. In Kuhn, H. W. and Tucker, A. W., editors, *Contributions to the Theory of Games II*, volume 28 of *Annals of Mathematics Studies*, pages 307–317. Princeton University Press, Princeton, NJ.
- Singh, K. (2018). Introducing expected threat (xt). Blog post – karun.in/blog/expected-threat.html.
- The Football Analyst (2022). Expected assists (xa) – football statistics explained. Accessed: 2025-08-01.

xCross: Um Modelo Probabilístico Para Avaliação de Cruzamentos através da Entropia

**Jalmir de Jesus Ferreira da Silva Junior¹ Vitor A. R. Azevedo¹,
Lucas V. S. Ramos¹, Eduardo A. Tomich¹, Bernardo S. B. de Oliveira¹**

¹ Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, MG – Brazil

Abstract. *Crosses are high-impact plays in football, often associated with significant scoring opportunities. However, their dynamic and collective nature poses challenges for statistical modeling, making it difficult to accurately capture player interactions. In this study, we introduce xCross, a probabilistic model designed to estimate the expected value of a cross. As a key contribution, we propose a novel positional entropy metric that quantifies the degree of defensive disorganization during the play, which plays a critical role in determining offensive success.*

Resumo. *Cruzamentos são jogadas de alto impacto no futebol, frequentemente associadas a oportunidades significativas de finalização. No entanto, sua natureza dinâmica e coletiva impõe desafios à modelagem estatística, dificultando a captura precisa das interações entre jogadores. Neste trabalho, apresentamos o xCross, um modelo probabilístico desenvolvido para estimar o valor esperado de um cruzamento. Como contribuição central, introduzimos uma nova métrica de entropia posicional, capaz de quantificar o grau de desorganização defensiva durante a jogada, fator que exerce papel fundamental no sucesso ofensivo.*

1. Introdução

Apesar de sua frequência no futebol, cruzamentos são estatisticamente pouco eficientes. A literatura (Vecer, 2014) aponta que, na Premier League, a média é de apenas um gol a cada 92 cruzamentos. Essa discrepância demonstra que as análises tradicionais, que se baseiam apenas em dados de eventos, não conseguem capturar a complexidade total da jogada. O valor de um cruzamento não se resume a um passe, mas sim ao contexto posicional e tático em que ele ocorre.

Com essa perspectiva, apresentamos o **xCross**, um modelo probabilístico que busca quantificar o valor esperado de um cruzamento. Nossa principal inovação é o uso da **entropia espacial**, uma medida que quantifica a desorganização defensiva. Acreditamos que, quanto maior a entropia no momento do cruzamento, mais vulnerável a defesa se encontra, e maior a chance de a jogada resultar em sucesso.

O modelo foi desenvolvido em duas versões, utilizando dados de rastreamento e eventos da temporada 2024 da Premier League:

- **xCross:** Estima o potencial da jogada no instante do cruzamento, avaliando o arranjo espacial dos jogadores sem considerar o destino final da bola.
- **xCross_on_Target:** Refina a estimativa inicial ao incluir a posição em que a bola realmente chegou, oferecendo uma análise mais precisa da trajetória.

O principal objetivo do nosso trabalho é demonstrar que a **entropia** pode ser uma métrica eficaz para medir a desorganização defensiva e, consequentemente, o potencial de uma jogada. Nosso modelo contribui para a evolução da análise esportiva, oferecendo uma ferramenta mais sofisticada e contextual para a avaliação de cruzamentos.

2. Metodologia

2.1. Padronização dos Dados

Para garantir consistência espacial entre as jogadas, padronizamos todos os lances para ocorrerem em uma única direção de ataque. As coordenadas dos eventos e do *tracking* foram transformadas de forma que todos os cruzamentos se desenvolvem da esquerda para a direita no campo. Essa abordagem nos permite comparar jogadas de forma consistente e focar em padrões espaciais, independentemente do lado do campo em que elas ocorreram.

2.2. Definição de Cruzamento

Adotamos uma definição sistemática de cruzamento, em vez de depender de anotações manuais. Consideramos como cruzamento qualquer passe realizado a partir de uma zona específica do campo (definida empiricamente) e direcionado para a área adversária. A delimitação exata das zonas de origem e destino será apresentada na Figura 6. Essa abordagem permite capturar cruzamentos que fogem do padrão convencional, mas que mantêm a mesma intenção tática.

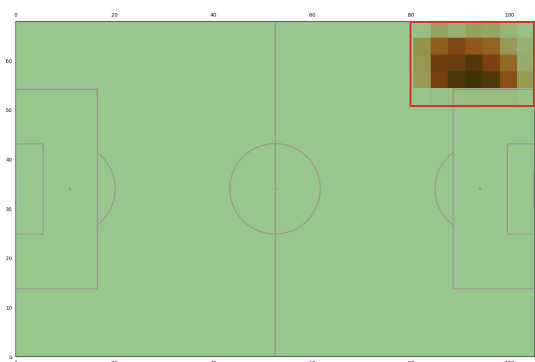


Figure 1. Região de início do cruzamento

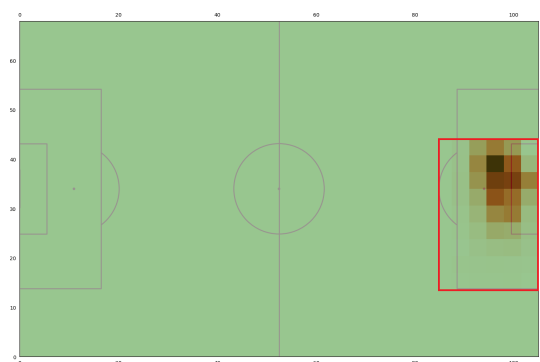


Figure 2. Região de destino do cruzamento

Figure 3. Zonas de origem e destino dos cruzamentos.

2.3. Definição de Sucesso

Um cruzamento é considerado bem-sucedido se a posse de bola imediata após a jogada for mantida pela equipe atacante. O sucesso é, portanto, determinado pela continuidade da posse de bola, independentemente de qual jogador a receba.

2.4. Features Extraídas a partir dos Dados de Eventos

A partir dos dados de eventos, que registram informações pontuais e posições de passes, extraímos as seguintes *features* para descrever a jogada:

- **action_1**: Tipo da ação principal (ex: cruzamento, passe curto).

- **start_x_1, start_y_1, end_x_1, end_y_1**: Coordenadas de início e fim da ação principal.
- **cross_region**: Região do campo de onde o cruzamento é originado.
- **polar_angle_entry**: Ângulo que representa a direção de entrada do jogador na zona de cruzamento.
- **polar_angle_cross**: Ângulo do vetor do cruzamento, indicando a direção do passe.
- **distance_start_from_goal**: Distância do jogador que cruzou até o gol.
- **distance_from_end_line**: Distância da bola até a linha de fundo.

2.5. Features Extraídas a partir dos Dados de Tracking

Além dos dados de eventos, utilizamos os dados de *tracking* para capturar informações dinâmicas e espaciais. As principais *features* extraídas incluem:

- **attackers_in_box** e **defenders_in_box**: Número de jogadores atacantes e defensores dentro da grande área.
- **attackers_in_zone** e **defenders_in_zone**: Número de jogadores em uma zona tática específica próxima à área.
- **num_attackers_near_action_line** e **num_defenders_near_action_line**: Quantidade de jogadores próximos à linha de ação do cruzamento.
- **box_ratio**: Razão entre a ocupação ofensiva e defensiva dentro da grande área.
- **zone_ratio**: Razão semelhante, calculada para a zona tática específica.

3. Entropia como Medida de Desorganização Defensiva

3.1. Motivação

No nosso trabalho, usamos a **entropia**, um conceito da teoria da informação, para medir a incerteza de um sistema. No futebol, isso nos permite quantificar a desorganização posicional de uma equipe, principalmente a defesa. A nossa ideia é simples: uma defesa mais desorganizada tem maior incerteza na ocupação de espaços importantes, o que cria mais oportunidades para o ataque.

Diferente de métricas mais tradicionais, como a simples contagem de jogadores, a entropia foca na distribuição e dispersão deles. Isso nos permite identificar nuances que uma análise mais superficial não capturaria, oferecendo um indicador mais completo da vulnerabilidade defensiva.

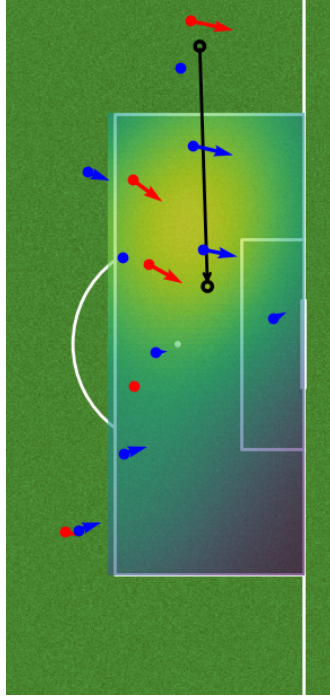


Figure 4. Entropia Geral

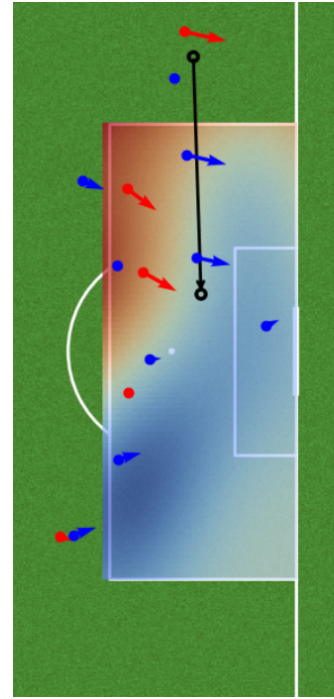


Figure 5. Diferença entre entropia ataque x defesa

Figure 6. Exemplos visuais de entropia no campo.

3.2. Cálculo da Entropia Espacial

Para calcular a entropia, dividimos o campo em uma grade de 120×80 células, focando na área ofensiva. Em cada *frame* do *tracking*, projetamos a posição dos jogadores para o segundo seguinte com base em sua velocidade. Para suavizar essa projeção e criar uma estimativa de densidade de presença, usamos *Kernel Density Estimation* (KDE) com uma função gaussiana.

A partir dessa densidade, calculamos a probabilidade de ocupação para cada célula da grade, separadamente para o ataque e para a defesa. A entropia posicional (H) de uma equipe é então calculada pela fórmula de Shannon:

$$H = - \sum_{i=1}^n p_i \log(p_i)$$

onde p_i é a probabilidade de presença da equipe na célula i .

3.3. Variáveis (Features) Derivadas da Entropia

A partir desses cálculos, geramos várias *features* para o modelo, que foram pensadas para capturar diferentes aspectos da entropia posicional. As principais categorias incluem:

- **Entropia Geral e por Equipe:** Calculamos a entropia média e a soma da entropia para o ataque, a defesa e o campo todo.
- **Entropia em Zonas Específicas:** Avaliamos a entropia em áreas críticas para cruzamentos, como as zonas ao redor da bola, no "primeiro pau", no "segundo pau" e na área central da grande área.

- **Diferença de Entropia:** Para entender o desequilíbrio entre as equipes, calculamos a diferença de entropia entre atacantes e defensores, tanto no campo como um todo quanto nas zonas específicas.
- **Gradiente de Entropia:** Medimos a variação da entropia em direção ao gol, o que nos ajuda a entender como a desorganização defensiva se comporta ao longo da progressão da jogada.

3.4. Integração ao Modelo

Incorporamos essas *features* de entropia aos nossos modelos. No *xCross*, elas ajudam a prever a probabilidade de um cruzamento resultar em uma finalização. Já no *xCrossOT*, usamos as *features* de entropia mais localizadas (*_in_zone*) para refinar a estimativa, focando na precisão da finalização. Acreditamos que a entropia é uma ferramenta poderosa para medir a vulnerabilidade momentânea da defesa, o que melhora a capacidade preditiva dos nossos modelos.

4. Resultados

Para avaliar o desempenho dos modelos, usamos um conjunto de 2.448 cruzamentos, sendo 562 deles considerados bem-sucedidos. Dividimos os dados em 80% para treino e 20% para teste, de forma estratificada. O treinamento foi feito com o *XGBoost*, ajustando os hiperparâmetros por validação cruzada e calibrando as previsões com o *Calibrated-ClassifierCV*.

Apresentamos os resultados de duas versões: o **xCross**, que avalia o potencial de qualquer cruzamento com base apenas na sua posição inicial, e o **xCrossOT**, que refina essa análise ao incluir a informação de onde a bola realmente foi, usando *features* locais adicionais.

4.1. Desempenho do Modelo xCross

O modelo **xCross** obteve uma área sob a curva (AUC) de 0,62 na curva ROC (Figura 7), um resultado razoável para a tarefa de distinguir entre cruzamentos que dão certo e os que não dão. A curva de calibração mostra que o modelo mantém boas previsões até uma probabilidade de 0,7 (Figura 4.1), e o histograma confirma uma boa distribuição das probabilidades.

A análise de importância das variáveis via SHAP (Figura 10) revelou um resultado importante: as *features* de **entropia espacial** foram cruciais. Variáveis como *entropy_diff_around* e *entropy_defense_around* tiveram grande destaque, o que valida nossa hipótese de que a desorganização defensiva é um fator-chave para o sucesso de um cruzamento.

Analisando o ranking de jogadores (Tabela 1), o modelo identificou que o lateral Kieran Trippier, com uma média de 0,392, não só cruza com frequência, mas também com alta qualidade. A lista inclui principalmente laterais e pontas de destaque, como Pervis Estupiñán e Gabriel Martinelli, confirmando que essas posições são as mais eficazes para a criação de jogadas a partir das laterais.

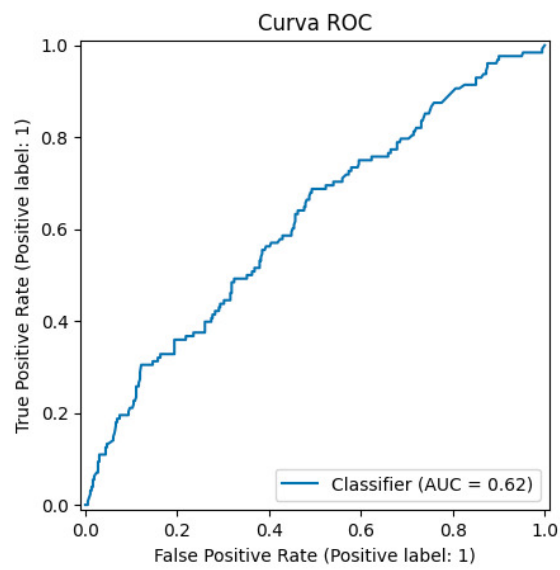


Figure 7. Curva ROC do modelo xCross.

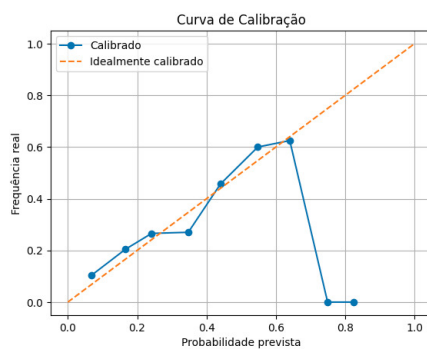


Figure 8. Curva de calibração do modelo xCross.

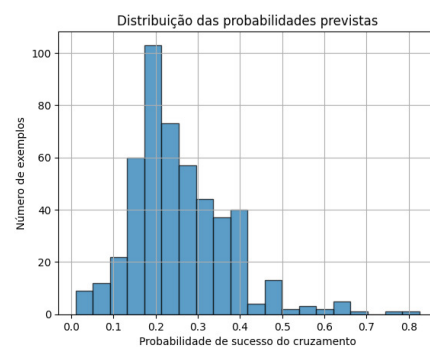


Figure 9. Histograma das probabilidades previstas pelo modelo xCross.

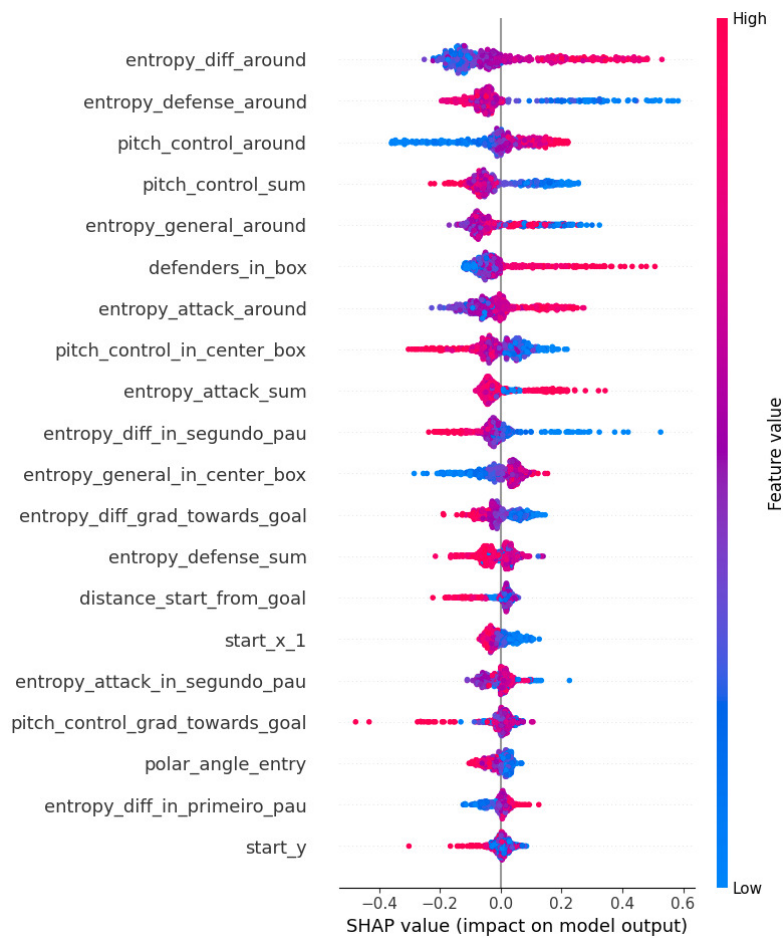


Figure 10. Importância das variáveis no modelo *xCross*, segundo SHAP.

Table 1. Top 10 jogadores segundo o modelo *xCross*

Jogador	Cruzamentos	Média <i>xCross</i>	Soma <i>xCross</i>
Kieran Trippier	25	0,392	9,789
Jack Harrison	24	0,319	7,646
Mads R. Rasmussen	20	0,315	6,302
Pervis Estupiñán	25	0,309	7,725
Malo Gusto	26	0,307	7,990
Diogo Dalot	27	0,303	8,176
Timothy Castagne	20	0,294	5,882
Gabriel Martinelli	26	0,290	7,546
Dejan Kulusevski	22	0,288	6,339
Dwight McNeil	32	0,286	9,138

4.2. Desempenho do Modelo *xCrossOT*

O modelo *xCrossOT* mostrou um desempenho superior, com uma AUC de 0,74 (Figura 11). Ele foi mais eficaz em prever se um cruzamento seria bem-sucedido, pois

teve acesso à informação da posição final da bola. A curva de calibração (Figura 4.2) confirma que as previsões são bem calibradas, e o histograma mostra uma distribuição mais equilibrada das probabilidades.

Neste modelo, as variáveis mais importantes (Figura 14) foram as coordenadas finais (`end_x` e `end_y`), o que faz sentido, já que o modelo leva em conta o destino da bola. No entanto, as *features* de entropia local, como `entropy_general_in_zone`, também tiveram um peso significativo, reforçando que a desorganização defensiva no ponto de chegada do cruzamento é fundamental.

A Tabela 2 apresenta os jogadores com melhor performance neste modelo. Per-vis Estupiñán e Kieran Trippier se destacam novamente, indicando que a qualidade dos seus cruzamentos não se resume apenas a uma boa posição inicial, mas também a uma execução precisa que posiciona a bola em uma área perigosa.

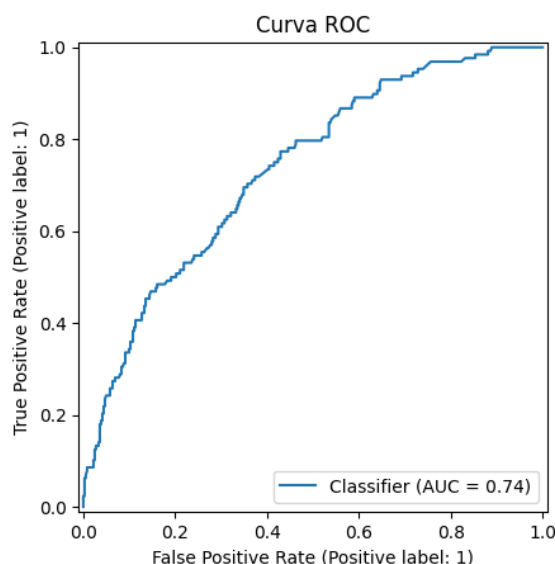


Figure 11. Curva ROC do modelo *xCrossOT*

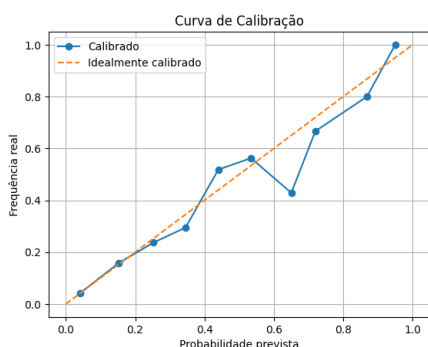


Figure 12. Curva de calibração do modelo *xCrossOT*

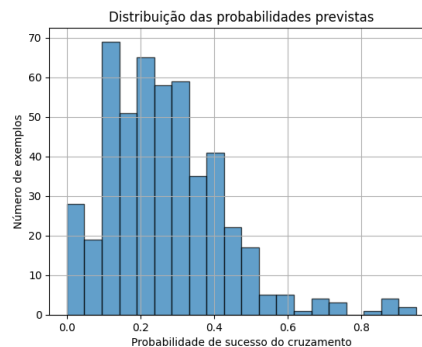


Figure 13. Histograma das probabilidades previstas pelo modelo

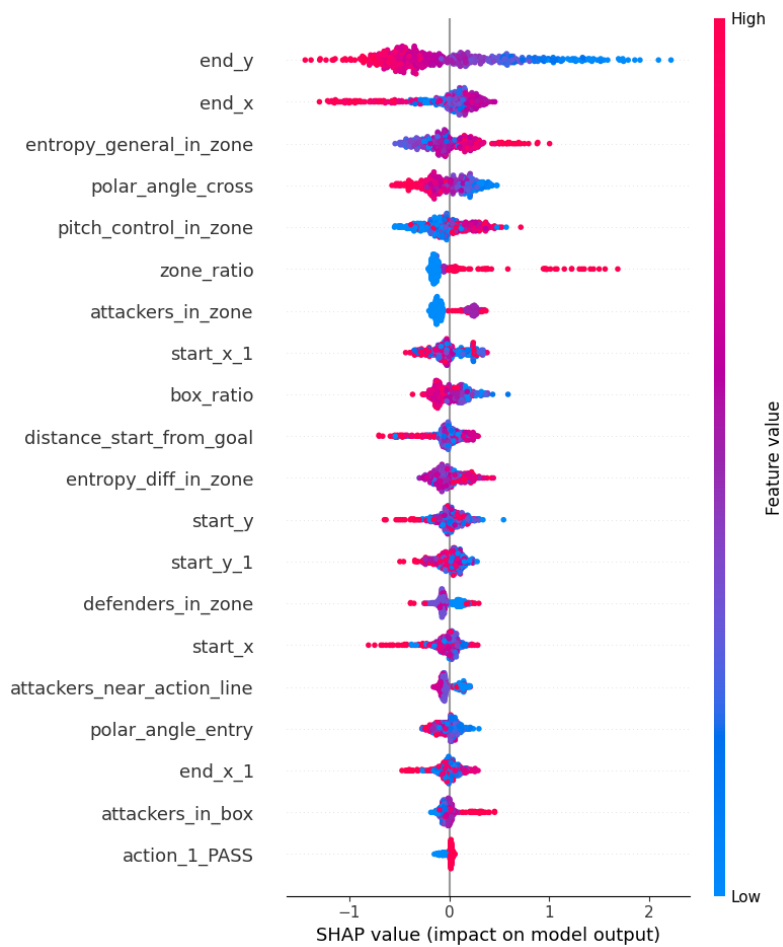


Figure 14. Importância das variáveis no modelo xCrossOT segundo SHAP

Table 2. Top 10 jogadores segundo o modelo xCrossOT

Jogador	Nº Cruzamentos	Valor Médio	Valor Total
Pervis Estupiñán	25	0,3625	9,06
Kieran Trippier	25	0,3477	8,69
Mohammed Kudus	33	0,3337	11,01
Malo Gusto	26	0,3246	8,44
Simon Adingra	20	0,3242	6,48
Gabriel Martinelli	26	0,3234	8,41
Jack Harrison	24	0,3227	7,74
Vladimir Coufal	38	0,3155	11,99
Kyle Walker	22	0,3073	6,76
Pedro Neto	39	0,3016	11,76

4.3. Análise Comparativa: xCross vs. xCrossOT

A Figura 15 mostra a relação entre as duas métricas. O eixo X representa o **valor médio de cruzamento** (*mean_xCross*), que mede o potencial da jogada no início, enquanto o eixo Y representa o **valor médio com informação de destino** (*mean_xCrossOT*).

Como esperado, a correlação entre as duas métricas é positiva, o que sugere que jogadores que iniciam cruzamentos de posições favoráveis tendem a ser mais eficazes em direcioná-los para a área perigosa.

No entanto, a dispersão dos pontos no gráfico é a parte mais importante. Ela mostra que as duas métricas não são iguais, e essa diferença é o ponto-chave do nosso trabalho.

- **Jogadores de alta eficácia** (canto superior direito do gráfico): Atletas como Estupinã e Trippier, por exemplo, se destacam nas duas métricas. Isso sugere que eles não só escolhem o momento certo para cruzar, mas também executam a jogada com alta precisão, direcionando a bola para uma zona de perigo.
- **Jogadores de eficácia variável**: Outros jogadores podem ter um *xCross* razoavelmente alto, mas um *xCrossOT* mais baixo. Isso indica que, embora eles tenham potencial para criar perigo, a execução do passe em si pode não ser tão precisa, posicionando a bola em uma área de menor risco.

A existência dessas variações confirma a importância do modelo **xCrossOT**. Ele nos dá uma visão mais refinada, capturando nuances da qualidade da execução do cruzamento que o modelo *xCross*, por si só, não conseguiria. A diferença entre as duas métricas para um mesmo jogador pode ser um bom indicador da sua precisão técnica e capacidade de finalizar a jogada.

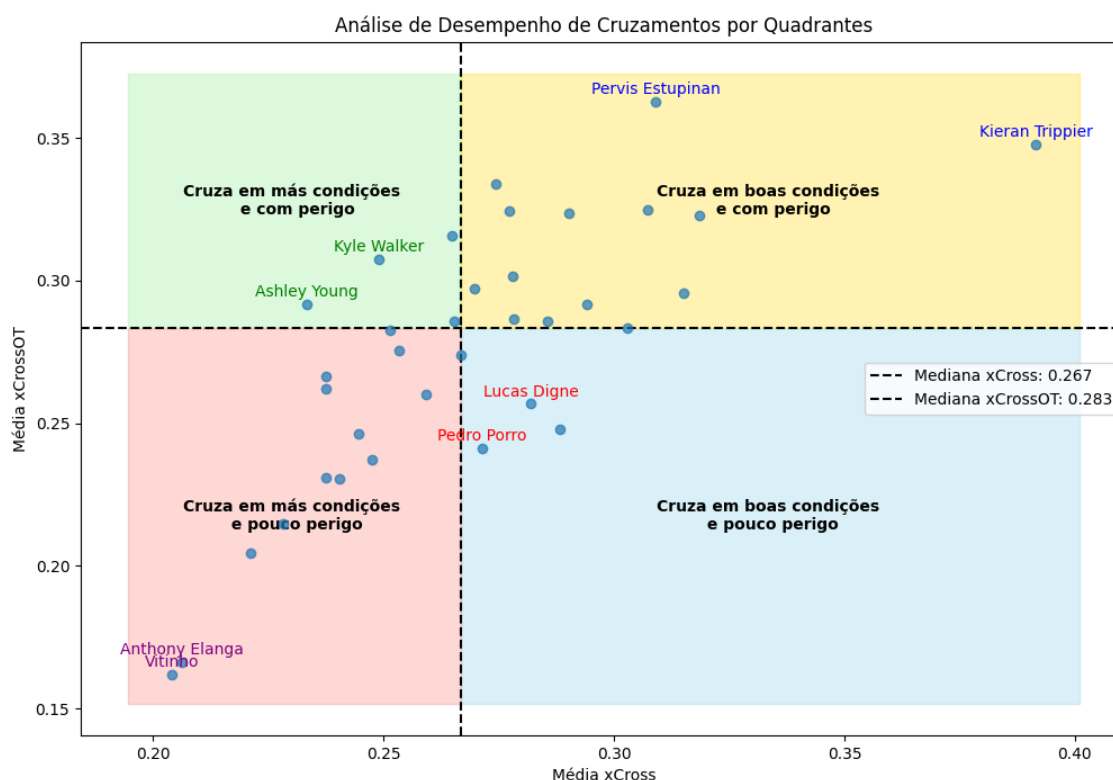


Figure 15. Relação entre a Média de Cruzamentos e a Média de Cruzamentos no Alvo.

5. Conclusão

Nosso trabalho propôs o **xCross**, um novo modelo para avaliar a qualidade de cruzamentos no futebol, utilizando a **entropia espacial** como uma medida de desorganização defensiva. O objetivo era ir além das métricas tradicionais, que ignoram o complexo contexto posicional das jogadas.

Os resultados obtidos com os modelos **xCross** e **xCrossOT** validam nossa hipótese central. A análise de importância das variáveis (SHAP) confirmou que as métricas de entropia, que quantificam a desorganização da defesa, são cruciais para a capacidade preditiva do modelo. Isso demonstra que a entropia é, de fato, um bom indicador da vulnerabilidade defensiva no momento de um cruzamento.

A comparação entre as duas versões do modelo foi outro ponto-chave. O **xCross**, que avalia o potencial da jogada com base apenas na posição inicial, e o **xCrossOT**, que incorpora a informação do destino da bola, mostraram uma correlação positiva. No entanto, a análise de dispersão revelou que as métricas não são idênticas. Essa diferença é o que torna o **xCrossOT** particularmente valioso, pois ele captura a qualidade da execução técnica de um jogador, algo que o primeiro modelo não consegue. Essa nuance nos permite diferenciar entre jogadores que têm bom posicionamento para cruzar e aqueles que, além de se posicionarem bem, também executam o passe com precisão para uma área perigosa.

Em resumo, nosso estudo oferece uma contribuição importante para a análise espacial no futebol, mostrando que métricas baseadas em entropia podem ser usadas de forma eficaz para avaliar o impacto tático e a técnica de jogadores. Para trabalhos futuros, seria interessante explorar o uso de redes neurais para capturar padrões espaciais ainda mais complexos ou aplicar a metodologia em outras jogadas, como passes longos e lançamentos, para expandir a compreensão sobre o valor posicional no futebol.

Referências

- Decroos, T., Bransen, L., Van Haaren, J., & Davis, J. (2019). Actions speak louder than goals: Valuing player actions in soccer. In Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (pp. 1851–1861).
- Fernández, J., Bornn, L., & Cervone, D. (2019). Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer. Sloan Sports Analytics Conference.
- Vecer, Jan, Crossing in Soccer has a Strong Negative Impact on Scoring: Evidence from the English Premier League the German Bundesliga and the World Cup 2014 (September 30, 2014).

Analysis of Off-Ball Movement in a Counterattack Using Pitch Control and Deep Learning Models

Bernardo Venancio¹, Bruno Silva¹, Luca Kosfeld¹, Lucas Albuquerque¹

¹ Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)

Belo Horizonte – MG – Brazil

bernardovenancio@dcc.ufmg.br, bruno.soares@dcc.ufmg.br,
lucakosfeld@ufmg.br, lucasalbuquerque@dcc.ufmg.br

Abstract. Counterattacks are one of the most effective offensive strategies in football, but their success is largely based on coordinated movement off the ball. This study proposes a computational approach to analyze player behavior off the ball during counterattacks, employing Field Control metrics, deep learning models, and sequence analysis to understand the spatial and temporal dynamics that contribute to successful plays. A pipeline involving data filtering, Autoencoder-based dimensionality reduction, and Long Short-Term Memory (LSTM) models is developed to predict the success of counterattacks. The results demonstrate the good ability of the model to represent movements and identify key spatial patterns of player positioning, providing valuable information to coaches and analysts. Future work includes simulating optimal player repositioning through ghosting techniques.

1. Introduction

Counterattacks represent one of the most significant avenues for goal scoring in football, but their efficacy is intrinsically linked to the intricate and coordinated movement of players without the ball. Understanding and optimizing these off-ball dynamics is crucial to gaining a competitive advantage. This study is motivated by the need to develop a trust model that can identify and predict successful counterattack patterns based on off-ball movement. The analysis of tactical plays from raw positional data presents substantial computational challenges due to the high dimensionality and inherent complexity of football movements. Traditional analytical methods often struggle to capture the nuanced temporal and spatial relationships that define successful counterattacks. Therefore, this research proposes an efficient approach that integrates advanced machine learning techniques to overcome these limitations and provide valuable insights for tactical planning.

2. Methods

2.1. Defining a Counterattack

The initial stage of this study consisted of establishing strict operational criteria to objectively identify counterattacks within the available match data. The adopted conditions are described below:

- The sequence must begin with a ball recovery in the defensive half of the field: Only sequences initiated by regaining possession within the team's own defensive zone are considered. This criterion ensures that the analyzed transitions originate from defensive actions.

- The transition to the offensive half must be rapid, direct, and significant in territorial gain: Following ball recovery, the ball must cross into the offensive half within a maximum of 10 seconds and the play must result in a forward advance of at least 10 meters. This ensures that only sequences with clear offensive intent are classified as counterattacks.
- Set-piece situations were excluded, only possessions initiated while the ball is in motion at a velocity exceeding a predetermined minimum threshold are considered.
- The final event of the possession sequence must occur shortly after the ball crosses the midfield line, within a predefined elliptical region: This spatial filter ensures that the analyzed counterattacks reflect sequences where the attacking team retained control and offensive structure after entering the opponent's territory.
- The possession sequence must culminate with the ball reaching the opponent's penalty area to be considered successful: Counterattacks that successfully penetrate the opponent's penalty area are recognized as effective.

2.2. Data Pipeline

The data processing pipeline consisted of multiple structured stages. An initial filtering process was applied to the raw dataset to isolate only those sequences that met the strict operational criteria for counterattacks. Following this, the tracking data underwent a structured adaptation process, where positional data for all players and the ball were augmented with derived velocity information. To ensure temporal consistency, the tracking dataset was synchronized with the corresponding event data. Finally, a cleaning and adjustment phase was conducted to correct erroneous speed measurements and standardize attacking directions, ensuring all plays progressed from left to right on the field for analytical consistency.

2.3. Tactical Representation: Pitch Control and Additional Metrics

To adequately quantify the spatial dynamics, this study adopts the **Pitch Control** model, a consolidated metric that estimates the probability of each team controlling different regions of the field. This estimation considers player positions, movement capabilities, and reaction time. In addition to Pitch Control, two complementary metrics enrich the analysis. The **Expected Threat (xT)** model quantifies the offensive potential based on the ball's location, while **Hot Spot** maps reveal regions of high player density and space generation. Together, these metrics provide a comprehensive representation of space control and offensive threat.

2.4. Dimensionality Reduction with Autoencoder

The positional and tactical data are inherently high-dimensional. To overcome this, an **Autoencoder** neural network is employed to compress the data into a lower-dimensional latent space, extracting the most relevant information. The Autoencoder consists of an encoder, which transforms the input data into a compact latent vector, and a decoder, which reconstructs the data during training. This latent vector serves as a condensed description of each moment, providing an efficient input for the subsequent temporal modeling.

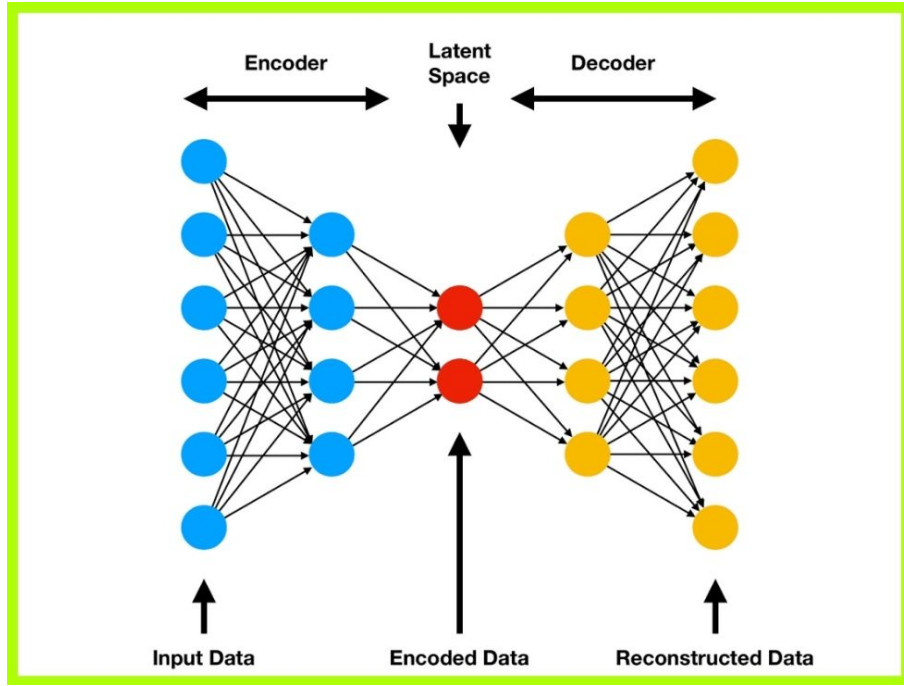


Figure 1. Architecture of the Autoencoder used for dimensionality reduction, highlighting the encoder, latent space, and decoder.

2.5. Temporal Modeling with LSTM

To model the evolution of player positioning over time, this study employs a **Long Short-Term Memory (LSTM)** network, a type of recurrent neural network designed to process sequential data. The key to the LSTM's effectiveness lies in its internal gating mechanisms. Each LSTM unit contains three gates: the *input gate*, which determines how much new information is incorporated; the *forget gate*, which selectively discards irrelevant information from the previous state; and the *output gate*, which controls how the current state influences the output. This structure allows the LSTM to capture complex temporal patterns. Once the latent vectors are generated by the Autoencoder, they are sequentially passed as inputs to a stacked LSTM network, which predicts the probability of success for the counterattack.

2.6. Model Training and Refinement

Several challenges were encountered during model development, including overfitting and class imbalance. To enhance model generalization, regularization techniques were applied. **Dropout** was used to randomly deactivate neurons during training, promoting robustness. Additionally, **L2 Regularization** was incorporated to penalize excessive weight magnitudes, preventing the model from overfitting to training noise.

3. Results

3.1. Model Performance: ROC Curve Analysis

The effectiveness of the model in distinguishing between successful and unsuccessful counterattacks was evaluated through the Receiver Operating Characteristic (ROC) curve. The ROC curve represents the relationship between the True Positive Rate and the False

Positive Rate, providing a global measure of the model's discriminative ability. As shown in Figure 2, the model achieved an **Area Under the Curve (AUC) of 0.85**. This value indicates a high capacity to correctly classify counterattacks, meaning there is an 85

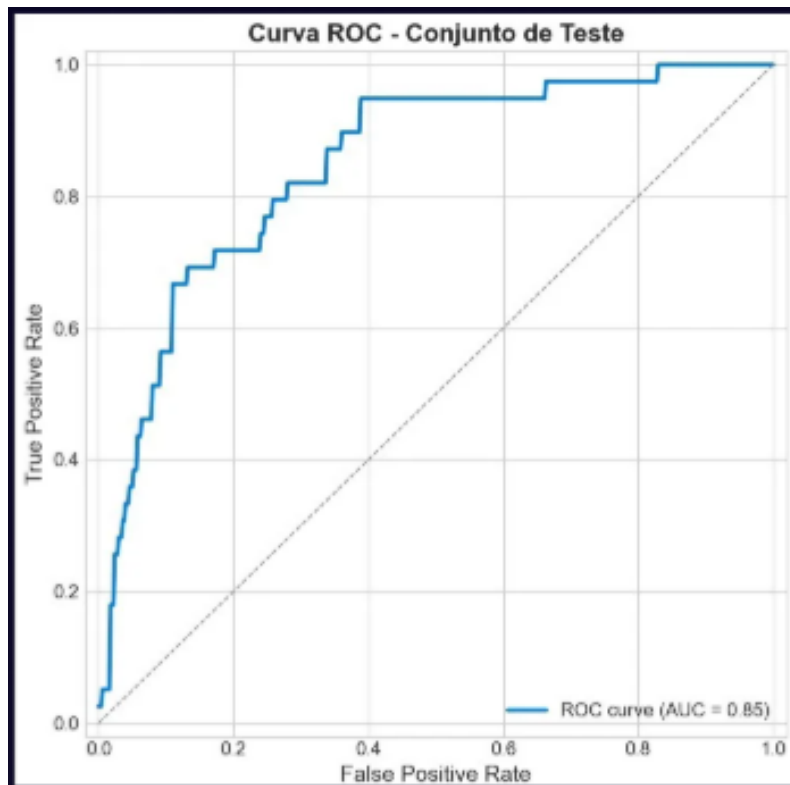


Figure 2. ROC curve for the test set, with an AUC of 0.85, indicating strong model performance in distinguishing successful counterattacks.

3.2. Model Training Behavior and Tactical Interpretability

The learning curves obtained during training provided important insights. As shown in Figure 3, although the model achieved rapid convergence, the validation curves exhibited instability, indicating the complexity of the task and a risk of overfitting. The adopted regularization strategies were crucial in stabilizing the training process and improving reliability. Beyond predictive performance, the model offered interpretable tactical insights. By analyzing the latent space and sequential outputs, it was possible to identify key player positioning patterns associated with successful transitions and quantify how space control evolved.

4. Ghosting: Simulating Optimal Player Positioning

To complement the analysis, this study incorporates **Ghosting**, a simulation-based approach designed to explore alternative player positioning. The core idea is to virtually reposition players and assess how these adjustments would affect the team's tactical structure. Ghosting was applied to simulate optimal positioning for off-ball players, estimating how adjustments could maximize space control or increase the probability of success. This technique provides valuable support for coaching staff, enabling data-driven evaluations of tactical decisions.

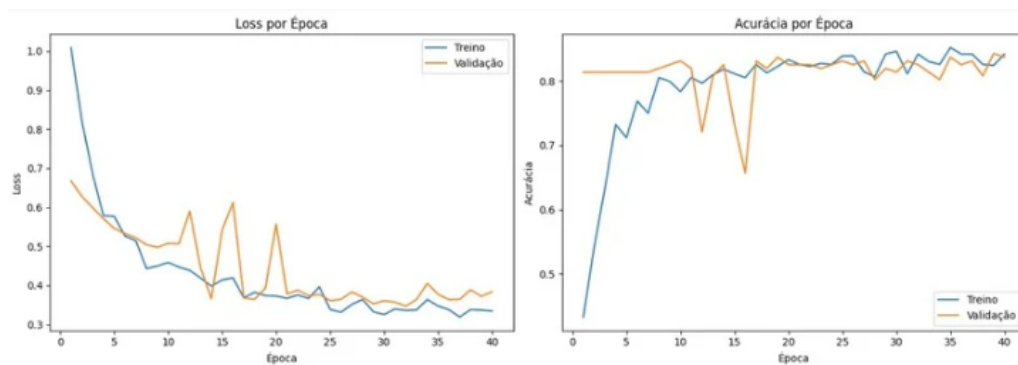


Figure 3. Learning curves of loss and accuracy for training and validation sets, illustrating rapid convergence, validation instability, and the need for regularization techniques.

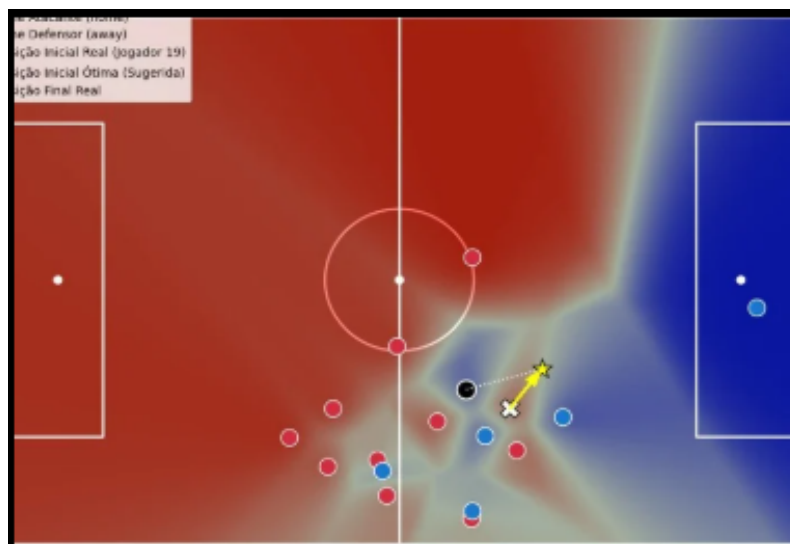


Figure 4. Example of Ghosting simulation suggesting optimal player repositioning to enhance space control during a counterattack.

5. Discussion

The integration of spatial control metrics and deep learning offers a novel framework for understanding off-ball dynamics in football. The Autoencoder-LSTM architecture successfully condenses complex spatial-temporal information, enabling both prediction and interpretation. However, several limitations remain:

- The reliance on predefined counterattack criteria may overlook unconventional transition patterns;
- The binary success label simplifies the inherently complex outcome spectrum of offensive plays;
- Interpretability of deep learning models, while improved, still requires domain expert validation.

Future work should explore:

- Incorporation of contextual variables (e.g., player fatigue, opponent formations);
- Simulation of alternative tactical scenarios via ghosting to optimize player positioning;

- Extension of the framework to other phases of play beyond counterattacks.

6. Conclusion

This study presents a computational framework for analyzing off-ball movement and predicting counterattack success in football. The results highlight the critical role of coordinated player positioning in maximizing offensive potential. The developed model not only predicts success with high accuracy but also provides actionable insights for coaches and analysts. By advancing the understanding of off-ball dynamics, this approach contributes to more data-driven and effective offensive strategies in modern football.

References

- [1] Spearman, W. (2018). “Beyond Expected Goals.” MIT Sloan Sports Analytics Conference.
- [2] Fernández, J., & Bornn, L. (2018). “Wide Open Spaces: A Statistical Technique for Measuring Space Creation in Professional Soccer.” MIT Sloan Sports Analytics Conference.
- [3] Kingma, D. P., & Welling, M. (2013). “Auto-Encoding Variational Bayes.” arXiv preprint arXiv:1312.6114.
- [4] Hochreiter, S., & Schmidhuber, J. (1997). “Long Short-Term Memory.” *Neural Computation*, 9(8), 1735–1780.
- [5] FIFA. (2022). FIFA Football Data Analysis Guidelines.
- [6] Decroos, T., Bransen, L., Van Haaren, J., & Davis, J. (2019). “Actions Speak Louder Than Goals: Valuing Player Actions in Soccer.” *Data Mining and Knowledge Discovery*, 33(5), 1350–1375.
- [7] Bekkers, J., & Sahasrabudhe, A. (2023). “A Graph Neural Network deep-dive into successful counterattacks.” MIT Sloan Sports Analytics Conference.



IV Football Analytics Modelling & Experience

03 de setembro de 2025

Belo Horizonte, Brasil

Compêndio de Artigos

Realização

SALab - Sports Analytics Lab

Organização

Departamento de Ciência da Computação - DCC - UFMG