Programa Institucional de Bolsas de Iniciação em Desenvolvimento Tecnológico e Inovação (PIBITI)

Insper

Engenharia Mecatrônica

Análise e Modelagem Estatística de Métricas de Desempenho no Futebol: Seleção de Atributos-Chave por Posição

João Vitor Oliveira Assunção

Orientador: Prof. Dr. André Luiz Maciel Santana

São Paulo

2025

Programa Institucional de Bolsas de Iniciação em Desenvolvimento Tecnológico e Inovação (PIBITI)

João Vitor Oliveira Assunção
Engenharia Mecatrônica

Análise e Modelagem Estatística de Métricas de Desempenho no Futebol: Seleção de Atributos-Chave por Posição

Relatório final apresentado ao Programa Institucional de Bolsas de Iniciação em Desenvolvimento Tecnológico e Inovação (PIBITI)

Orientador: Prof. Dr. André Luiz Maciel Santana

São Paulo

2025

AGRADECIMENTOS

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro, concedido por meio da bolsa no Programa Institucional de Bolsas de Iniciação em Desenvolvimento Tecnológico e Inovação (PIBITI), que viabilizou a realização deste estudo.

Resumo

Este relatório apresenta o processo de modelagem estatística do desempenho de jogadores de futebol, com o objetivo de identificar e quantificar as principais métricas que influenciam nas classificações numéricas atribuídas aos desempenhos de atletas em partidas oficiais, conhecidas como *ratings*. Embora plataformas como o SofaScore sejam consolidadas na atribuição de notas ao desempenho de jogadores (*ratings*), não há transparência no processo de atribuição dos *ratings*. Esta pesquisa se propõe a identificar os atributos estatísticos mais relevantes para descrever o desempenho de atletas por posição, construindo modelos capazes de transformar métricas individuais em um índice quantitativo de avaliação (*rating*). Para alcançar esse objetivo, foram extraídos dados da plataforma SofaScore, seguindo as etapas de coleta automática via API, tratamento dos dados, estudo das métricas disponíveis, exclusão de métricas redundantes e agrupamento dos dados de acordo com a função tática exercida pelo atleta. A metodologia consistiu em estruturar o conjunto de dados em quatro segmentações, sendo goleiros, defensores, meio-campistas e atacantes, aplicar o método de regressão linear múltipla e analisar os resultados a fim de identificar as métricas de maior relevância por posição e os seus respectivos coeficientes de impacto. O modelo foi avaliado considerando o desempenho em prever os *rating* de dois atletas por posição, garantindo a diversidade de estilos de jogo, características físicas, ligas de atuação e origem dos atletas, com análise de métricas como o R^2, que mede a capacidade explicativa do modelo, e da raiz do erro quadrático médio (RMSE). Os resultados indicam que métricas de natureza ofensiva, como gols, assistências, chutes no gol, finalizações na trave e cobranças de pênaltis convertidas impactam positivamente no *rating*, enquanto métricas de natureza defensiva não contribuem com a mesma relevância. Os resultados reforçam a importância de contextualizar os dados com base no enredo da partida e na função desempenhada pelo atleta.

Palavras-chave: Análise de desempenho; Futebol; Ciência de dados; Regressão linear; Métricas esportivas.

Abstract

This report presents the statistical modeling process for soccer player performance, aiming to identify and quantify the key metrics that influence the numerical classifications assigned to athletes' performances in official matches, known as ratings. Although platforms like SofaScore are well-established in assigning ratings to player performance (ratings), the rating process lacks transparency. This research aims to identify the most relevant statistical attributes for describing athlete performance by position, building models capable of transforming individual metrics into a quantitative evaluation index (rating). To achieve this goal, data was extracted from the SofaScore platform, following the steps of automatic collection via API, data processing, analysis of available metrics, exclusion of redundant metrics, and grouping of data according to the tactical role played by the athlete. The methodology consisted of structuring the dataset into four segments: goalkeepers, defenders, midfielders, and attackers. It applied multiple linear regression and analyzed the results to identify the most relevant metrics by position and their respective impact coefficients. The model was evaluated based on its performance in predicting the ratings of two players per position, ensuring diversity in playing styles, physical characteristics, leagues, and player origins. It also analyzed metrics such as R^2, which measures the model's explanatory power, and the root-mean-square error (RMSE). The results indicate that offensive metrics such as goals, assists, shots on target, shots on goal, and penalty kicks converted have a positive impact on ratings, while defensive metrics do not contribute as significantly. The results reinforce the importance of contextualizing the data based on the match's storyline and the player's role.

Keywords: Performance analysis; Soccer; Data science; Linear regression; Sports metrics.

Sumário

1. INTRODUÇÃO 6
2. REVISÃO TEÓRICA 7
2.1 Métricas e Avaliação de Desempenho no Futebol 7
2.2 Regressão Linear Múltipla na modelagem esportiva 9
3. METODOLOGIA 12
3.1 Extração dos Dados 12
3.2 Tratamento dos Dados 14
3.3 Modelagem Estatística 15
3.4 Avaliação do Modelo 22
4. RESULTADOS 23
4.1 Capacidade Preditiva 23
4.2 Atributos com Maior Impacto no Rating 26
5. CASOS DE USO NO FUTEBOL 30
6. CONSIDERAÇÕES FINAIS 33
Referências 34

1. INTRODUÇÃO

No contexto do futebol contemporâneo, observa-se uma crescente profissionalização dos clubes em relação ao uso de dados como instrumento de apoio à tomada de decisões estratégicas. Conforme destacado por Nakamura e Cerqueira (2021), um fator determinante para a formação de equipes sólidas e competitivas reside nos investimentos em infraestrutura tecnológica, tanto por meio da aquisição de equipamentos avançados para a coleta eficiente de dados quanto pela constituição de equipes qualificadas, capazes de realizar a extração, análise e interpretação das informações, promovendo decisões fundamentadas em evidências.

No cenário nacional, observa-se uma adoção crescente de plataformas como o SofaScore, que oferecem métricas detalhadas de desempenho, incluindo gols, assistências, passes, ações defensivas e índices de precisão. Esses dados reduzem o viés subjetivo na avaliação de atletas, favorecendo análises quantitativas e qualitativas de performance. Trabalhos como os de Ortolan e Silva (2020), Assunção (2024) e Shamah (2021) reforçam a importância de estudos baseados em métricas objetivas para compreender padrões de jogo, eficiência individual e coletiva, além de apoiar decisões táticas.

No entanto, um desafio recorrente está em considerar as particularidades de cada posição em campo. Goleiros, defensores, meio-campistas e atacantes desempenham funções distintas, demandando métricas específicas para uma avaliação justa e precisa. Estudos como o de Ball et al. (2025) indicam que ações ofensivas, como gols e assistências, tendem a ser supervalorizadas em relação a métricas defensivas, o que pode distorcer a percepção real do desempenho.

Diante desse contexto, este trabalho tem como objetivo identificar os atributos estatísticos mais relevantes para descrever o desempenho de atletas por posição, construindo modelos capazes de transformar métricas individuais em um índice quantitativo de avaliação (*rating*). A abordagem adotada busca não apenas prever notas, mas também explicar de forma transparente a contribuição de cada métrica, permitindo insights úteis para analistas, clubes e pesquisadores.

2. REVISÃO TEÓRICA

Nesta seção, apresenta-se os principais conceitos e estudos relacionados à análise esportiva no futebol, destacando o uso da ciência de dados e técnicas estatísticas para a compreensão e previsão do desempenho dos atletas.

2.1 Métricas e Avaliação de Desempenho no Futebol

A crescente necessidade de mensurar o desempenho dos atletas tem impulsionado o interesse por análises quantitativas no esporte, especialmente na avaliação estatística de partidas de futebol profissional. Conforme destacado por Ali (2011), o desempenho técnico de um jogador, que pode ser mensurado por fundamentos como passes, finalizações, dribles e desarmes, é um dos principais determinantes da performance da equipe, sendo, portanto, objeto de interesse em avaliações objetivas. Nesse contexto, plataformas de análise automatizada, como SofaScore, WhoScored e FotMob, vêm ganhando destaque ao fornecer métricas detalhadas e notas quantitativas baseadas em algoritmos proprietários, contribuindo para a redução do viés subjetivo nas avaliações (Miranda, 2021). Entre essas plataformas, o SofaScore destaca-se como a principal referência no cenário atual, especialmente em contextos profissionais e acadêmicos, sendo amplamente utilizado por clubes, analistas de desempenho, veículos da mídia esportiva e pesquisadores (Ball et al., 2025). Sua popularidade e profundidade estatística tornaram-no uma fonte consolidada para estudos baseados em métricas objetivas no futebol. Além disso, metodologias como o modelo plus-minus, métrica de esportes coletivos que mostra a diferença no placar quando um jogador está em quadra, adaptado ao futebol, proposto por Kharrat, López Peña e McHale (2017), destacam a importância de métricas que avaliam a contribuição real do jogador para os resultados da equipe, reforçando a necessidade de modelos mais interpretáveis e comparáveis.

Entretanto, a utilização de métricas agregadas e generalistas pode gerar distorções na avaliação do desempenho individual, sobretudo quando se desconsidera a função tática exercida pelo atleta. Ball et al. (2025), ao comparar diferentes plataformas de

rating automático, observaram que métricas ofensivas como gols e assistências são, com frequência, supervalorizadas em relação a ações defensivas, como interceptações e desarmes. Tal assimetria tende a favorecer jogadores de ataque em detrimento de defensores e goleiros, comprometendo a fidedignidade das notas atribuídas, sendo importante equilibrar o peso de cada métrica de acordo com a função tática que o jogador desempenha.

Nesse sentido, Nakamura e Cerqueira (2021) defendem que uma estrutura analítica eficaz depende não apenas da coleta automatizada de dados, mas também da capacidade analítica de interpretar tais dados conforme as atribuições específicas de cada posição em campo. O investimento em infraestrutura tecnológica, associado à formação de equipes técnicas capacitadas, é apontado pelos autores como um diferencial competitivo para clubes que buscam decisões estratégicas fundamentadas em evidências para a formação de equipes sólidas e competitivas. Paralelamente, estudos como o de Wunderlich e Memmert (2018) propõem abordagens alternativas para a avaliação de desempenho, explorando o uso de grafos de confrontos entre equipes para prever resultados futuros com base em ratings anteriores. Embora centrado em análises coletivas, o estudo evidencia a utilidade de índices sintéticos de desempenho, reforçando a importância da construção de métricas interpretáveis.

No contexto brasileiro, trabalhos como o de Assunção et al. (2024) exploram o uso de ciência de dados e visão computacional para a análise do desempenho físico de atletas, destacando a relevância de variáveis como distância percorrida, aceleração e intensidade dos deslocamentos. Embora tais variáveis não estejam amplamente disponíveis em plataformas públicas, sua menção reforça a complexidade envolvida na avaliação objetiva do desempenho esportivo.

Por fim, vale mencionar a advertência feita por Afonso et al. (2024) quanto ao uso indiscriminado de modelos estatísticos, como a regressão linear múltipla, em contextos puramente preditivos. Os autores ressaltam que, quando o objetivo é a explicação de fenômenos, a ênfase deve recair sobre a interpretabilidade dos coeficientes, a análise das suposições do modelo e a identificação da influência relativa de cada variável independente sobre a variável resposta. Essa necessidade de interpretabilidade também é destacada por Kharrat, López Peña e McHale (2017), que defendem a adoção de modelos transparentes e comparáveis para que a

avaliação do impacto individual de cada atleta esteja ancorada em métricas compreensíveis e replicáveis.

Diante dessas pesquisas, observa-se uma lacuna na literatura quanto à construção de modelos explicativos que considerem as particularidades posicionais dos atletas e que permitam compreender, de forma transparente, quais atributos estatísticos mais influenciam a nota atribuída em plataformas de avaliação. Este trabalho propõe-se a suprir tal lacuna, por meio do desenvolvimento de modelos segmentados por posição e com foco na interpretabilidade dos resultados, respeitando os pressupostos estatísticos necessários para garantir validade e confiabilidade à análise.

2.2 Regressão Linear Múltipla na modelagem esportiva

A Regressão Linear (Rodrigues, 2012) é uma técnica estatística que permite modelar e analisar a relações entre uma ou mais variáveis dependentes, partindo de um conjunto de variáveis independentes, a fim de predizer valores das variáveis dependentes.

A diferenciação entre os tipos de variáveis é feita a partir dos seguintes critérios:

- **Variáveis Dependentes:** Variável que se tem o interesse de predizer ou explicar seus valores. No contexto desta pesquisa, trata-se do *rating* do jogador na partida.
- **Variáveis Independentes (ou Preditoras):** Variáveis utilizadas para predizer ou explicar as variáveis dependentes.

No futebol, podem ser métricas como número de gols, passes precisos, desarmes, entre outros. A regressão linear pode ser classificada em duas abordagens principais:

- **Regressão Linear Simples:** Utiliza-se apenas de uma variável independente para prever a variável dependente. Ideal para análises de relações lineares (Rodrigues, 2012).
- **Regressão Linear Múltipla:** Utiliza-se duas ou mais variáveis independentes para prever a variável dependente. Ideal para quando se tem interesse em

estudar a relação entre diversas variáveis e a variável de interesse (Rodrigues, 2012).

A regressão linear múltipla é dada pela equação 1:

$$y = \beta_0 + \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots + \beta_n \times x_n + \varepsilon \qquad (1)$$

Nesta equação, y representa a variável dependente, ou seja, o valor que se deseja prever com base nas demais variáveis do modelo. Os termos $x_1, x_2, \ldots, x_n$ correspondem às variáveis independentes (ou preditoras), que influenciam diretamente o comportamento de y. O coeficiente β_0 é o termo intercepto, que representa o valor esperado de y quando todas as variáveis independentes assumem valor zero.

Os parâmetros $\beta_1, \beta_2, \ldots, \beta_n$ são os coeficientes de regressão, que quantificam a contribuição individual de cada variável x_i na variação de y, mantendo-se constantes as demais variáveis do modelo. Já o termo ε (épsilon) representa o erro aleatório ou termo residual, o qual captura as variações não explicadas pelo modelo, assumindo-se que ele possui média zero e variância constante.

A estimativa dos coeficientes β é realizada por meio do método dos mínimos quadrados ordinários (MQO), que busca minimizar a soma dos quadrados dos resíduos, garantindo assim o melhor ajuste possível entre os valores observados e os valores preditos pelo modelo.

A determinação dos coeficientes β constitui o principal objetivo deste estudo, que se propõe a identificar, com base em técnicas de ciência de dados, as variáveis que mais influenciam o *rating* atribuído ao desempenho dos atletas. Para tanto, adota-se como premissa metodológica a segmentação dos dados de acordo com a posição em campo, permitindo uma análise mais refinada e contextualizada das métricas de desempenho relevantes para cada função específica.

No contexto esportivo, sua aplicação tem se mostrado útil tanto em tarefas preditivas quanto explicativas, permitindo estimar desempenhos futuros e identificar os fatores que mais influenciam os resultados observados. Sua ampla utilização decorre, em grande parte, da facilidade de interpretação dos coeficientes estimados, o que facilita a comunicação e interpretação dos resultados com profissionais não estatísticos, como técnicos e analistas esportivos.

Estudos recentes têm utilizado essa metodologia para avaliar o desempenho de jogadores com base em métricas extraídas de plataformas como o FIFA Dataset (Wang et al., 2023) e o SofaScore (Ball et al., 2025), aplicando-a com ajustes específicos, como transformações logarítmicas e *Box-Cox*, para lidar com problemas de não linearidade. Adicionalmente, tais abordagens frequentemente empregam estratégias de seleção de variáveis, como *forward selection*, a fim de eliminar variáveis irrelevantes ou redundantes, melhorar o ajuste do modelo e evitar sobre ajuste. Nesse sentido, práticas como a análise da multicolinearidade por meio da matriz de correlação também têm sido recomendadas para garantir a robustez do modelo e a estabilidade dos coeficientes.

Contudo, é importante destacar que, em aplicações esportivas, a multicolinearidade entre variáveis não deve ser considerada como uma justificativa para a exclusão automática. Certas métricas altamente correlacionadas, como gols e finalizações, são inerentemente interdependentes e carregam, em conjunto, importante significado tático e técnico. A remoção arbitrária de tais variáveis pode comprometer a representatividade do modelo, motivo pelo qual a análise de colinearidade deve considerar o contexto do domínio antes de qualquer decisão de exclusão. Ainda que a presença de correlação elevada possa inflacionar erros-padrão, o objetivo explicativo e a necessidade de interpretabilidade justificam, em muitos casos, sua permanência no modelo.

Nesse sentido, a presente pesquisa adota a regressão linear múltipla como ferramenta central para explicar, de forma transparente, quais métricas de desempenho contribuem para a nota atribuída ao jogador. A análise será conduzida por posição tática, com atenção especial à validação dos pressupostos do modelo, ao controle da multicolinearidade e à clareza na interpretação dos resultados.

3. METODOLOGIA

Nesta seção são apresentados os métodos empregados ao longo desses estudos, destacando os critérios e as técnicas utilizadas para a criação de um modelo capaz de auxiliar na previsibilidade e explicabilidade de uma nota de classificação de desempenho de atletas (rating) em uma partida de futebol. A condução desse estudo pode ser dividida em quatro etapas: (i) obtenção dos dados, (ii) tratamento dos dados, (iii) modelagem estatística e (iv) análise dos resultados.

A pesquisa foi realizada utilizando dados de desempenho de jogadores de futebol, obtidos a partir de um conjunto de partidas disputadas em competições oficiais ao longo da temporada passada. Os arquivos analisados estão no formato CSV, contendo as variáveis dependentes e independentes associadas às ações do jogador em campo.

Os modelos estatísticos foram desenvolvidos em Python (versão 3.9.7), utilizando as bibliotecas pandas para manipulação de dados, *numpy* para operações vetoriais e matemáticas, *scipy* para testes estatísticos, *statsmodels* para ajuste da regressão linear múltipla e *matplotlib* para a geração de gráficos e visualizações, além de outras bibliotecas utilizados de apoio.

O ambiente computacional utilizado foi em sistema operacional Windows, com o código armazenado e executados no Jupyter Notebook localmente. Além disso, os resultados da análise foram exportados em formato Excel, para facilitar a visualização e manipulação dos coeficientes extraídos do modelo.

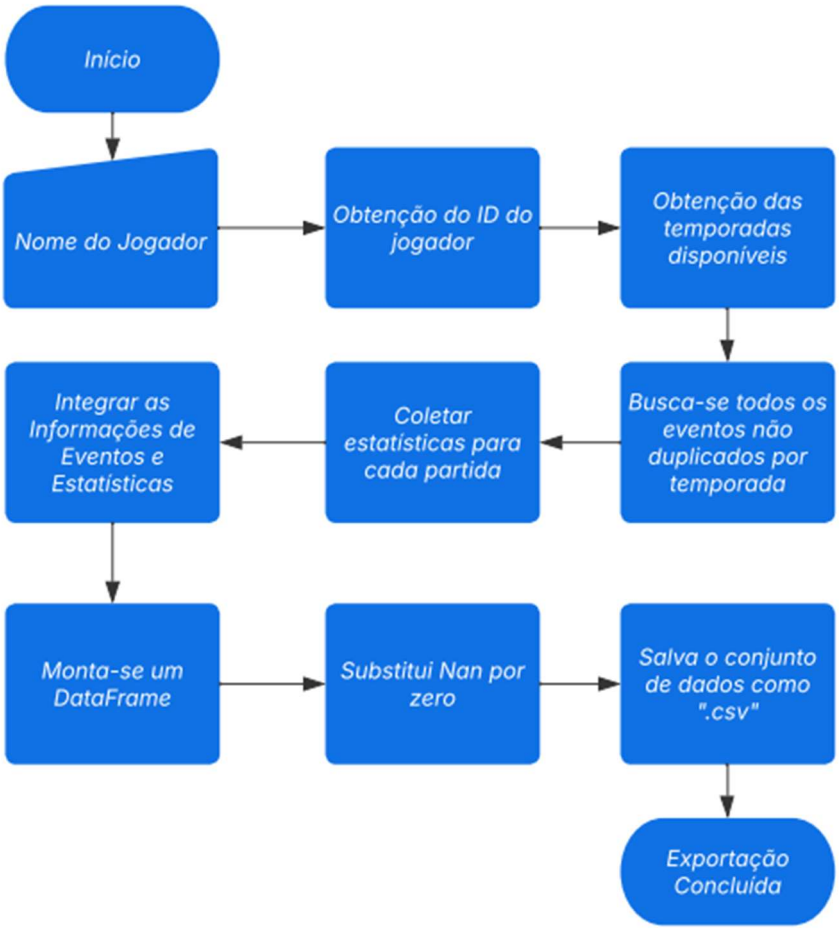
3.1 Extração dos Dados

Essa seção descreve o processo de obtenção dos dados que serviram de matéria-prima para a execução do presente estudo, descrevendo a automatização necessária para pegar dados em larga escala.

Os dados foram obtidos a partir da plataforma *SofaScore*, que fornece dados detalhados de partidas de futebol profissional, segmentadas por atleta e por jogo. Foram coletados dados referentes as últimas 5 temporadas de cada jogador,

abrangendo quatro categorias de posição: goleiros, defensores, meio-campistas e atacantes. As variáveis extraídas incluíram métricas de desempenho ofensivo, defensivo e de transição, bem como a nota global (rating) atribuída pela plataforma, utilizada como variável dependente. A coleta foi realizada por meio de acessos a API do SofaScore, disponibilizadas na plataforma RapidAPI e codificados em Python, utilizando a biblioteca *pandas* para manipulação das tabelas e *requests* para acesso a API.

Figura 1 – Fluxograma do Processo de Extração de Dados



Fonte: elaborado pelo autor (2025)

A Figura 1 ilustra o processo de coleta e organização dos dados estatísticos de um jogador de futebol. O processo se inicia na definição do jogador que se deseja avaliar,

informando o nome e clube, que serão processados para obter o identificador único (ID). Uma vez obtido o ID do jogador, são recuperadas as temporadas em que ele atuou, assim como os campeonatos disputados. Esses dados passam por um processo de estruturação, garantindo a coerência e veracidade. Após isso, identifica-se todas as partidas que o jogador atuou dentro dos campeonatos listados, garantindo que não haja duplicação dos dados. Dessa forma, já obtidas as partidas em que o jogador atuou e que se tem interesse em analisar, coleta-se os dados individuais do atleta. Desse processo obtém-se métricas detalhadas, como o número de passes, finalizações, gols, assistências, desarmes, precisão nos passes, cruzamentos, chutes a gol e rating, métrica de interesse desse estudo. Posteriormente, atribui-se as métricas obtidas às partidas em que ocorreram e os valores ausentes (NaN – Not a Number, representando dados não disponíveis) são substituídos pelo valor zero, representando zero ocorrências, uma vez que o evento não foi observado. Por fim, os dados são armazenados em um arquivo CSV.

3.2 Tratamento dos Dados

Esta seção descreve o processo de tratamento dos dados. Após a sua obtenção, esse procedimento é realizado para padronizar e preparar as informações de modo a servirem de entrada aos modelos estatísticos, tanto na fase de aprendizado quanto na de teste, assegurando a integridade e a coerência da análise.

Para garantir a integridade da análise, algumas variáveis foram descartadas por não se enquadrarem na posição tática do atleta, por exemplo, o estudo de quantos desarmes um goleiro faz durante uma partida não agrega valor, assim como o número de defesas feitas por um volante durante a partida não resulta em aprendizados relevantes. Entretanto, o número de defesas feitas por um goleiro ao longo da partida possui muita relevância, assim como o número de desarmes feitos por um zagueiro, lateral ou volante. Diante disso, o processo de verificação e remoção de atributos que não possuem significado para a posição do jogador em análise constitui uma etapa essencial da pesquisa. Além disso, também foram removidos dados comumente apresentados para análises de apostas esportivas, como gols esperados, probabilidade de marcar gol, probabilidade de dar assistências. Também foram

removidos os dados meramente informativos, como o nome das equipes, nome do estádio, juiz responsável por apitar a partida e data do jogo. Por fim, valores ausentes foram substituídos por zero, evitando distorções no modelo estatístico.

3.3 Modelagem Estatística

Uma vez obtidos e tratados os dados, a metodologia adotada para a análise dos dados foi a regressão linear múltipla, implementada utilizando a técnica de Mínimos Quadrados Ordinários (OLS) da biblioteca *statsmodels*. Essa abordagem permite estimar o impacto de diferentes estatísticas individuais sobre a nota de desempenho do jogador (*rating*), identificando quais variáveis apresentam maior relevância.

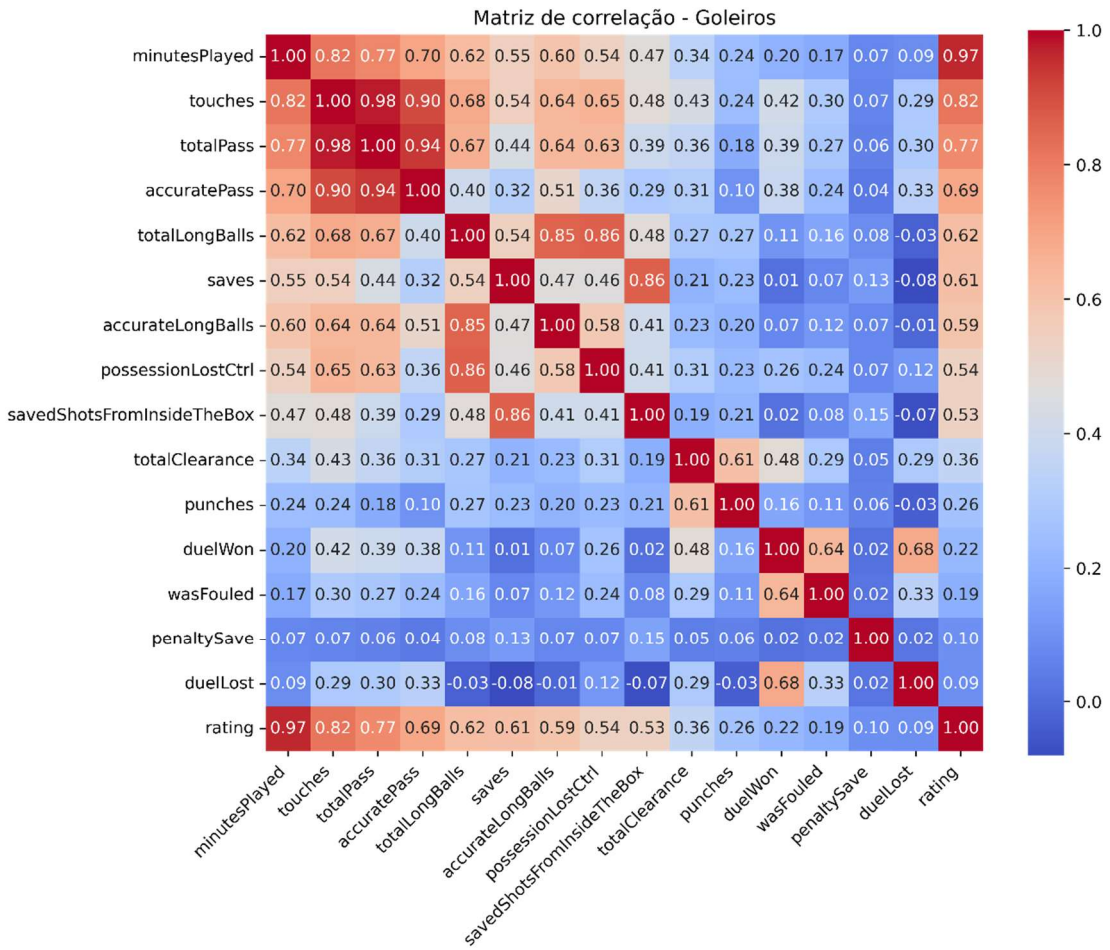
A implementação seguiu os seguintes passos:

1. Definição da variável dependente (*rating*) e das variáveis independentes (estatísticas do jogador);
2. Inclusão de um termo de intercepto no modelo (adição de uma constante ao grupo de variáveis independentes, para que o modelo não passe, obrigatoriamente, pela origem);
3. Ajuste do modelo aos dados utilizando *sm.OLS(Y, X)*;
4. Extração dos coeficientes das variáveis e ordenação por relevância.

Antes do ajuste dos modelos, foi realizada a análise de colinearidade entre as variáveis independentes por meio do cálculo da matriz de correlação de Pearson, considerando, para cada grupo posicional, as quinze métricas com maiores valores absolutos de correlação com a variável dependente (*rating*). Essa etapa visou identificar relações lineares fortes entre variáveis, que, embora possam indicar redundância informacional, também carregam, em conjunto, significado tático relevante. Assim, a eventual manutenção de métricas altamente correlacionadas foi avaliada à luz do contexto técnico de cada posição, evitando a exclusão automática de atributos cujo valor explicativo seja reforçado pela sua relação com outras métricas. As Figuras 2, 3, 4 e 5 apresentam, respectivamente, as matrizes de correlação para

goleiros, defensores, meio-campistas e atacantes, evidenciando os padrões de interdependência entre as métricas mais associadas ao desempenho global.

Figura 2 – Matriz de Correlação dos Dados de Goleiros

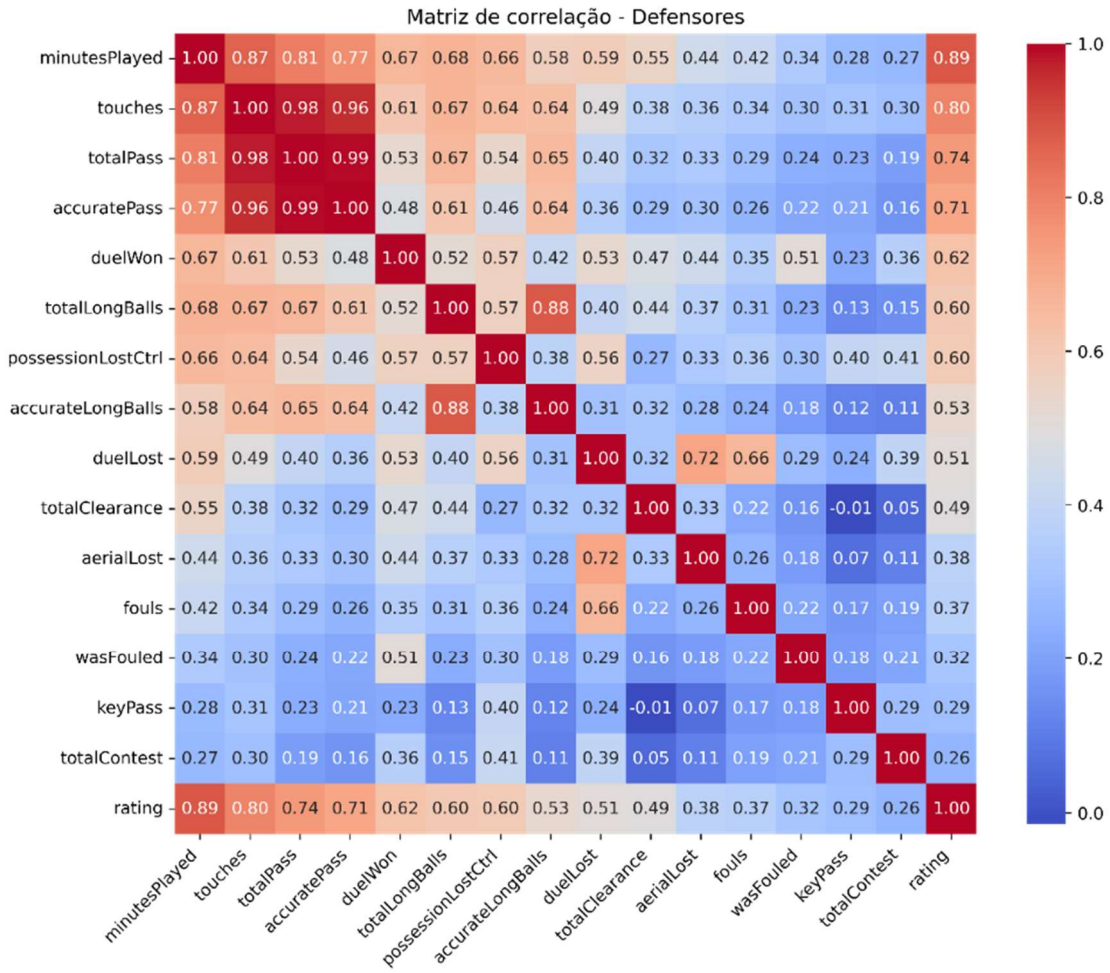


Fonte: elaborado pelo autor (2025)

A partir da análise da Figura 2, verifica-se que variáveis como minutos jogados, número de toques, total de passes, acurácia dos passes, total de bolas longas e número de defesas apresentam coeficientes de correlação superiores a 0,6 nos dados referentes aos goleiros. Todavia, ainda que a matriz de correlações aponte elevados coeficientes para essas métricas, a interpretação contextual demonstra que não seria adequado desconsiderá-las, pois evidentemente o tempo de jogo, o volume de defesas, a participação em passes e demais indicadores estão intrinsecamente

associados ao desempenho de um goleiro, refletindo de maneira significativa em sua avaliação. Dessa forma, optou-se por manter tais variáveis na construção do modelo preditivo.

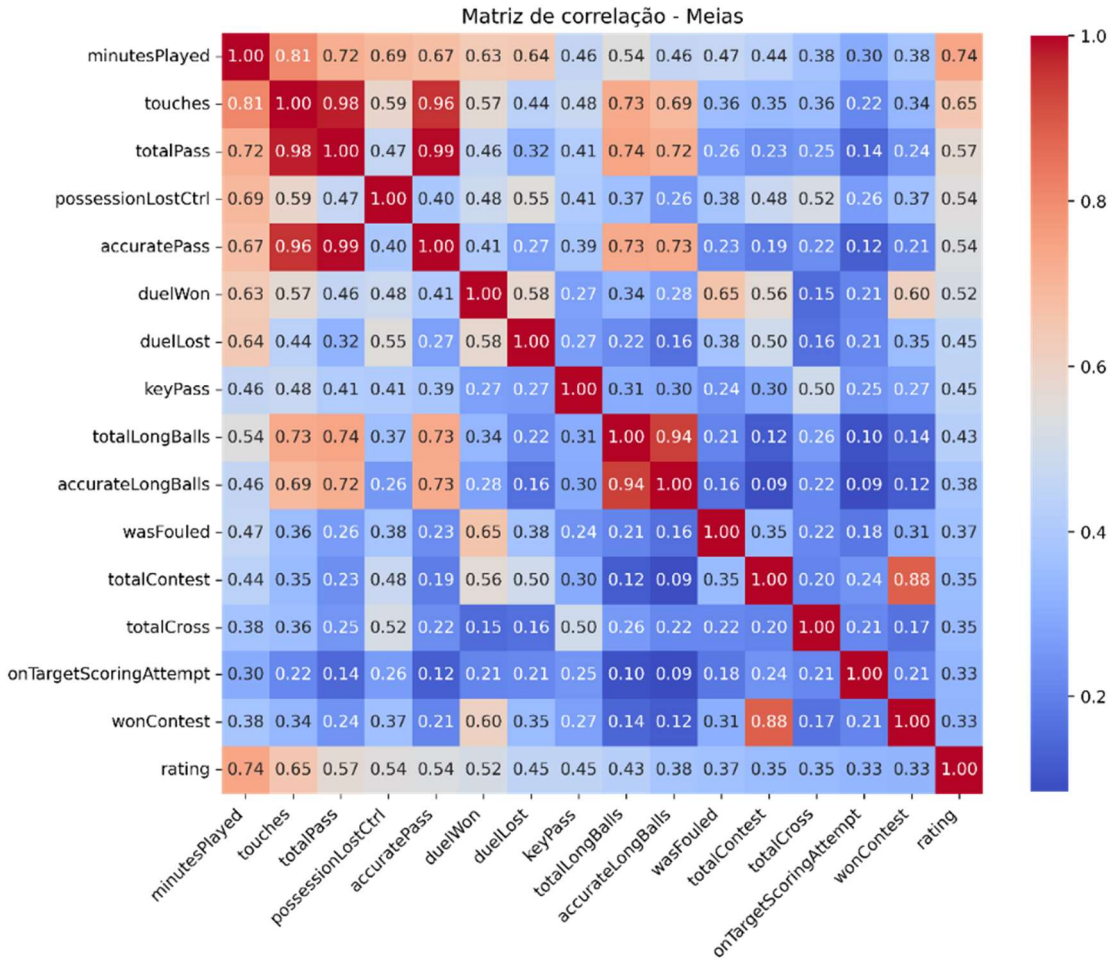
Figura 3 – Matriz de Correlação dos Dados de Defensores



Fonte: elaborado pelo autor (2025)

A Figura 3 apresenta a matriz de correlação referente aos dados dos defensores, evidenciando que variáveis como minutos jogados, número de toques, total de passes, acurácia dos passes e duelos ganhos apresentam elevados coeficientes de correlação. Tal resultado é coerente com a dinâmica da posição, uma vez que essas métricas representam aspectos centrais da atuação de um zagueiro, destacando tanto sua participação ativa nas ações defensivas quanto sua contribuição na construção do jogo.

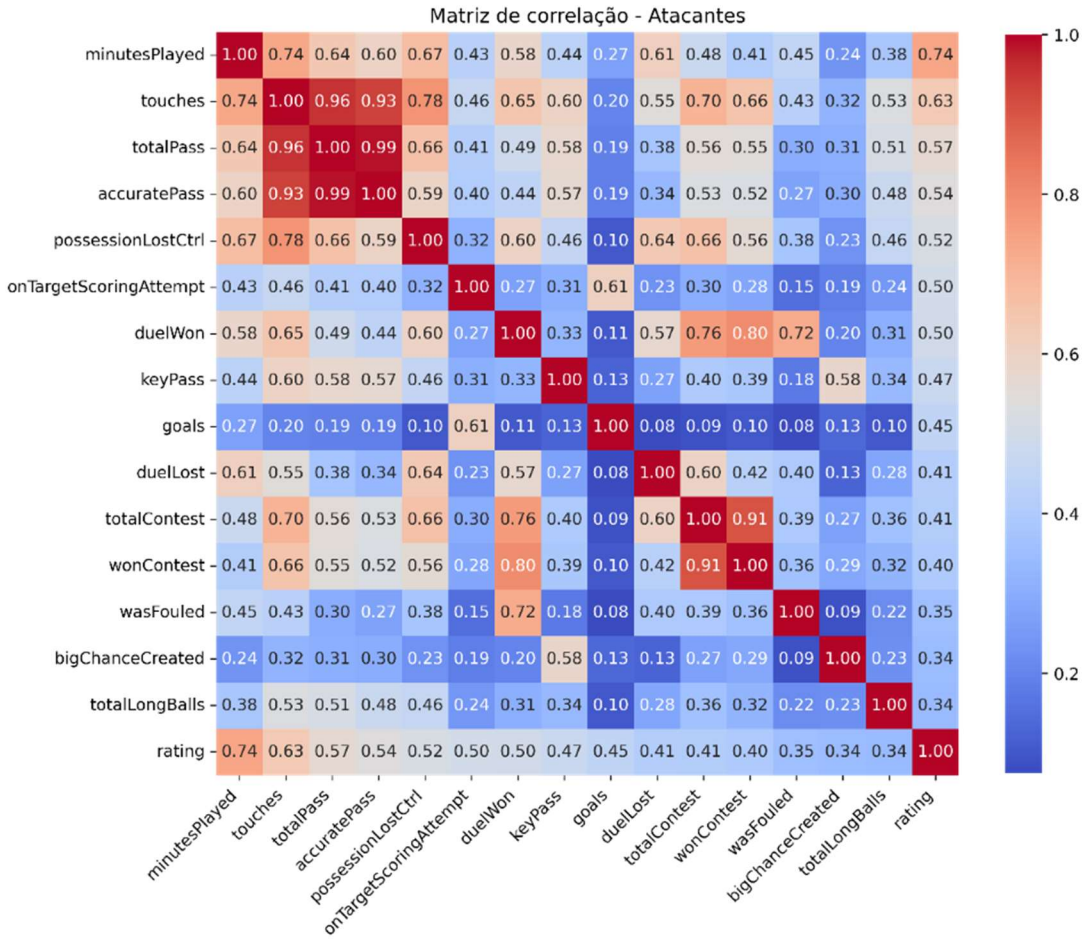
Figura 4 – Matriz de Correlação dos Dados de Meio-Campistas



Fonte: elaborado pelo autor (2025)

A Figura 4 apresenta a matriz de correlação referente aos meio-campistas, evidenciando que variáveis como minutos jogados, número de toques e total de passes apresentam coeficientes de correlação elevados. Esses resultados são consistentes com as funções desempenhadas pelos atletas dessa posição, uma vez que o meio-campo atua na área de transição entre defesa e ataque, demandando alta participação na circulação da bola, precisão no passe e eficiência em ações de combate e recuperação. Tais métricas, portanto, refletem adequadamente a relevância dessa função para a avaliação do desempenho.

Figura 5 – Matriz de Correlação dos Dados de Atacantes



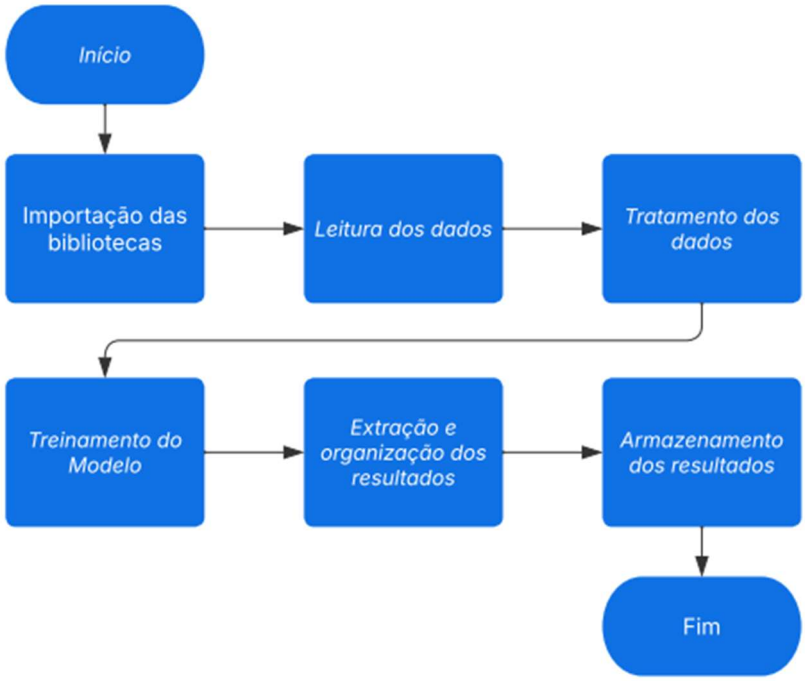
Fonte: elaborado pelo autor (2025)

A Figura 5, por sua vez, apresenta a matriz de correlação dos atacantes, destacando que métricas como minutos jogados, número de toques, total de passes, acurácia dos passes, finalizações e gols obtêm correlações expressivas. A interpretação é coerente com a lógica da posição, dado que a contribuição ofensiva de um atacante está diretamente associada à sua capacidade de participar ativamente da construção ofensiva e, sobretudo, de converter oportunidades em gols. Dessa forma, a manutenção dessas variáveis na modelagem é fundamental para captar a representatividade estatística e contextual do desempenho desses atletas.

Após a análise de correlação das variáveis, os dados obtidos via SofaScore foram separados em diferentes grupos, definidos pelas posições dos jogadores em campo, antes que fossem implementados. São eles:

- **Goleiros:** Grupo de dados que comportam todos os atletas que atuam como goleiros em suas equipes;
- **Defensores:** Grupo de dados que comportam todos os atletas que atuam como lateral direito, lateral esquerdo e zagueiros em suas equipes;
- **Meio-Campistas:** Grupo de dados que comportam todos os atletas que atuam como volante ou meia em suas equipes;
- **Atacantes:** Grupo de dados que comportam todos os atletas que atuam como centroavante, ponta-direita ou ponta-esquerda em suas equipes.

Figura 6 – Fluxograma do Processo de Treinamento e Testes do Modelo



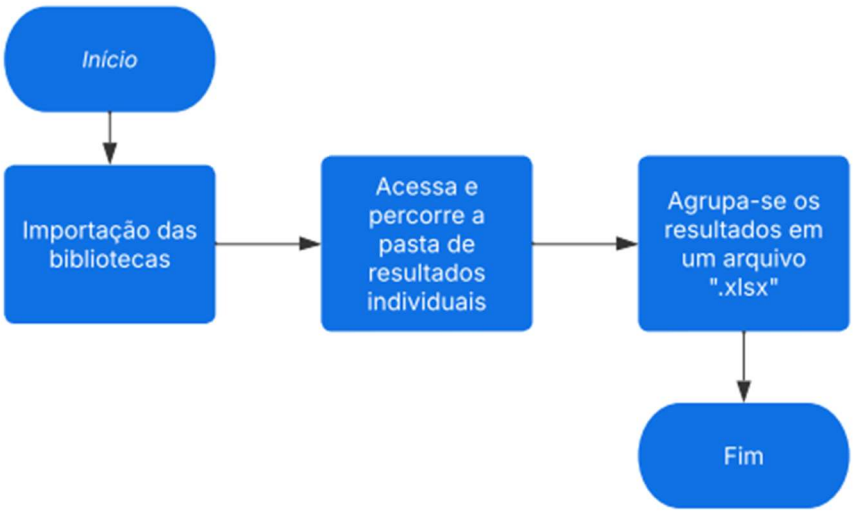
Fonte: elaborado pelo autor (2025)

A Figura 6 ilustra o processamento dos dados obtidos. Inicialmente, importam-se as bibliotecas necessárias, listadas abaixo:

- pandas versão 1.3.4;
- numpy versão 1.20.3;
- statsmodels versão 0.12.2;
- os (módulo da biblioteca padrão do Python 3.9.7);
- scipy (submódulo scipy.stats) versão 1.7.1;
- matplotlib (submódulo matplotlib.pyplot) versão 3.4.3;

Importadas as bibliotecas, parte-se para a leitura dos dados, feitos a partir do caminho do arquivo CSV que contêm os dados que serão analisados. Após a leitura dos dados, realiza-se o processo de tratamento, descartando variáveis que não interessam ao modelo e substituindo os dados faltantes por zero. Uma vez que se tem os dados lidos, organizados e tratados, define-se o modelo de regressão linear múltipla, definindo as variáveis independentes (atributos) e a variável dependente (*rating*). Após a aplicação do modelo, extrai-se os coeficientes das variáveis independentes, cria-se um dicionário associando as variáveis aos seus respectivos coeficientes, que servirá para a criação de um DataFrame que será ordenado pelos coeficientes de forma decrescente. Dessa forma, salva-se os dados dos coeficientes em uma planilha excel individual do jogador.

Figura 7 – Processo de Agrupamento dos Resultados Individuais dos Atletas



Fonte: elaborado pelo autor (2025)

Após o processamento individual, no mesmo arquivo, executa-se a etapa de consolidação dos resultados individuais dos atletas. Conforme é mostrado na Figura 7, o script percorre o diretório que contém as planilhas de resultados por atleta, importa cada arquivo, concatena os registros em uma única planilha e, ao final, calcula e adiciona uma linha com a média dos resultados.

3.4 Avaliação do Modelo

Para a avaliação da performance do modelo de regressão linear múltipla proposto, foram utilizados indicadores estatísticos consagrados, como o coeficiente de determinação (R^2) e a raiz do erro quadrático médio (RMSE). O R^2 mensura a proporção da variância total da variável dependente que é explicada pelas variáveis independentes do modelo, funcionando como um indicador de qualidade do ajuste. Seus valores variam entre 0 e 1, sendo que valores próximos de 0 indicam baixo poder explicativo do modelo, enquanto valores próximos de 1 refletem elevada capacidade de explicação da variabilidade dos dados. Já o RMSE quantifica, em média, o desvio entre os valores observados e os valores previstos pelo modelo, sendo especialmente útil para avaliar o erro de predição em unidades compatíveis com a variável dependente. Os dois indicadores são utilizados para avaliar a capacidade de generalização dos modelos treinados por posição, aplicando cada um deles a jogadores específicos e comparando seu desempenho preditivo. Dessa forma, analisa-se, por exemplo, como o modelo treinado com dados de uma determinada posição se comporta ao ser aplicado na predição do desempenho de atletas de outras funções, possibilitando verificar a pertinência da segmentação posicional adotada e a robustez dos coeficientes estimados.

4. RESULTADOS

4.1 Capacidade Preditiva

Esta seção descreve os principais resultados obtidos após o treinamento e teste dos modelos, para cada uma das quatro separações táticas pré-definidas, de modo que se busque discutir os resultados obtidos e a qualidade de generalização dos modelos.

Tabela 1 – Conjunto de Treinamento dos Modelos

Grupo	Número de Jogadores	Número de dados
Goleiro	20	2.949
Defensores	20	3.691
Meio-campistas	20	2.890
Atacantes	20	3.158
Total	**80**	**12.688**

Fonte: elaborado pelo autor (2025)

A aplicação da regressão linear múltipla por posição gerou modelos com diferentes níveis de ajuste e capacidade de generalização. O conjunto de treinamento, composto por 80 jogadores distribuídos entre as quatro funções táticas, resultou em um banco de mais de 12 mil registros de partidas, conforme detalhado na Tabela 1. Essa amostra ampla permitiu explorar padrões específicos de cada função em campo.

Após a realização do treinamento do modelo com dados de 20 diferentes atletas por posição, foram realizados 8 testes, sendo dois para cada posição pré-definida. Os atletas dos testes foram os goleiros Yann Sommer (Suíça) e Yassine Bounou (Marrocos), os defensores Ruben Dias (Portugal) e Léo Ortiz (Brasil), os meio-campistas Declan Rice (Inglaterra) e João Gomes (Brasil) e os atacantes Messi (Argentina) e Darwin Núnez (Uruguai), de modo a garantir a diversidade de estilos de jogo, características físicas, ligas de atuação e origem dos atletas, permitindo avaliar

a capacidade de generalização dos modelos em contextos distintos de atuação em campo.

Tabela 2 – Resultados dos Modelos por Posição

		Posições do Conjunto de Treinamento							
		Goleiros		Defensores		Meio-campistas		Atacantes	
Conjunto de Teste		R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
Goleiros	Yann Sommer	**0,9830**	**0,4444**	0,9672	0,6169	0,9629	0,6566	0,9332	0,8809
	Yassine Bounou	**0,9809**	**0,4364**	0,9719	0,5298	0,9666	0,5778	0,9318	0,8249
Defensores	Ruben Dias	0,8973	1,0320	**0,9939**	**0,2508**	0,9927	0,2749	0,9912	0,3013
	Léo Ortiz	0,8098	1,2058	**0,9793**	**0,3982**	0,9791	0,3997	0,9711	0,4704
Meio-campistas	Declan Rice	0,4632	1,5234	0,9818	0,2808	**0,9913**	**0,1944**	0,9825	0,2751
	João Gomes	0,1754	2,6329	0,9856	0,3475	**0,9935**	**0,2340**	0,9920	0,2586
Atacantes	Lionel Messi	-0,4964	3,2566	0,9255	0,7264	0,9821	0,3562	**0,9834**	**0,3426**
	Darwin Núnez	0,3733	2,0610	0,9337	0,6702	0,9377	0,6496	**0,9361**	**0,6581**

Fonte: elaborado pelo autor (2025)

Na avaliação dos modelos (Tabela 2), observou-se desempenho satisfatório para goleiros, com valores de R^2 superiores a 0,98 quando testados com o modelo treinado com os dados dos goleiros. Além disso, os testes de Yann Sommer e Yassine Bounou obtiveram R^2 inferior a 0,98 em todos os demais modelos (defensores, meio-campistas e atacantes), indicando que variáveis como número de defesas, saídas do gol e defesas de pênalti explicam de forma consistente as notas atribuídas a atletas dessas posições. O baixo RMSE reforça a capacidade preditiva do modelo nesses casos.

No caso dos defensores Ruben Dias e Léo Ortiz, os valores de R^2 também foram os maiores para a função de origem desses atletas, sendo 0,9939 e 0,9793, respectivamente. Além disso, o RMSE apresenta erro baixo em ambos os casos, indicando a assertividade do modelo de defensores em predizer notas de dois atletas

da mesma função tática. Quando utilizados para os modelos de goleiros, meio-campistas e atacantes, o desempenho apresentado foi inferior (R^2 menor) e a raiz do erro quadrático médio (RMSE) apresentou valores superiores.

Entre os meio-campistas, os resultados apresentaram maior divergência em função do modelo utilizado. Enquanto testados com o modelo de meio-campistas, Declan Rice e João Gomes apresentaram valores de R^2 próximos de 0,99. Quando os mesmos atletas foram utilizados no teste do modelo de goleiros, o valor de R^2 apresentou grande redução, sendo de 0,46 e 0,17, o que confirma a pertinência da segmentação posicional. O valor do RMSE também confirma a necessidade da segmentação dos dados, uma vez que foi de 0,19 e 0,23 para o modelo de meio-campistas e de 1,52 e 2,63 para o modelo de goleiros.

Para os atacantes, os resultados também confirmaram a pertinência da segmentação posicional, uma vez que Lionel Messi apresentou R^2 negativo quando avaliado pelo modelo de goleiros, evidenciando a inadequação de métricas defensivas para explicar o desempenho ofensivo. Já no modelo específico de atacantes, tanto Messi (R^2 = 0,9834; RMSE = 0,3426) quanto Darwin Núñez (R^2 = 0,9361; RMSE = 0,6581) alcançaram valores elevados, indicando boa capacidade de predição. Contudo, ao comparar com os modelos de defensores e meio-campistas, nota-se que o modelo de atacantes apresentou valores de R^2 e RMSE semelhantes, se destacando como o mais consistente somente em relação ao modelo de goleiros dentre as quatro funções táticas. Esse comportamento pode estar relacionado a diferenças de estilo de jogo entre os atacantes do conjunto de teste e aqueles do conjunto de treinamento, ou ainda a necessidade de ampliar a base de dados para fortalecer a robustez do modelo nesta posição.

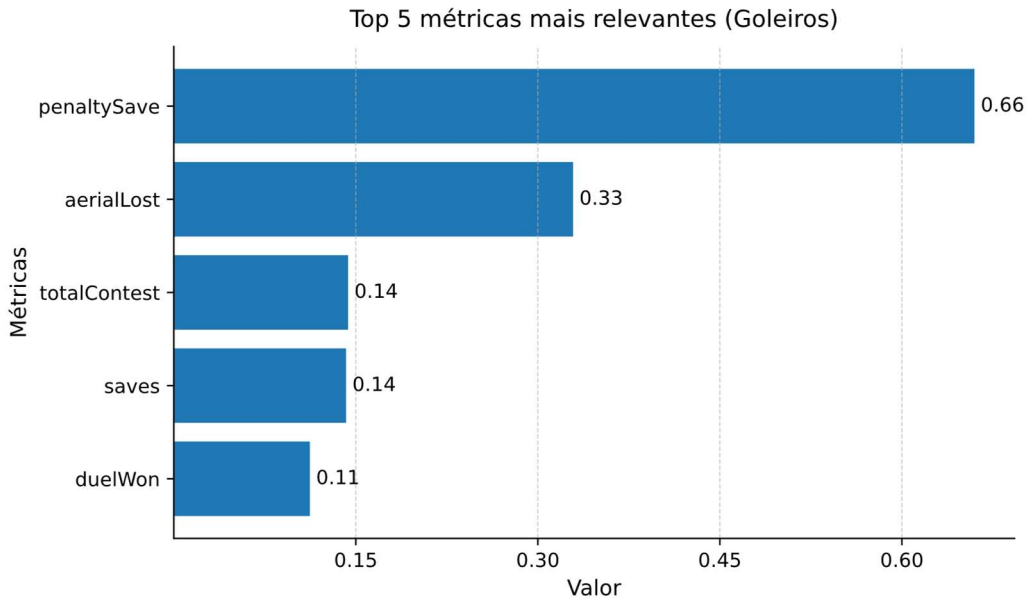
De modo geral, os modelos confirmaram que métricas comumente associadas a determinadas posições, como defesas para goleiros, desarmes para defensores, passes certos para meio-campistas e finalizações para atacantes, explicam de forma consistente as notas atribuídas aos atletas. Observou-se que os modelos de goleiros, defensores e meio-campistas apresentaram maior estabilidade, com elevados valores de R^2 e baixos erros médios, enquanto o modelo de atacantes, embora também tenha atingido bom desempenho, mostrou-se menos robusto em comparação, destacando-se apenas em relação ao modelo de goleiros. Esses resultados evidenciam que a segmentação posicional é fundamental para a construção de modelos explicativos,

mas também sugerem que determinadas funções táticas, especialmente no ataque, exigem o desenvolvimento de abordagens complementares ou a ampliação da base de dados para captar variabilidades que não são plenamente explicadas pelas métricas tradicionais.

4.2 Atributos com Maior Impacto no Rating

A análise dos coeficientes estimados pelos modelos de regressão linear múltipla permitiu identificar, para cada posição tática, os atributos com maior impacto sobre o rating dos atletas. As Figuras 8, 9, 10 e 11 apresentam os cinco preditores mais relevantes em cada caso, destacando a magnitude de sua influência.

Figura 8 – Cinco Métricas com Maior Relevância para Goleiros

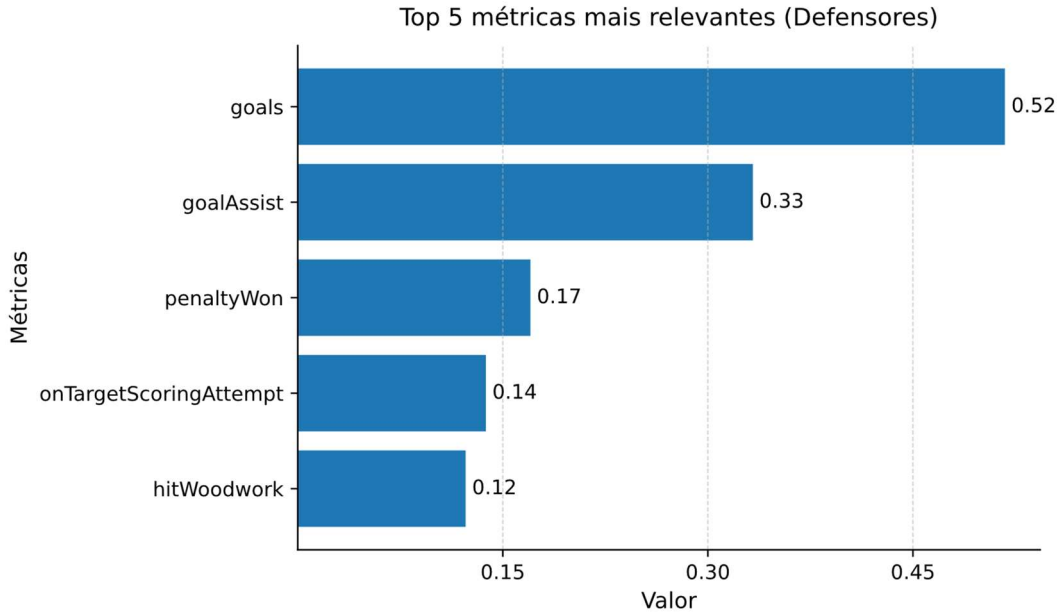


Fonte: elaborado pelo autor (2025)

A Figura 8 apresenta as cinco métricas relacionadas diretamente à função de proteção do gol que apresentaram maior peso no modelo. Defesas realizadas, saídas bem-sucedidas e defesas de pênalti destacaram-se como fatores determinantes para a

avaliação, reforçando a predominância de métricas ligadas à eficácia na prevenção de gols. Isso evidencia a aderência do modelo ao papel central do goleiro e confirma achados prévios de que indicadores defensivos específicos são fundamentais para a mensuração de seu desempenho.

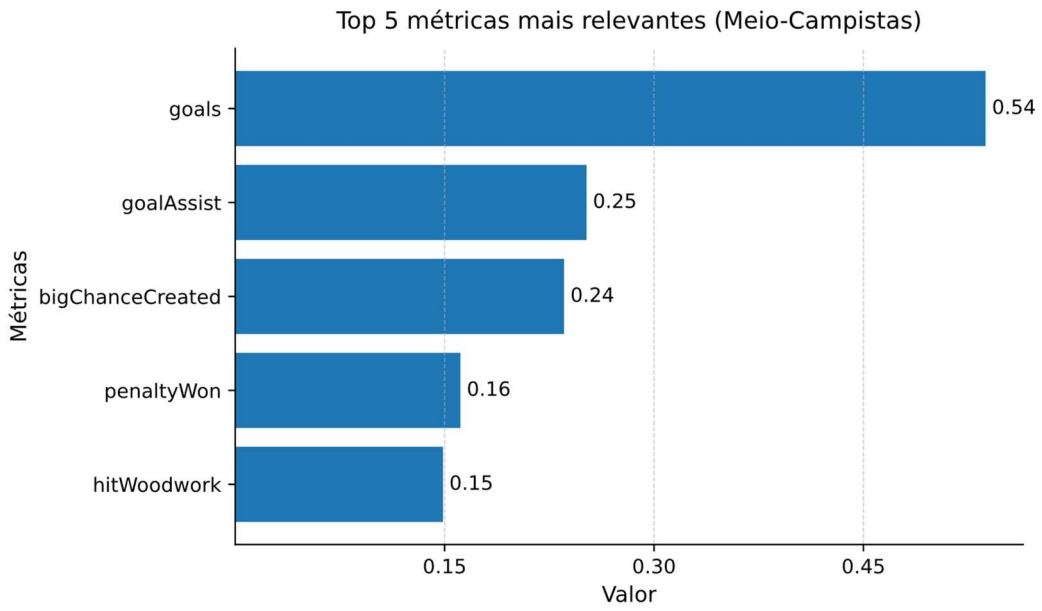
Figura 9 – Cinco Métricas com Maior Relevância para Defensores



Fonte: elaborado pelo autor (2025)

Entre os defensores (Figura 9), os atributos de caráter defensivo, como desarmes, interceptações e cortes, não assumiram maior relevância. Observa-se também impacto de variáveis ofensivas, como gols, assistências, gols de pênalti, chutes no gol e finalizações na trave, que refletem o papel híbrido de defensores modernos, mas principalmente a supervalorização das ações de ataque no futebol. Esses resultados indicam que a avaliação estatística do desempenho defensivo não pode ser restrita às ações defensivas, devendo considerar também a capacidade de manter a posse e contribuir para a progressão ofensiva do jogo.

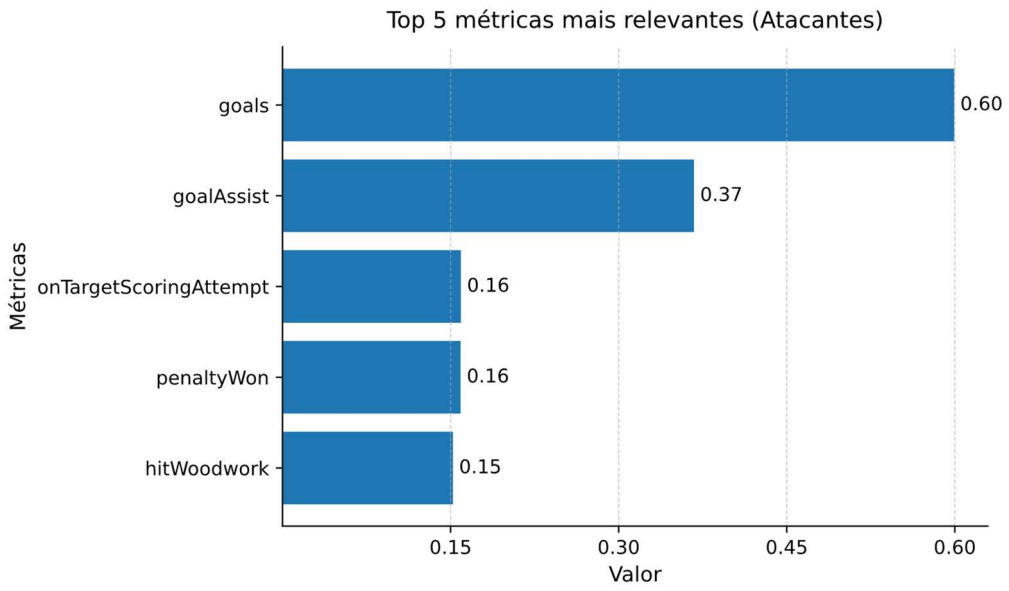
Figura 10 – Cinco Métricas com Maior Relevância para Meio-Campistas



Fonte: elaborado pelo autor (2025)

No caso dos meio-campistas (Figura 10), verificou-se predominância de métricas ligadas à distribuição e controle de jogo, como número assistências e grandes chances criadas. Ressalta-se que 4 das 5 métricas mais relevantes são comuns aos defensores e, embora estejam mais aderentes ao estilo de jogo de um meio-campista e por isso apresentam coeficientes mais altos, reforça a valorização de ações ofensivas em uma partida de futebol.

Figura 11 – Cinco Métricas com Maior Relevância para Atacantes



Fonte: elaborado pelo autor (2025)

Por fim, entre os atacantes (Figura 11), o modelo apontou os gols, as assistências, os chutes no gol, os gols de pênalti e as finalizações na trave como os atributos mais significativos para o rating. Esses coeficientes refletem a natureza da função de um atacante, destacando a importância de ações ofensivas e a produção direta de gols. Além disso, destaca a importância de um atacante servir a sua equipe com assistências e não somente com gols.

De maneira geral, a análise evidencia um conjunto semelhante de variáveis predominantes para as diferentes funções táticas, com exceção para os goleiros que predominantemente dependem das ações defensivas em campo para uma boa performance. Entretanto, os diferentes valores dos coeficientes destacam a crescente importância de ações ofensivas em posições que naturalmente se espera ações diretas na produção de gols de uma equipe, confirmando a pertinência da segmentação posicional adotada. A diferença dos coeficientes e, no caso dos goleiros, as diferentes métricas de maior relevância, destacam que avaliações globais de desempenho, sem a diferenciação tática, tendem a distorcer a contribuição real de jogadores, penalizando funções de natureza defensiva.

5. CASOS DE USO NO FUTEBOL

Os resultados obtidos nesta pesquisa possuem relevância prática que excedem o ambiente acadêmico, encontrando aplicações diretas no cotidiano de diferentes agentes do futebol e de áreas correlatas.

Figura 12 – Ideação da Jornada do Usuário

Fonte: elaborado pelo autor (2025)

A Figura 12 apresenta a ideação da jornada do usuário interno dos clubes profissionais de futebol, como um analista de desempenho de uma equipe. O profissional responsável por fornecer análises que apoiem a escalação da equipe, detectar possíveis alvos para contratação e atletas abaixo do desempenho esperado para liberação no mercado ou venda, busca no uso dos dados o embasamento necessário para ser assertivo em suas decisões. Dessa forma, recorrendo ao modelo previamente treinado, o processo de análise e apresentação dos resultados ocorre de forma automatizada e fundamentada em dados e ferramentas estatísticas.

Tabela 3 – Aplicações do Modelo de Análise no Cotidiano

Stakeholder	Caso de uso	Valor gerado	Artefatos e Visualização
Scout/Analista de desempenho (usuário primário)	Escolher e aplicar métricas segmentadas por posição; gerar relatórios comparativos; interpretar resultados.	Avaliação objetiva, rápida e comparável entre atletas.	Dashboard com ranking por posição; relatórios automáticos; assistente de métricas-chave.
Comissão técnica (treinadores e auxiliares)	Compreender pontos fortes e fracos dos jogadores; apoiar escalações e estratégias de jogo.	Melhor tomada de decisão tática; uso de evidências complementando percepção subjetiva.	Relatórios visuais; comparações entre atletas; simulações de impacto no time.
Diretoria e gestão do clube	Apoiar decisões de contratação e venda de jogadores com base em evidências estatísticas.	Redução de risco financeiro; contratações mais assertivas.	Relatórios; gráficos para análise de custo-benefício; filtros por posição.
Jogadores (atletas avaliados)	Acompanhar desempenho individual; identificar áreas de melhoria.	Guia estatístico sobre evolução; foco em treinamentos específicos.	Guias individuais; feedback de desempenho; relatórios de progresso ao longo da temporada.
Mídia esportiva e torcedores	Acessar dados comparativos para análises e narrativas.	Conteúdo rico e fundamentado em estatísticas.	Infográficos interativos; rankings públicos; comparação de jogadores.

Fonte: elaborado pelo autor (2025)

Além das aplicações no cotidiano de analistas de desempenho, a Tabela 3 apresenta possíveis aplicações para outros agentes do esporte como comissões técnicas, diretoria e gestores dos clubes, jogadores em atividade, mídia e torcedores.

No âmbito da comissão técnica, a aplicação do modelo possibilita compreender de forma objetiva os pontos fortes e fracos de cada atleta, auxiliando tanto na definição

das escalações quanto na elaboração de estratégias de jogo mais consistentes. Já para a diretoria e gestores do clube, os resultados obtidos com o modelo representam uma ferramenta estratégica na condução do mercado de transferências. Ao apoiar decisões de contratação e venda de jogadores com base em evidências estatísticas, o processo torna-se mais transparente e fundamentado, reduzindo riscos financeiros e favorecendo contratações mais assertivas.

Do ponto de vista dos atletas, a utilização dessas métricas contribui para o autoconhecimento e o desenvolvimento profissional. A análise individualizada de desempenho permite que o jogador compreenda de forma clara quais ações em campo possuem maior peso para sua avaliação, possibilitando o direcionamento de treinamentos específicos que favoreçam sua evolução.

No contexto midiático, plataformas esportivas e veículos de comunicação podem se apropriar de tais resultados para enriquecer a cobertura jornalística, oferecendo análises mais consistentes e embasadas em evidências quantitativas. Esse movimento favorece a aproximação entre ciência de dados e torcedores, promovendo uma compreensão mais ampla do desempenho esportivo.

Por fim, cabe destacar a aplicabilidade em ambientes educacionais e tecnológicos. A sistematização de métodos de coleta, tratamento e modelagem de dados esportivos pode ser incorporada em projetos de iniciação científica e tecnológica, fomentando a formação de novos profissionais capacitados a atuar na interface entre esporte, estatística e ciência de dados.

Assim, observa-se que a pesquisa realizada não apenas cumpre um papel de inovação metodológica, mas também apresenta resultados com impacto concreto no cotidiano do futebol e em áreas relacionadas, reforçando a relevância da integração entre ciência de dados e prática esportiva.

6. CONSIDERAÇÕES FINAIS

Portanto, o presente estudo evidenciou a relevância da aplicação de técnicas de modelagem estatística, em especial a regressão linear múltipla, na análise de desempenho de jogadores de futebol em diferentes funções táticas. O processo de segmentação posicional mostrou-se essencial para identificar métricas específicas de cada função em campo, além de quantificar a contribuição de diferentes ações, reduzindo vieses e permitindo avaliações mais justas e contextualizadas, bem como a atribuição de ratings consistentes à performance dos atletas.

Os resultados indicaram que métricas defensivas, como interceptações, desarmes e defesas, explicam melhor o desempenho de goleiros, zagueiros e laterais. De forma análoga, ações ofensivas como gols, assistências, finalizações na trave, chutes no gol e pênaltis convertidos se mostraram determinantes para os ratings de meio-campistas e atacantes. Embora as métricas ofensivas apresentem relevância em todas as posições, goleiros e defensores têm coeficientes menores para essas ações, reforçando a importância da abordagem segmentada.

Adicionalmente, o estudo evidencia o potencial da integração da ciência de dados no esporte como suporte às tomadas de decisão baseadas em evidências. A sistematização da coleta, tratamento, interpretação e apresentação dos resultados permite que analistas de desempenho, comissões técnicas, gestores de clubes, departamentos médicos e torcedores se beneficiem de análises pautadas em dados consistentes, modelos estatísticos consolidados e ferramentas de automação, ampliando a objetividade e a eficiência no cotidiano do futebol profissional.

Em síntese, este trabalho não apenas cumpre o objetivo de identificar os atributos estatísticos mais relevantes para cada posição no futebol, como também reforça a importância de análises baseadas em evidências no esporte moderno. Ao mesmo tempo em que oferece contribuições práticas imediatas para scouts, analistas e gestores, abre caminho para investigações futuras que incorporem bases de dados mais amplas, métricas físicas e contextuais, bem como a exploração de novos modelos estatísticos e de aprendizado de máquina. Dessa forma, consolida-se como um passo importante na integração entre ciência de dados e futebol, fortalecendo a tomada de decisões objetivas em um cenário cada vez mais competitivo.

Referências

AFONSO, José et al. Regression models are not inherently predictive: an educational review on why sports exercise and medicine research should be more careful with the term “predictors”. **SportRxiv**, 2024. DOI: 10.51224/SRXIV.405. Disponível em: https://sportrxiv.org/index.php/server/preprint/view/405/851. Acesso em: 15 mar. 2025.

ALI, Ajmol. Measuring soccer skill performance: a review. **Scandinavian Journal of Medicine & Science in Sports**, v. 21, n. 2, p. 170-183, 2011. DOI: 10.1111/j.1600-0838.2010.01256.x.

ASSUNÇÃO, João Vitor Oliveira; SANTANA, André Luiz Maciel; DURÃO, Luiz Fernando Cardoso dos Santos. Construção de métricas físicas no futebol: uma abordagem por visão computacional e ciência de dados Insper. In**: SIMPÓSIO DE INICIAÇÃO CIENTÍFICA E TECNOLÓGICA DO INSPER – SEMANA DE PESQUISA NA GRADUAÇÃO**, 2., 2024, São Paulo. Anais…. São Paulo: Insper, 2024. p. [não paginado].

BALL, James; HUYNH, Minh; VARLEY, Matthew C. Comparing player rating systems as a metric for assessing individual performance in soccer. **Journal of Sports Sciences**, v. 43, n. 7, p. 676-686, 2025. DOI: 10.1080/02640414.2025.2471208.

FERNANDES FILHO, Ricardo Palmieri. **O processo decisório nas organizações esportivas:** análise de dados ou empirismo em clubes de futebol do Brasil. 2018. Trabalho de Conclusão de Curso (Bacharelado em Administração) – Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2018. Disponível em: http://hdl.handle.net/11422/8434. Acesso em: 15 mar. 2025.

KOLOGLU, Yunus et al. A multiple linear regression approach for estimating the market value of football players in forward position. **arXiv preprint**, arXiv:1807.01104, 2018. DOI: 10.48550/arXiv.1807.01104.

MIRANDA, Leonardo. Empresa de análise e estatística detalha trabalho e revela planos de expansão ao Brasil. **Globo Esporte**, 17 fev. 2021. Disponível em: https://ge.globo.com/blogs/paineltatico/post/2021/02/17/empresa-de-analise-e-estatistica-detalha-trabalho-e-revela-planos-deexpansao-ao-brasil.ghtml. Acesso em: 15 mar. 2025.

NAKAMURA, Wilson Toshiro; CERQUEIRA, Sérgio de Almeida. A nova era do futebol brasileiro e clubes geridos como negócio. **Revista de Administração Contemporânea**, Rio de Janeiro, v. 25, n. 4, art. e210055, 2021. DOI: 10.1590/1982-7849rac2021210055. Acesso em: 11 mar. 2025.

ORTOLAN, Luís; SILVA, Diego. Descoberta de conhecimento em dados de scout do Campeonato Brasileiro de Futebol. In: **ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL (ENIAC)**, 2020. Anais…. Porto Alegre: Sociedade Brasileira de Computação, 2020. p. 626-637.

RODRIGUES, Sandra Cristina Antunes. **Modelo de regressão linear e suas aplicações**. 2012. Dissertação (Mestrado em Estatística) – Universidade da Beira Interior, Covilhã, 2012.

SHAMAH, Manoel Eduardo do Prado. **Análise de desempenho no futebol:** a prática do analista de desempenho nas categorias de base dos clubes brasileiros da Série A. 2021. Trabalho de Conclusão de Curso – [s.l.: s.n.], 2021. SOFASCORE. Sobre o rating do Sofascore. [s.d.]. Disponível em: https://corporate.sofascore.com/br/about/rating. Acesso em: 22 maio 2025.

WANG, Siyuan; CHEN, Jinming; CHEN, Haolong. Player score prediction based on multiple linear regression model. **Advances in Engineering Technology Research**, v. 4, n. 1, p. 246-246, 2023.

WUNDERLICH, Fabian; MEMMERT, Daniel. The betting odds rating system: using soccer forecasts to forecast soccer. **PLOS ONE**, v. 13, n. 6, e0198668, 2018. DOI: 10.1371/journal.pone.0198668.