UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

PEDRO AUGUSTO CARNEIRO DE OLIVEIRA

UM SISTEMA DE PREVISÃO DE DESEMPENHO DE JOGADORES DE FUTEBOL BASEADO EM DADOS ESTATÍSTICOS DE JOGOS

Ouro Preto
2026

PEDRO AUGUSTO CARNEIRO DE OLIVEIRA

UM SISTEMA DE PREVISÃO DE DESEMPENHO DE JOGADORES DE FUTEBOL BASEADO EM DADOS ESTATÍSTICOS DE JOGOS

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Reinaldo Silva Fortes
Coorientador: Prof. M.Sc. Renato Lopes Moreira

Ouro Preto
2026

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Pedro Augusto Carneiro de Oliveira

Um sistema de previsão de desempenho de jogadores de futebol baseado em dados estatísticos de jogos

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 3 de Março de 2026

Membros da banca

Reinaldo Silva Fortes (Orientador) - Doutor - Universidade Federal de Ouro Preto
Renato Lopes Moreira (Coorientador) - Mestre - Universidade Federal de Ouro Preto
Rodrigo César Pedrosa Silva (Examinador) - Doutor - Universidade Federal de Ouro Preto
Leandro Vinhas de Paula (Examinador) - Doutor - Universidade Federal de Ouro Preto

Reinaldo Silva Fortes, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 03/03/2026

Documento assinado eletronicamente por **Reinaldo Silva Fortes**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 03/03/2026, às 14:26, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **1062480** e o código CRC **B5A892C4**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

Agradecimentos

Gostaria de expressar minha sincera gratidão a todos que, de alguma forma, contribuíram para que este trabalho fosse possível e para que a minha trajetória acadêmica na UFOP fosse tão significativa.

Agradeço, em especial, aos professores Reinaldo e Renato, pela disposição e por aceitarem participar deste projeto, compartilhando conhecimento, orientação e confiança.

À minha família, deixo meu mais profundo reconhecimento pelo apoio incondicional, pelo incentivo constante e pela compreensão nos momentos de maior dedicação à graduação.

Sou grato também aos meus amigos de período, que tornaram os desafios do curso mais leves e transformaram o cotidiano em uma experiência de aprendizado coletivo. Aos colegas do Centro Acadêmico, pela parceria, convivência e por contribuírem para a minha formação dentro e fora da sala de aula.

De modo geral, agradeço a todos que estiveram ao meu lado ao longo deste tempo na UFOP. Cada pessoa, à sua maneira, deixou uma marca importante e contribuiu para que eu chegasse até aqui, não apenas como estudante, mas como indivíduo em constante evolução.

Resumo

Este trabalho tem como objetivo desenvolver um sistema de recomendação de jogadores de futebol baseado na predição de desempenho, utilizando dados do Campeonato Brasileiro Série A (2022-2024). A proposta previa duas etapas: a modelagem preditiva de métricas individuais e a geração de recomendações. Este trabalho concentrou-se na primeira etapa, implementando e avaliando três modelos de *machine learning* (*Random Forest*, *XGBoost* e MLP) com registros de todos os jogadores do campeonato através da plataforma *FootyStats*. A análise exploratória revelou desafios significativos: distribuições *zero-inflated* em métricas ofensivas (35–40% com zero gols), tamanho amostral reduzido para goleiros (21 no treino) e correlação temporal moderada (r=0,584 para gols), estabelecendo R^2 máximo teórico de 0,341. Os resultados mostraram R^2=0,181 para gols (MLP), representando 53% do limite teórico, e R^2 negativo para a maioria das demais métricas. A análise demonstrou que 40–60% da variação é fundamentalmente imprevisível devido ao componente estocástico do futebol. A capacidade preditiva alcançada mostrou-se insuficiente para viabilizar recomendações confiáveis, inviabilizando a segunda etapa. O trabalho contribui para estabelecer limites realistas para predição de desempenho esportivo e identificar desafios metodológicos (distribuições *zero-inflated*, tamanho amostral, estocasticidade) que precisam ser superados para viabilizar sistemas de recomendação robustos no futebol.

Palavras-chave: Predição de Desempenho. *Machine Learning*. Futebol. Análise de Dados Esportivos.

Abstract

This work aimed to develop a football player recommendation system based on performance prediction, using data from the Brazilian Championship Serie A (2022-2024). The proposal envisioned two stages: predictive modeling of individual metrics and generation of recommendations. This work focused on the first stage, implementing and evaluating three machine learning models (Random Forest, XGBoost, and MLP) with records of all championship players through the FootyStats platform. The exploratory analysis revealed significant challenges: zero-inflated distributions in offensive metrics (35–40% with zero goals), reduced sample size for goalkeepers (21 in training) and moderate temporal correlation (r=0.584 for goals), establishing a theoretical maximum R^2 of 0.341. The results showed R^2=0.181 for goals (MLP), representing 53% of the theoretical limit, and negative R^2 for most other metrics. The analysis demonstrated that 40–60% of the variation is fundamentally unpredictable due to the stochastic component of football. The predictive capacity achieved proved insufficient to enable reliable recommendations, making the second stage unfeasible. The work contributes by establishing realistic limits for sports performance prediction and identifying methodological challenges (zero-inflated distributions, sample size, stochasticity) that need to be overcome to enable robust recommendation systems in football.

Keywords: Performance Prediction. Machine Learning. Football. Sports Data Analysis.

Lista de Figuras

Sumário

1 Introdução

O futebol é muito mais que um esporte no Brasil, é parte fundamental da identidade cultural do país. Presente em praças, ruas e estádios, ele mobiliza milhões de torcedores e movimenta uma das indústrias mais influentes da economia nacional. O Campeonato Brasileiro, em especial, reúne alguns dos principais clubes do continente e é palco de grandes disputas, rivalidades históricas e revelação de talentos que ganham projeção internacional. Essa paixão coletiva pelo futebol faz com que cada jogo seja acompanhado de perto, com análises, debates e expectativas que vão muito além dos 90 minutos em campo (HADAMA, 2015).

Nos últimos anos, o futebol mundial passou por uma profunda transformação, impulsionada pelo avanço da tecnologia e pela crescente disponibilidade de dados. Se antes o desempenho de um time ou jogador era avaliado principalmente pela observação e pela experiência de treinadores e analistas, hoje métricas estatísticas, modelos preditivos e sistemas automatizados desempenham papel central na compreensão e no aprimoramento do jogo. Essa tendência também se reflete no cenário brasileiro, onde clubes e profissionais têm investido cada vez mais em análise de dados como ferramenta estratégica para obter vantagem competitiva (Doentes por Futebol, 2022).

Entre as temporadas de 2022 e 2024 do Campeonato Brasileiro da Série A, a ampla coleta de informações sobre jogos, equipes e atletas abriu espaço para novas possibilidades de estudo. É nesse contexto que se insere este trabalho, cujo objetivo foi desenvolver um sistema de recomendação de jogadores baseado na predição do desempenho individual. A proposta original previa duas etapas: (1) modelagem preditiva de métricas individuais futuras e (2) geração de recomendações fundamentadas nessas predições. Este trabalho concentrou-se na implementação e avaliação da primeira etapa, investigando a viabilidade da predição como base para sistemas de recomendação no futebol.

Para contextualizar, a Confederação Brasileira de Futebol (CBF) (2025) disponibiliza informações detalhadas sobre todas as partidas do Campeonato Brasileiro, incluindo resultados, escalações, cartões, gols e outros eventos ocorridos em campo, servindo como referência importante para estudos acadêmicos e análises de desempenho no futebol nacional. Neste trabalho, especificamente, foram utilizados dados da plataforma *FootyStats*, que agrega estatísticas individuais de jogadores em formato estruturado adequado à modelagem preditiva.

Ao integrar conceitos de análise de desempenho esportivo, estatística e ciência de dados, este trabalho investigou como técnicas de *machine learning* podem ser aplicadas à predição de desempenho no futebol. Os resultados evidenciaram tanto potencialidades quanto limitações fundamentais, contribuindo para o estabelecimento de expectativas realistas sobre previsibilidade de desempenho individual em esportes caracterizados por alta estocasticidade.

No restante deste capítulo, a Seção 1.1 apresenta as justificativas que fundamentam a escolha do tema e evidenciam sua relevância acadêmica, prática e pessoal. Em seguida, na Seção 1.2, são descritos o objetivo geral e os objetivos específicos que orientaram o desenvolvimento do trabalho. Por fim, a Seção 1.3 detalha a organização da monografia, apresentando a estrutura dos capítulos que compõem este estudo.

1.1 Justificativa

A análise de dados tem se consolidado como uma das áreas mais promissoras da Ciência da Computação, com aplicações em praticamente todos os setores da sociedade. No esporte, especialmente no futebol, essa abordagem tem se mostrado cada vez mais relevante para embasar decisões estratégicas, melhorar o desempenho das equipes e potencializar a identificação de talentos. Ao trazer ferramentas computacionais para um campo tradicionalmente guiado pela intuição e experiência, abre-se um leque de possibilidades para a inovação e o avanço do conhecimento aplicado ao esporte.

No contexto brasileiro, o futebol não é apenas um entretenimento, mas também um importante componente cultural e econômico, movimentando grandes investimentos e impactando diretamente a vida de milhões de pessoas. Desenvolver soluções tecnológicas que contribuam para a qualificação das análises e para a tomada de decisões nesse cenário representa não apenas um avanço técnico, mas também um impacto social significativo.

Este projeto explora conceitos técnicos avançados como modelagem preditiva, análise de séries temporais e métricas de desempenho esportivo, investigando as possibilidades e limitações da aplicação de *machine learning* em dados de futebol. Os resultados contribuem tanto para o avanço acadêmico na área de análise esportiva quanto para o estabelecimento de expectativas realistas quanto à viabilidade de sistemas de recomendação baseados em predição.

Além da relevância prática, este trabalho contribui academicamente ao estabelecer limites realistas para a predição do desempenho esportivo. Enquanto muitos estudos reportam apenas casos de sucesso, esta pesquisa documenta de forma transparente tanto potencialidades quanto limitações fundamentais, fornecendo *insights* valiosos sobre os desafios inerentes à predição no futebol: distribuições *zero-inflated*, tamanho amostral limitado e alta estocasticidade. Esses achados são essenciais para orientar trabalhos futuros e evitar expectativas irrealistas quanto à aplicabilidade de técnicas de *machine learning* em esportes.

1.2 Objetivos

Este trabalho teve como objetivo geral desenvolver e avaliar modelos de *machine learning* para a predição do desempenho individual de jogadores de futebol, como etapa fundamental para viabilizar sistemas de recomendação, utilizando dados das temporadas de 2022 a 2024 do

Campeonato Brasileiro Série A. A proposta original previa duas etapas: (1) modelagem preditiva de métricas individuais futuras e (2) geração de recomendações fundamentadas nessas predições. Este trabalho concentrou-se na implementação e na avaliação da primeira etapa, com foco na viabilidade técnica da predição como base para sistemas de recomendação.

Para atingir este objetivo geral, foram estabelecidos os seguintes objetivos específicos:

- Coletar, organizar e preparar dados de desempenho de jogadores da Série A do Campeonato Brasileiro referentes às temporadas de 2022 a 2024, provenientes da plataforma *FootyStats*;
- Realizar análise exploratória dos dados, identificando padrões distribucionais, correlações entre variáveis e persistência temporal de desempenho através de análise de pares temporais;
- Implementar e comparar três modelos de *machine learning* (*Random Forest*, *XGBoost* e Redes Neurais MLP) para predição de 17 métricas individuais (11 para jogadores de linha, 6 para goleiros);
- Avaliar o desempenho preditivo dos modelos através de métricas quantitativas (MAE, RMSE, R^2) no conjunto de validação, comparando com limites teóricos estabelecidos pela análise de pares temporais;
- Identificar limitações e desafios metodológicos específicos da predição de desempenho esportivo, incluindo distribuições *zero-inflated*, tamanho amostral reduzido e estocasticidade inerente ao futebol.

1.3 Organização do Trabalho

Esta monografia está organizada nos seguintes capítulos:

Capítulo 1: Introdução. Apresenta o contexto do trabalho, destacando a relevância do futebol no Brasil, a importância crescente da análise de dados no esporte e o problema de pesquisa que motiva o estudo de predição de desempenho. Também são expostos a justificativa, os objetivos e a organização da monografia.

Capítulo 2: Revisão Bibliográfica. Reúne o referencial teórico necessário para embasar o estudo, abordando conceitos de ciência de dados, *machine learning*, métricas de desempenho esportivo e trabalhos relacionados que exploram o uso da análise de dados no futebol.

Capítulo 3: Desenvolvimento. Descreve o processo metodológico adotado, incluindo a coleta e preparação dos dados, a seleção das métricas de análise, as técnicas de modelagem preditiva utilizadas (*Random Forest*, *XGBoost*, MLP) e os procedimentos de avaliação de desempenho.

Capítulo 4: Resultados. Dividido em duas partes: (1) Análise Exploratória das Variáveis Preditivas, caracterizando distribuições, correlações e persistência temporal para estabelecer limites de previsibilidade; e (2) Resultados da Modelagem Preditiva, apresentando o desempenho dos três modelos implementados através de métricas quantitativas, análise por métrica individual e comparação entre abordagens.

Capítulo 5: Considerações Finais. Resume as principais contribuições do trabalho, destacando os achados sobre os limites da previsibilidade do desempenho esportivo e os desafios metodológicos identificados. Discute as implicações dos resultados para a viabilidade de sistemas de recomendação baseados em predição, aponta limitações do estudo e sugere caminhos para trabalhos futuros.

2 Revisão Bibliográfica

Este capítulo está organizado em duas seções principais: a Seção 2.1 reúne a fundamentação teórica, abordando os conceitos e técnicas essenciais para o entendimento e desenvolvimento da solução proposta, já a Seção 2.2 apresenta estudos que exploram a aplicação de métodos computacionais na recomendação de jogadores e no apoio à tomada de decisões estratégicas no esporte.

2.1 Fundamentação Teórica

A recomendação de itens, serviços ou conteúdos é um campo consolidado da Inteligência Artificial e da Ciência de Dados, com aplicações que vão desde o comércio eletrônico até o esporte. Sistemas de recomendação têm como objetivo identificar padrões e preferências dos usuários, oferecendo sugestões personalizadas que aumentam a relevância da informação apresentada (RICCI; ROKACH; SHAPIRA, 2015).

De maneira geral, esses sistemas podem ser classificados em três categorias principais: baseados em conteúdo, que analisam as características dos itens e o histórico do usuário; baseados em filtragem colaborativa, que utilizam similaridade entre usuários ou itens para sugerir novas opções; e híbridos, que combinam as duas abordagens para superar limitações individuais (BURKE, 2002).

Na área esportiva, a recomendação de atletas pode apoiar decisões estratégicas, como contratações e substituições, auxiliando técnicos e analistas a identificarem jogadores com características semelhantes ou complementares às de interesse. Tais aplicações dependem de dados confiáveis, obtidos a partir de estatísticas de desempenho, histórico de partidas e métricas físicas e técnicas (LIU et al., 2016).

O avanço de técnicas de aprendizado de máquina e processamento de grandes volumes de dados tem ampliado as possibilidades desses sistemas, permitindo análises mais precisas e contextualizadas. Nesse contexto, a presente pesquisa busca integrar conceitos consolidados da área de sistemas de recomendação ao cenário esportivo, com foco no futebol brasileiro.

2.2 Trabalhos Relacionados

Romano (2023) apresenta um sistema completo para auxiliar no processo de *scouting* de jogadores de futebol. A abordagem proposta combina métodos de recuperação de informação com inteligência artificial generativa em um fluxo de duas etapas. Primeiramente, o sistema utiliza dados estatísticos da temporada 2022-2023 para encontrar os jogadores mais similares a

um perfil de referência, empregando o método da Similaridade de Cosseno, que se mostrou mais eficaz que a clusterização *K-Means* em testes comparativos.

Em seguida, essa lista de atletas similares é usada como base para o segundo componente do sistema, que utiliza o modelo de linguagem GPT-3.5 Turbo para gerar um relatório de *scouting* detalhado e contextualizado. Romano (2023) avaliou diferentes formas de instruir a IA (*prompting*), concluindo que fornecer dados estatísticos dos jogadores diretamente ao modelo resultou em relatórios de maior qualidade e fidelidade. Essa eficácia foi validada tanto por avaliação humana quanto por métricas automáticas como *ROUGE* e *BERTScore*, que indicaram alta similaridade semântica com relatórios de referência.

A principal contribuição do estudo é a demonstração de uma arquitetura funcional que une com sucesso a análise de similaridade e a geração de texto por IA para uma aplicação prática. Como limitação, o próprio autor destaca que o sistema se baseia em dados de uma única temporada, o que pode não refletir a carreira completa ou a consistência de um jogador.

Moya et al. (2025) conduziram revisão sistemática abrangendo 172 artigos publicados entre 2019 e 2024 sobre aplicação de *Machine Learning* (ML) em futebol profissional, identificando duas áreas principais de investigação: análise de desempenho de jogadores e times, e predição de resultados de partidas. O estudo revelou predominância de aprendizado supervisionado, *deep learning* e modelos híbridos que integram múltiplas técnicas de ML. Entre os algoritmos mais amplamente utilizados, destacam-se árvores de decisão, *Extreme Gradient Boosting* (XGBoost) e redes neurais artificiais, todos com foco na otimização de performance esportiva e predição de resultados. A revisão identificou desafios significativos na área, incluindo a limitada disponibilidade de *datasets* públicos devido a restrições de acesso e custo, o uso restrito de ferramentas avançadas de visualização, e a integração incipiente entre diferentes fontes de dados. Os autores concluem que, apesar do crescimento acelerado da área, a escassez de dados acessíveis e a dificuldade de reproduzir os estudos representam barreiras importantes ao avanço da pesquisa em análise esportiva baseada em ML.

Mills et al. (2024) propuseram um *framework* inovador para predição de resultados de partidas de futebol, utilizando algoritmos de ML e *deep learning*, aplicado inicialmente à *Dutch Eredivisie League* e, posteriormente, expandido para as ligas escocesa (*Scottish Premiership*) e belga (*Belgian Jupiler Pro League*). A metodologia incluiu pré-processamento de dados, engenharia de *features*, treinamento e teste de múltiplos modelos: Regressão Logística, XGBoost, *Random Forest*, SVM, *Naive Bayes*, Rede Neural *Feedforward* e Rede Neural Recorrente. Um diferencial do estudo foi a incorporação de *features* em tempo real, como resultados do intervalo e gols marcados até determinado momento da partida, em contraste com abordagens convencionais que utilizam apenas estatísticas finais. Os modelos foram avaliados através de métricas de acurácia, *recall*, precisão, *F1-score* e área sob a curva ROC. O principal achado foi que um modelo de votação (*voting model*) integrando *Random Forest* e XGBoost apresentou acurácia superior aos modelos individuais, reforçando a eficácia de abordagens *ensemble* quando combinadas com

engenharia de *features* específicas do domínio esportivo.

A aplicação de ML em análise esportiva enfrenta desafios particulares relacionados à natureza dos dados disponíveis e às características intrínsecas do domínio. Conforme destacado por Analytics Vidhya (2025), muitas equipes dispõem de dados históricos limitados, o que naturalmente aumenta o risco de *overfitting* dos modelos. O autor ressalta que sistemas de ML para esportes devem ser construídos considerando o contexto real do jogo e as práticas de treinamento, uma vez que o conhecimento especializado do domínio é fundamental para a interpretação adequada dos resultados. Adicionalmente, eventos imprevisíveis inerentes ao futebol, como lesões súbitas, decisões arbitrais controversas e fatores psicológicos, limitam a capacidade de generalização e a acurácia das predições. O estudo também aponta disparidades de recursos, observando que clubes menores frequentemente carecem tanto de orçamento quanto de *expertise* técnica para implementar soluções de ML em escala. Esses fatores evidenciam que a utilização eficaz de ML em esportes demanda não apenas competência técnica em ciência de dados, mas também julgamento criterioso e profundo entendimento do contexto esportivo específico.

Os trabalhos apresentados abordam diferentes aspectos da aplicação de ML no futebol, mas diferem substancialmente desse estudo quanto ao escopo e ao objetivo. Romano (2023) foca na similaridade entre jogadores e na geração de relatórios qualitativos com base em uma única temporada, enquanto este trabalho realiza predição quantitativa multivariada ao longo de três temporadas. Moya et al. (2025) oferece uma síntese bibliográfica sem aplicação empírica, ao passo que este estudo implementa e compara três modelos de ML calibrados para o futebol brasileiro. Mills et al. (2024) prediz resultados binários de partidas em ligas europeias, ao contrário da predição baseada em onze métricas individuais de desempenho aqui proposta. Por fim, Analytics Vidhya (2025) discute conceitualmente os desafios de dados limitados, enquanto este trabalho os enfrenta metodologicamente por meio de validação temporal e regularização.

Esse trabalho diferencia-se, portanto, por combinar predição multivariada de desempenho individual, aplicação ao contexto sub-representado do Campeonato Brasileiro Série A, tratamento empírico de limitações de dados e análise granular separada para jogadores de linha e goleiros.

3 Desenvolvimento

O presente capítulo detalha a metodologia adotada para preparação e análise dos dados, desde a obtenção dos dados brutos até a construção de uma base consolidada adequada para modelagem preditiva. São apresentadas as etapas de coleta, organização, análise exploratória, tratamento de dados e seleção de variáveis, destacando decisões adotadas para simplificar e otimizar o *dataset* de jogadores.

O trabalho iniciou-se com a obtenção de dados brutos de jogadores provenientes da plataforma FootyStats (2025), referentes às temporadas de 2022 a 2024. Os *datasets* foram adquiridos mediante assinatura da plataforma e incluem métricas detalhadas de desempenho individual, abrangendo aspectos técnicos, táticos e estatísticos.

Este capítulo apresenta os procedimentos adotados para a preparação dos dados e modelagem preditiva. A Seção 3.1 descreve o processo de coleta dos dados e a estruturação inicial do conjunto analisado. Em seguida, a Seção 3.2 detalha a Análise Exploratória de Dados e as etapas de preparação necessárias para garantir a qualidade das informações utilizadas. A Seção 3.3 detalha os modelos implementados.

3.1 Coleta e Estrutura Inicial dos Dados

Os dados utilizados neste trabalho foram obtidos da plataforma FootyStats (2025), que disponibiliza informações detalhadas de jogos e jogadores de diferentes ligas. A escolha dessa fonte justifica-se pela abrangência das estatísticas oferecidas e pela possibilidade de utilização acadêmica, apresentando dados consistentes para análises exploratórias e modelagem preditiva.

O *dataset* de jogadores foi organizado em grupos temáticos de variáveis, facilitando a análise e manipulação: Identificadores, Métricas Gerais e de Participação, Goleiros, Ataque, Meio-Campo, Defesa, Dribles e Movimentação, Pressão e Recuperação, Cartões e Disciplina, e Avaliação e Ranking.

3.2 Análise Exploratória de Dados e Preparação

Esta seção descreve o processo de análise exploratória e de preparação dos dados, com o objetivo de estruturar e depurar o conjunto bruto coletado, eliminando variáveis irrelevantes ou problemáticas e produzindo um subconjunto consistente de 66 variáveis para análises subsequentes.

A etapa de tratamento dos dados de jogadores seguiu um fluxo sistemático:

- **Análise de Qualidade:** verificação dos tipos de dados, contagem de valores nulos e cálculo do percentual de ausência por variável;
- **Estatísticas Descritivas:** cálculo de média, mediana, desvio padrão, valores mínimo e máximo para variáveis numéricas;
- **Cardinalidade:** contagem e análise dos valores únicos em variáveis categóricas.

Durante a análise, identificou-se que aproximadamente 33,5% dos jogadores não possuíam métricas avançadas preenchidas, reflexo da segmentação natural entre atletas com participação relevante e aqueles com poucos minutos jogados. Também foram detectadas variáveis com variância zero, métricas redundantes e valores atípicos indicativos de erros de cálculo ou de códigos especiais (ex.: -1 para "não ranqueado").

O processo de refinamento envolveu:

- Remoção de variáveis com 100% de valores nulos, variância zero ou redundância evidente;
- Padronização das métricas, mantendo preferencialmente estatísticas totais para volume e por 90 minutos para eficiência;
- Eliminação de duplicidades e colunas menos informativas;
- Exclusão de variáveis descritivas sem valor analítico direto para a modelagem, como informações pessoais ou administrativas (ex.: *date of birth*, *squad number*).

Esse processo reduziu significativamente a dimensionalidade do *dataset*, que passou de 277 variáveis iniciais para 66, garantindo um conjunto mais enxuto, consistente e adequado às análises subsequentes.

3.2.1 Seleção de Variáveis para Modelagem Preditiva

Das 66 variáveis resultantes do processo de preparação descrito na seção anterior, uma seleção adicional foi realizada para identificar o subconjunto mais relevante à modelagem preditiva. Testes preliminares indicaram desempenho inferior ao utilizar todas as 66 variáveis disponíveis, devido a *overfitting* e multicolinearidade, motivando a eliminação de redundâncias entre versões totais e normalizadas por 90 minutos da mesma métrica (ex: *goals_overall* vs *goals_per_90_overall*), de variáveis derivadas recalculáveis a partir de seus componentes, e de métricas menos informativas dado o tamanho limitado do *dataset* (290 amostras para jogadores de linha após filtro de 450 minutos; 21 para goleiros).

Variáveis-alvo (targets). Foram selecionadas 17 métricas como variáveis-alvo para predição, distribuídas entre jogadores de linha (11 métricas) e goleiros (6 métricas). Para jogadores de linha, as métricas ofensivas incluem gols, assistências, pênaltis convertidos, pênaltis perdidos e

bolas na trave. As métricas defensivas compreendem desarmes, interceptações e cortes, enquanto as disciplinares envolvem cartões amarelos e vermelhos. A métrica coletiva selecionada foi jogos sem sofrer gols (*clean sheets*). Para goleiros, as métricas incluem *clean sheets*, gols sofridos, defesas realizadas, chutes enfrentados, defesas dentro da área e socos/rebotes.

Variáveis de entrada (features). Como features preditivas, utilizaram-se seis variáveis contextuais (idade, minutos jogados, partidas disputadas, jogos como titular, entradas como substituto e saídas durante a partida), as mesmas métricas da temporada anterior listadas como targets (11 para linha, 6 para goleiros), o clube atual codificado via one-hot encoding (24 variáveis binárias), e a posição para jogadores de linha codificada via label encoding (1 variável ordinal). Assim, o conjunto final de features consiste em 42 variáveis para jogadores de linha (6 contextuais + 11 métricas + 24 clubes + 1 posição) e 36 para goleiros (6 contextuais + 6 métricas + 24 clubes).

Justificativa da redução. A escolha desse subconjunto específico baseou-se em testes preliminares, que mostraram desempenho inferior ao usar todas as 66 variáveis disponíveis, provavelmente devido a *overfitting* e multicolinearidade. Ao focar em métricas fundamentais e não-redundantes, buscou-se melhorar o poder preditivo e facilitar a interpretação dos resultados. As demais variáveis foram preservadas no *dataset* para análises futuras, mas não foram utilizadas na modelagem preditiva apresentada neste estudo.

3.3 Modelagem Preditiva

A análise exploratória apresentada nas seções anteriores permitiu compreender o comportamento histórico de jogadores ao longo das temporadas de 2022 a 2024. Para avançar na direção de sistemas de recomendação, torna-se necessário não apenas analisar o desempenho passado, mas também estimar o desempenho futuro dos atletas através de técnicas de aprendizado de máquina. Este capítulo descreve a preparação dos dados para modelagem temporal, as arquiteturas dos modelos implementados e os critérios de avaliação adotados.

3.3.1 Motivação e Abordagem

A hipótese principal é que características contextuais (idade, minutos jogados, clube) combinadas ao histórico estatístico recente possam fornecer indícios sobre a trajetória futura do atleta. A abordagem adotada enquadra-se como um problema de regressão multivariada supervisionada, no qual múltiplas métricas de desempenho são previstas simultaneamente. A formulação temporal é definida da seguinte forma: dado o conjunto de estatísticas de um jogador no ano N, o objetivo é prever suas estatísticas no ano N+1, desde que o jogador tenha permanecido ativo em ambas as temporadas.

Devido às diferenças fundamentais entre as métricas de jogadores de linha e goleiros, optou-se por treinar dois modelos independentes: um especializado em jogadores de linha (defensores, meio-campistas e atacantes) e outro dedicado aos goleiros. Essa separação permite

que cada modelo capture padrões específicos de cada grupo, evitando que o aprendizado seja prejudicado pela mistura de estatísticas incompatíveis.

Para avaliar qual abordagem de aprendizado de máquina é mais adequada ao problema, foram implementados três modelos distintos, cada um representando uma família algorítmica diferente: Random Forest (ensemble baseado em bagging), XGBoost (ensemble baseado em boosting) e Redes Neurais Artificiais (MLP). A comparação entre essas abordagens permite identificar se a complexidade adicional de redes neurais justifica-se em datasets de tamanho limitado ou se métodos baseados em árvores produzem resultados comparáveis ou superiores.

3.3.2 Preparação dos Dados para Modelagem Temporal

A construção de um dataset adequado para aprendizado supervisionado temporal requer a criação de pares de exemplos (X, y), onde X representa o estado do jogador no ano base e y representa seu desempenho no ano seguinte. Para cada jogador presente no dataset consolidado, foram identificadas sequências de temporadas consecutivas. Apenas pares de anos consecutivos foram considerados válidos (transições $2022 \rightarrow 2023$ ou $2023 \rightarrow 2024$), uma vez que gaps temporais introduziriam ruído na modelagem.

A divisão entre conjuntos de treino e validação seguiu uma estratégia temporal, respeitando a natureza sequencial dos dados: o conjunto de treino incluiu todos os jogadores com ano base em 2022 prevendo desempenho em 2023, enquanto o conjunto de validação incluiu jogadores com ano base em 2023 prevendo 2024. Essa abordagem simula o cenário real de uso do modelo, no qual dados futuros não estão disponíveis durante o treinamento, evitando assim o vazamento de informação (data leakage) que ocorreria com uma divisão aleatória.

As features utilizadas como entrada dos modelos consistem em três grupos principais. Primeiro, seis variáveis contextuais: idade, minutos jogados, partidas disputadas, jogos como titular, entradas como substituto e saídas durante a partida. Segundo, as métricas da temporada atual (ano base): 11 para jogadores de linha (gols, assistências, pênaltis convertidos e perdidos, bolas na trave, clean sheets, desarmes, interceptações, cortes, cartões amarelos e vermelhos) e 6 para goleiros (clean sheets, gols sofridos, defesas, chutes enfrentados, defesas dentro da área e socos/rebotes). Terceiro, informações categóricas: clube atual codificado via one-hot encoding (24 variáveis binárias) e posição para jogadores de linha codificada via label encoding (Defender $\rightarrow$ 0, Forward $\rightarrow$ 1, Midfielder $\rightarrow$ 2). Isso resulta em 42 features para jogadores de linha (6 contextuais + 11 métricas + 24 clubes + 1 posição) e 36 features para goleiros (6 contextuais + 6 métricas + 24 clubes).

Os targets correspondem às mesmas métricas estatísticas utilizadas como features, porém referentes à temporada seguinte, permitindo que o modelo aprenda a mapear o estado atual para o estado futuro do jogador. Uma etapa crítica no pré-processamento foi a normalização tanto das features quanto dos targets, utilizando o StandardScaler do scikit-learn. Esse procedimento

transforma as variáveis para que tenham média aproximadamente zero e desvio padrão unitário, acelerando a convergência do treinamento e melhorando a estabilidade numérica. A normalização foi aplicada separadamente: features numéricas ajustadas com base no conjunto de treino e aplicadas ao conjunto de validação, e targets também normalizados com base no treino, permitindo a posterior desnormalização para a interpretação dos resultados na escala original.

Após todas as transformações, os conjuntos resultantes apresentaram as seguintes dimensões. Para jogadores de linha: X_{treino} com $(290, 42)$ e y_{treino} com $(290, 11)$, $X_{validacao}$ com $(281, 42)$ e $y_{validacao}$ com $(281, 11)$. Para goleiros: X_{treino} com $(21, 36)$ e y_{treino} com $(21, 6)$, $X_{validacao}$ com $(23, 36)$ e $y_{validacao}$ com $(23, 6)$. A discrepância no número de amostras entre jogadores de linha e goleiros reflete a realidade da distribuição de atletas na liga, sendo um fator que influenciará diretamente os resultados da modelagem, conforme será discutido no Capítulo 4.

3.3.3 Modelos Implementados

Esta seção descreve as três abordagens de aprendizado de máquina implementadas, começando pelos métodos baseados em árvores de decisão e concluindo com redes neurais artificiais.

Random Forest. *Random Forest* é um método de aprendizado *ensemble* que constrói múltiplas árvores de decisão durante o treinamento e produz, como saída, a média das previsões individuais (para regressão). Proposto por (BREIMAN, 2001), o algoritmo combina dois conceitos fundamentais: *bagging* (*bootstrap aggregating*) e seleção aleatória de features.

O treinamento de um *Random Forest* segue os seguintes passos: primeiro, para cada árvore t a ser construída, é criado um conjunto de treino D_t através de amostragem com reposição (*bootstrap*) do *dataset* original D, introduzindo diversidade entre as árvores. Segundo, cada árvore é construída recursivamente, dividindo os dados em cada nó com base na *feature* que melhor reduz a variância dos targets. Crucialmente, em cada divisão, apenas um subconjunto aleatório de m *features* é considerado (onde $m = \sqrt{p}$ para p *features* totais). Terceiro, na predição, cada árvore produz uma estimativa independente, e o resultado final é a média aritmética de todas as previsões:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} \hat{y}_t \tag{3.1}$$

onde T é o número de árvores e $\hat{y}_t$ é a predição da árvore t.

O *Random Forest* foi escolhido por apresentar características particularmente adequadas ao contexto deste trabalho: robustez com poucos dados, não requerendo grandes volumes de dados para generalizar adequadamente, reduzindo significativamente o risco de *overfitting*; independência de escala, pois as árvores de decisão são invariantes a transformações monotônicas das *features*, eliminando a necessidade de normalização; e interpretabilidade, permitindo calcular

a importância de *features* (*feature importance*), facilitando a compreensão de quais fatores são mais preditivos.

Os hiperparâmetros do *Random Forest* foram definidos através de uma abordagem conservadora, priorizando valores que reduzem *overfitting* dado o tamanho amostral limitado (290 para linha, 21 para goleiros). Não foi realizada busca exaustiva via *grid search* ou validação cruzada devido ao foco exploratório do trabalho; ao invés disso, utilizaram-se valores recomendados pela literatura, por Breiman (2001), e pela documentação da biblioteca `scikit-learn`, com ajustes manuais para adequação ao contexto. A Tabela 3.1 apresenta os hiperparâmetros configurados e suas respectivas justificativas.

Tabela 3.1 – Hiperparâmetros do *Random Forest*.

Hiperparâmetro	**Valor**	**Justificativa**
`n_estimators`	100	Valor padrão; mais árvores reduzem variância sem aumentar risco de *overfitting*
`max_depth`	10	Limitação conservadora para evitar árvores excessivamente complexas; profundidade ilimitada seria arriscada com 21 amostras (goleiros)
`min_samples_split`	10	Requer no mínimo 10 amostras para permitir split, evitando divisões em ruído estatístico
`min_samples_leaf`	5	Garante que folhas tenham pelo menos 5 amostras, aumentando representatividade das predições
`max_features`	sqrt	$\sqrt{p}$ features por split (recomendação padrão para regressão; promove decorrelação entre árvores)
`random_state`	42	Garante reprodutibilidade dos resultados

Esses valores representam um equilíbrio entre a capacidade de modelagem e a prevenção de *overfitting*. Em trabalhos futuros, técnicas de otimização automática de hiperparâmetros poderiam ser exploradas, mas requerem um volume maior de dados para uma validação confiável.

XGBoost (Extreme Gradient Boosting). *XGBoost* (CHEN; GUESTRIN, 2016) é uma implementação otimizada e escalável de *gradient boosting*, uma técnica *ensemble* que constrói modelos sequencialmente, onde cada novo modelo tenta corrigir os erros do anterior.

Gradient boosting constrói um modelo aditivo da forma:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{3.2}$$

onde cada f_k é uma árvore de decisão adicionada sequencialmente para minimizar uma função de perda L. Cada nova árvore foca nos exemplos em que o modelo atual erra mais, permitindo correção progressiva dos erros.

O *XGBoost* se distingue de implementações tradicionais de *gradient boosting* por incorporar otimizações algorítmicas que melhoram desempenho e reduzem *overfitting*: (1) regularização

explícita através de termos L1 e L2 na função objetivo, penalizando complexidade das árvores; (2) tratamento automático de valores ausentes, aprendendo a melhor direção para dados faltantes durante treinamento; (3) poda de árvores de trás para frente (*backward pruning*), removendo *splits* que não contribuem significativamente; (4) paralelização e otimização computacional, com estruturas de dados eficientes e suporte a GPUs; e (5) *subsampling* de exemplos e *features*, similar ao *Random Forest*, aumentando diversidade entre árvores.

A Tabela 3.2 apresenta as principais diferenças conceituais entre os dois métodos *ensemble*.

Tabela 3.2 – Comparação entre *Random Forest* e *XGBoost*.

Aspecto	***Random Forest***	***XGBoost***
Estratégia	*Bagging*: árvores treinadas independentemente em paralelo	*Boosting*: árvores treinadas sequencialmente, cada uma corrigindo erros da anterior
Dependência	Árvores independentes entre si	Árvores dependentes (sequenciais)
Profundidade	Árvores profundas (`max_depth=10`) para capturar complexidade individualmente	Árvores rasas (`max_depth=3–5`) compensadas por múltiplas iterações
Regularização	Robustez natural via agregação; não requer regularização explícita	Requer regularização L1/L2 para evitar *overfitting*
Velocidade	Paralelizável; treinamento rápido	Sequencial; ligeiramente mais lento, mas com otimizações

Os hiperparâmetros do *XGBoost* foram configurados de forma conservadora, priorizando a prevenção de *overfitting* dado o tamanho amostral reduzido (especialmente para goleiros, com apenas 21 amostras no treino). Não foi realizada busca exaustiva via *grid search*; utilizaram-se valores recomendados pela documentação oficial e pela literatura, com ajustes manuais para adequação ao contexto. A estratégia privilegia árvores rasas (`max_depth=3`) compensadas por múltiplas iterações, seguindo a recomendação de que algoritmos de *boosting* funcionam melhor com árvores pequenas (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Adicionalmente, o *shrinkage* (taxa de aprendizado) reduz a influência de cada árvore individual (FRIEDMAN, 2001).

A Tabela 3.3 apresenta os hiperparâmetros configurados e suas justificativas. Esses valores representam um equilíbrio entre a capacidade de modelagem e a prevenção de *overfitting*. O tempo de treinamento foi aproximadamente 0,25 segundos por modelo, ligeiramente superior ao *Random Forest* devido à natureza sequencial do algoritmo, mas ainda aceitável para o tamanho do *dataset*.

Redes Neurais Artificiais (MLP). A arquitetura escolhida para o terceiro modelo é um *Multi-Layer Perceptron* (MLP), também conhecido como rede neural totalmente conectada (*feedforward neural network*). Essa escolha justifica-se pela capacidade de modelar relações não-lineares complexas e pela facilidade de implementação em *frameworks* modernos como *PyTorch*.

Tabela 3.3 – Hiperparâmetros do *XGBoost*.

Hiperparâmetro	**Valor**	**Justificativa**
`n_estimators`	100	Número de árvores sequenciais; balanço entre poder preditivo e tempo de treinamento
`max_depth`	3	Árvores rasas típicas de *boosting*; complexidade é alcançada pela sequência, não por árvores individuais profundas
`learning_rate`	0,1	Taxa de aprendizado (padrão); controla quanto cada árvore contribui para o modelo final
`subsample`	0,8	Usa 80% das amostras por árvore; introduz aleatoriedade e reduz *overfitting*
`colsample_bytree`	0,8	Usa 80% das *features* por árvore; estratégia similar ao *Random Forest* para aumentar diversidade
`reg_alpha`	0,1	Regularização L1 (*lasso*); promove esparsidade nos pesos das folhas
`reg_lambda`	1,0	Regularização L2 (*ridge*, valor padrão); penaliza pesos grandes, reduzindo sensibilidade a ruído
`random_state`	42	Garante reprodutibilidade dos resultados

A arquitetura da rede neural foi definida seguindo princípios conservadores adequados ao tamanho amostral limitado (290 para linha, 21 para goleiros). Redes muito profundas ou largas apresentariam alto risco de *overfitting*, uma vez que o número de parâmetros treináveis cresceria rapidamente em relação ao número de amostras disponíveis. A rede é composta por três tipos de camadas principais: camada de entrada que recebe as *features* processadas (42 para jogadores de linha, 36 para goleiros), correspondendo às métricas do ano anterior mais o *one-hot encoding* do clube; duas camadas ocultas densas sequenciais com 32 e 16 neurônios, respectivamente; e camada de saída linear (sem função de ativação) que produz previsões para as 11 métricas (jogadores de linha) ou 6 métricas (goleiros).

A escolha de duas camadas ocultas representa um compromisso entre capacidade expressiva e prevenção de *overfitting*, considerando o tamanho amostral limitado (290 para linha, 21 para goleiros). A redução progressiva de neurônios ($42 \rightarrow 32 \rightarrow 16 \rightarrow 11/6$) cria um afunilamento que força a rede a aprender representações cada vez mais compactas, atuando como regularização implícita. Ambas as camadas utilizam a função de ativação ReLU (*Rectified Linear Unit*), padrão para camadas ocultas por sua simplicidade computacional e capacidade de mitigar o problema de gradientes desvanecentes. A ausência de função de ativação na saída é apropriada para problemas de regressão, permitindo que a rede produza valores em toda a reta real.

A arquitetura resultante possui aproximadamente 2.091 parâmetros treináveis para jogadores de linha e 1.750 para goleiros, resultando em razões amostras/parâmetros de 0,14 (linha) e 0,012 (goleiros), valores que justificam a necessidade de regularização agressiva via *dropout* e *early stopping*. Arquiteturas mais profundas (3–4 camadas ocultas) ou mais largas (64–128 neurônios) foram deliberadamente evitadas. Com apenas 290 amostras para linha e, criticamente,

apenas 21 para goleiros, redes complexas tenderiam a memorizar os dados de treino ao invés de aprender padrões generalizáveis. A arquitetura escolhida representa um balanço entre capacidade expressiva (suficiente para capturar não-linearidades) e simplicidade (necessária para generalização com poucos dados).

A função de ativação ReLU (*Rectified Linear Unit*), definida como $f(x) = \max(0, x)$, foi escolhida para as camadas ocultas por evitar o problema de *vanishing gradient* e ser computacionalmente eficiente. Foi aplicada a técnica de *dropout* com probabilidade de 0,3 (30%) após cada camada oculta, desativando aleatoriamente 30% dos neurônios em cada iteração para forçar a rede a aprender representações mais robustas. A função de perda utilizada foi o Erro Quadrático Médio (MSE, do inglês *Mean Squared Error*):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.3}$$

O otimizador foi Adam (*Adaptive Moment Estimation*) com taxas de aprendizado de 0,001 (linha) e 0,0005 (goleiros), e regularização L2 (*weight decay*) com coeficiente $\lambda = 0,0001$. Foi implementado `ReduceLROnPlateau` que reduz a taxa pela metade se não houver melhora por 15 épocas. O *early stopping* foi configurado com paciência de 25 épocas, interrompendo o treinamento se o *validation loss* não melhorar. O *batch size* foi de 16 amostras (linha) e 8 amostras (goleiros). A arquitetura possui 2.091 parâmetros treináveis (linha) e 1.750 parâmetros (goleiros).

O treinamento seguiu o algoritmo de retropropagação (*backpropagation*): embaralhamento dos dados, iteração em mini-lotes, propagação direta (*forward pass*), cálculo da perda, propagação reversa (*backward pass*) com gradientes calculados usando a regra da cadeia, *gradient clipping* (norma máxima 1,0), atualização dos pesos, validação, ajuste da taxa de aprendizado, salvamento do melhor modelo, e verificação de *early stopping*.

Durante a propagação direta, cada camada aplica a Equação 3.4.

$$h^{(l)} = \text{ReLU}(W^{(l)}h^{(l-1)} + b^{(l)}) \tag{3.4}$$

Na propagação reversa, o gradiente é calculado:

$$\frac{\partial \mathcal{L}}{\partial W^{(l)}} = \frac{\partial \mathcal{L}}{\partial h^{(l)}} \cdot \frac{\partial h^{(l)}}{\partial W^{(l)}} \tag{3.5}$$

Os modelos foram treinados por até 100 épocas. Para jogadores de linha, o treinamento foi interrompido na época 26 (melhor modelo na época 1). Para goleiros, o treinamento completo de 100 épocas foi executado.

Comparação conceitual dos modelos. Os três modelos implementados representam paradigmas algorítmicos distintos, cada um com características próprias que influenciam di-

retamente seu comportamento em cenários de dados limitados como o presente trabalho. A Tabela 3.4 sintetiza as principais diferenças conceituais entre as abordagens, destacando aspectos relevantes para a escolha e interpretação dos resultados.

Tabela 3.4 – Características Comparativas dos Modelos

Característica	**MLP**	**RF**	**XGBoost**
Tipo	Rede Neural	Bagging	Boosting
Interpretabilidade	Baixa	Alta	Moderada
Requer Normalização	Sim	Não	Não
Robustez c/ Poucos Dados	Baixa	Alta	Moderada
Tempo Treinamento	Longo	Rápido	Rápido

Observa-se que o *Random Forest* destaca-se pela robustez com poucos dados e alta interpretabilidade, características desejáveis no contexto deste trabalho (290 amostras para linha, 21 para goleiros). Sua estratégia de *bagging*, treinamento paralelo de árvores independente naturalmente reduz variância sem exigir normalização ou ajustes complexos de hiperparâmetros. O *XGBoost*, por sua vez, oferece posicionamento intermediário: mais sofisticado que o *Random Forest* através de *boosting* sequencial, mas ainda baseado em árvores, mantendo boa robustez e dispensa de normalização. Já o MLP apresenta menor robustez com dados escassos e maior tempo de treinamento, mas possui, em tese, maior capacidade de modelar relações não-lineares complexas, justificando sua inclusão como contraponto às abordagens baseadas em árvores. Essas diferenças fundamentais permitem avaliar se a complexidade adicional do MLP compensa suas desvantagens em cenários de dados limitados ou se métodos mais simples alcançam desempenho comparável ou superior.

3.3.4 Métricas de Avaliação

Foram adotadas três métricas complementares para avaliar o desempenho dos modelos, cada uma capturando aspectos distintos da qualidade preditiva.

O Erro Absoluto Médio (*Mean Absolute Error* — MAE) quantifica o erro médio das previsões em termos absolutos, calculado pela Equação 3.6.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.6}$$

onde n é o número total de amostras, y_i é o valor real da i-ésima observação, $\hat{y}_i$ é o valor previsto pelo modelo para a i-ésima observação, e $|y_i - \hat{y}_i|$ é o valor absoluto do erro individual. O MAE fornece uma medida direta e intuitiva do erro típico nas unidades originais da variável. Por exemplo, MAE = 1,7 gols significa que, em média, as previsões do modelo erram por 1,7 gols. Por tratar todos os erros de forma linear (sem elevar ao quadrado), o MAE é menos sensível a *outliers* que o RMSE, representando melhor o erro "típico" ou "mediano".

A Raiz do Erro Quadrático Médio (*Root Mean Squared Error* — RMSE) é calculada pela Equação 3.7.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.7}$$

onde os termos são os mesmos definidos para o MAE, com a diferença de que os erros são elevados ao quadrado antes da média, e posteriormente aplica-se a raiz quadrada. Ao elevar os erros ao quadrado, o RMSE penaliza desproporcionalmente erros grandes em relação a erros pequenos. Por exemplo, um erro de 4 gols contribui 16 vezes mais para o RMSE que um erro de 1 gol, enquanto no MAE contribui apenas 4 vezes mais. Matematicamente, RMSE $\geq$ MAE sempre, com igualdade apenas quando todos os erros têm a mesma magnitude. A diferença entre RMSE e MAE indica a presença de erros grandes: quanto maior a diferença, mais heterogêneos são os erros do modelo.

O Coeficiente de Determinação (R^2) mede a proporção da variação total dos dados que é explicada pelo modelo, calculado pela Equação 3.8.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SS_{res}}{SS_{tot}} \tag{3.8}$$

onde $\bar{y}$ é a média dos valores reais, $SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ é a soma dos quadrados dos resíduos (erros do modelo), e $SS_{tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ é a soma total dos quadrados (variação total dos dados). O R^2 quantifica o ganho do modelo em relação a uma predição trivial pela média. Seus valores podem ser interpretados como: $R^2 = 1,0$ indica modelo perfeito, captura 100% da variação (todas as previsões exatas); $R^2 = 0,5$ significa que o modelo explica 50% da variação total, reduzindo o erro em relação à média pela metade; $R^2 = 0,0$ indica que o modelo equivale a prever sempre a média, não há ganho preditivo; $R^2 < 0$ significa modelo pior que prever a média simples, indica falha grave na modelagem ou inadequação do modelo ao problema. Diferentemente de MAE e RMSE, o R^2 é adimensional e independe da escala da variável, permitindo comparação direta entre métricas de magnitudes diferentes (ex: gols vs cartões). É particularmente útil para avaliar se o modelo captura padrões reais ou apenas memoriza ruído.

Como os modelos de Redes Neurais foram treinados com dados normalizados (média 0, desvio padrão 1), suas previsões em escala normalizada requerem transformação inversa para interpretação nas unidades originais:

$$\hat{y}_{\text{original}} = \hat{y}_{\text{normalizado}} \times \sigma_y + \mu_y \tag{3.9}$$

onde μ_y e σ_y são, respectivamente, a média e o desvio padrão dos *targets* do conjunto de treino, armazenados durante a normalização inicial. Os modelos baseados em árvores (*Random Forest* e

XGBoost) não requerem essa etapa, pois são invariantes a transformações monotônicas.

As métricas foram calculadas em dois níveis complementares: global (consolidado), considerando todas as previsões de todas as métricas simultaneamente, fornecendo visão geral do desempenho do modelo; e por métrica individual, calculadas separadamente para cada variável prevista (gols, assistências, desarmes, etc.), permitindo identificar quais métricas são mais ou menos previsíveis e onde cada modelo se destaca ou falha. Essa abordagem dual permite tanto comparações globais entre modelos quanto análises granulares de desempenho por tipo de estatística, essenciais para compreender as limitações específicas de cada abordagem.

4 Resultados

Este capítulo apresenta os resultados obtidos neste trabalho, organizados em duas etapas complementares que refletem a natureza progressiva da investigação: (1) Compreensão dos dados antes da modelagem, na Seção 4.1; (2) Avaliação do desempenho preditivo, na Seção 4.2.

A primeira etapa (**Análise Exploratória das Variáveis Preditivas**) caracteriza o comportamento das 17 métricas selecionadas como *targets*: 11 para jogadores de linha e 6 para goleiros. Esta análise preliminar é essencial para estabelecer expectativas realistas sobre a capacidade preditiva dos modelos, uma vez que a previsibilidade de uma variável é limitada, em última instância, pela sua própria persistência temporal. A análise abrange distribuições, correlações e, crucialmente, a persistência de desempenho entre anos consecutivos.

A segunda etapa (**Resultado da Modelagem Preditiva**) apresenta os resultados dos três modelos implementados, *Random Forest*, *XGBoost* e MLP aplicados à predição das métricas caracterizadas na etapa anterior. Os resultados são apresentados de forma progressiva: (1) desempenho global agregado, fornecendo visão geral da capacidade preditiva; (2) análise detalhada por métrica individual, identificando quais variáveis são melhor previstas; (3) indicadores de *overfitting*, avaliando a capacidade de generalização; e (4) síntese comparativa entre os três modelos, destacando *trade-offs* de desempenho, tempo computacional e robustez.

A estrutura do capítulo reflete, portanto, um processo investigativo em dois momentos: primeiro, compreender o que é previsível (análise exploratória); depois, avaliar quão bem os modelos conseguem prever (avaliação de desempenho). Essa abordagem permite não apenas reportar métricas de desempenho, mas contextualizá-las adequadamente: valores de R^2 que seriam considerados insatisfatórios em outros domínios podem representar resultados razoáveis em predição de desempenho esportivo, dado o alto componente estocástico inerente ao futebol. A análise exploratória prévia fornece, assim, a base para interpretação informada dos resultados de modelagem.

4.1 Análise Exploratória das Variáveis Preditivas

Antes de apresentar os resultados da modelagem preditiva, esta seção caracteriza o comportamento das 17 métricas utilizadas como *targets* (11 para linha, 6 para goleiros), identificando padrões distribucionais e estimando viabilidade de predição através da análise de persistência temporal.

4.1.1 Impacto do Filtro de Participação Mínima

O filtro de 450 minutos (aproximadamente 5 partidas completas) foi aplicado para garantir robustez estatística. A Figura 4.1 evidencia o impacto dessa decisão.

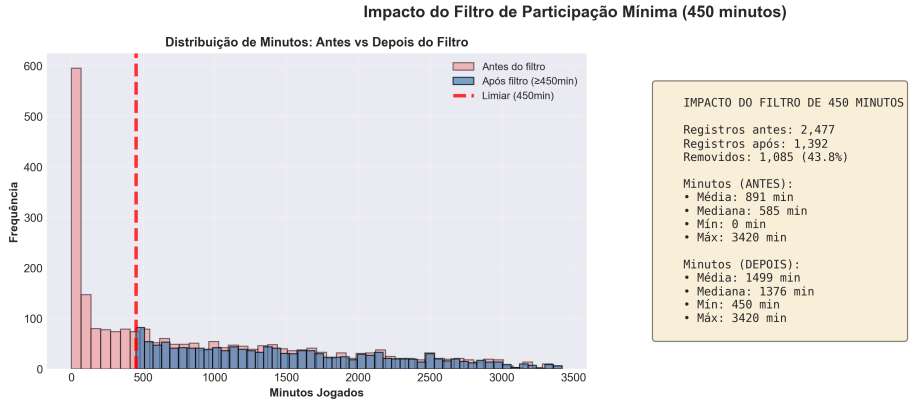

Figura 4.1 – Impacto do filtro de participação mínima (450 minutos) sobre a distribuição de minutos jogados e estatísticas descritivas do dataset.

Fonte: Elaborado pelo autor.

O histograma revela concentração extrema no dataset original: 24% dos jogadores (600 de 2.477) disputaram menos de 200 minutos, introduzindo instabilidade estatística. O filtro removeu 43,8% dos registros, mas transformou a qualidade amostral: média de minutos aumentou 68% (891 $\rightarrow$ 1.499 min) e mediana mais que dobrou (585 $\rightarrow$ 1.376 min). Esse *trade-off*, perda quantitativa versus ganho qualitativo foi considerado essencial para modelagem robusta.

4.1.2 Distribuição das Métricas

As Figuras 4.2 e 4.3 apresentam as distribuições das métricas de linha e goleiros.

Jogadores de Linha: Gols (média 1,97) e assistências (média 1,36) apresentam distribuições *zero-inflated*, termo que descreve distribuições com concentração anormalmente alta em zero devido à especialização funcional: 35–40% dos jogadores (defensores e volantes) não marcam gols, enquanto atacantes apresentam cauda longa até 27 gols. Esse padrão desafia modelos de regressão tradicionais que assumem distribuição normal: a variância não é constante (jogadores com 0 gols têm variância nula, artilheiros têm alta variabilidade) e a relação entre preditores e *target* tende a ser não-linear. Métricas defensivas (desarmes, interceptações, cortes) exibem distribuições mais equilibradas com boa dispersão. Pênaltis e cartões vermelhos são eventos raros (média 0,06–0,23) com limitada informação discriminante.

Goleiros: Clean sheets (média 6,5) apresenta distribuição bimodal, caracterizada por dois picos distintos ao invés de um único pico central: o primeiro pico em 0–2 jogos corresponde a goleiros reservas ou de times fracos, enquanto o segundo pico em 6–8 jogos representa titulares de times médios/fortes. Essa bimodalidade evidencia clara separação entre perfis de participação. Defesas totais (média 47,4) e chutes enfrentados (média 65,1) exibem amplitude extrema (10–140),

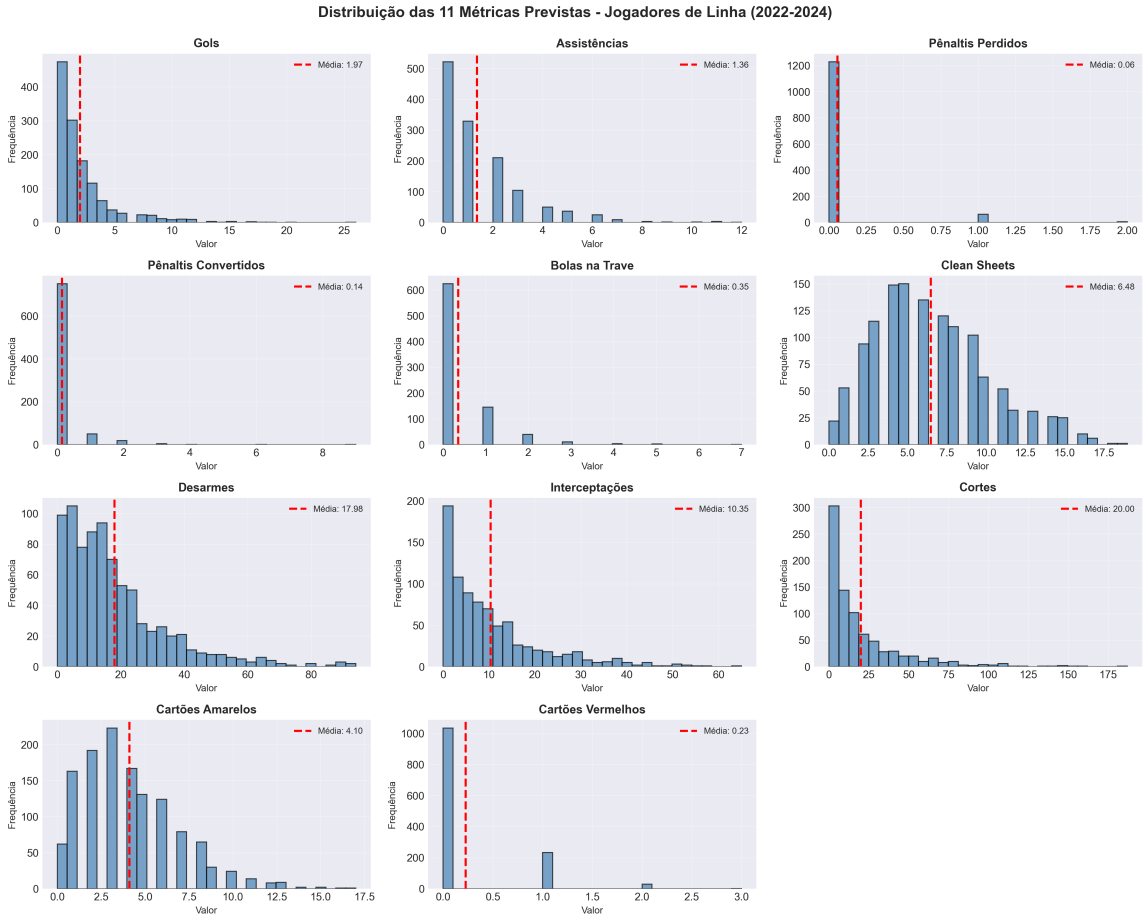

Figura 4.2 – Distribuição das 11 métricas previstas para jogadores de linha (n=1.297, 2022–2024).

Fonte: Elaborado pelo autor.

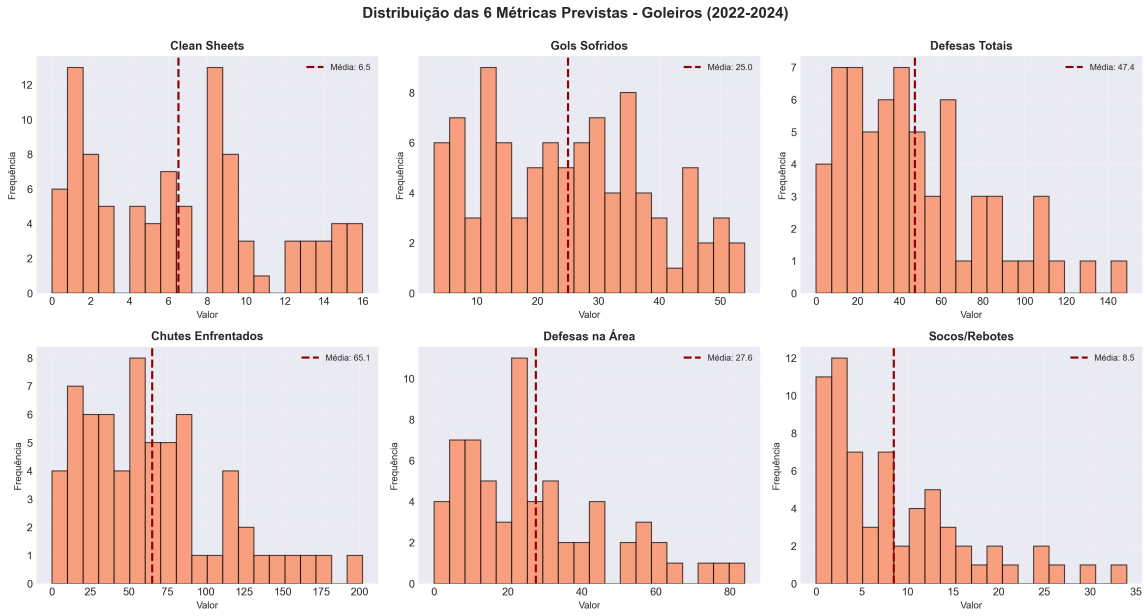

Figura 4.3 – Distribuição das 6 métricas previstas para goleiros (n=95, 2022–2024).

Fonte: Elaborado pelo autor.

refletindo contexto tático ao invés de habilidade individual: goleiros de times fracos enfrentam maior volume. Taxa média defesas/chutes de 72,8% indica aproveitamento relativamente estável. O reduzido tamanho amostral (n=95, apenas 21 no treino) representa limitação crítica para modelagem de goleiros.

4.1.3 Estrutura de Correlação

A Figura 4.4 revela estrutura clara de especialização posicional.

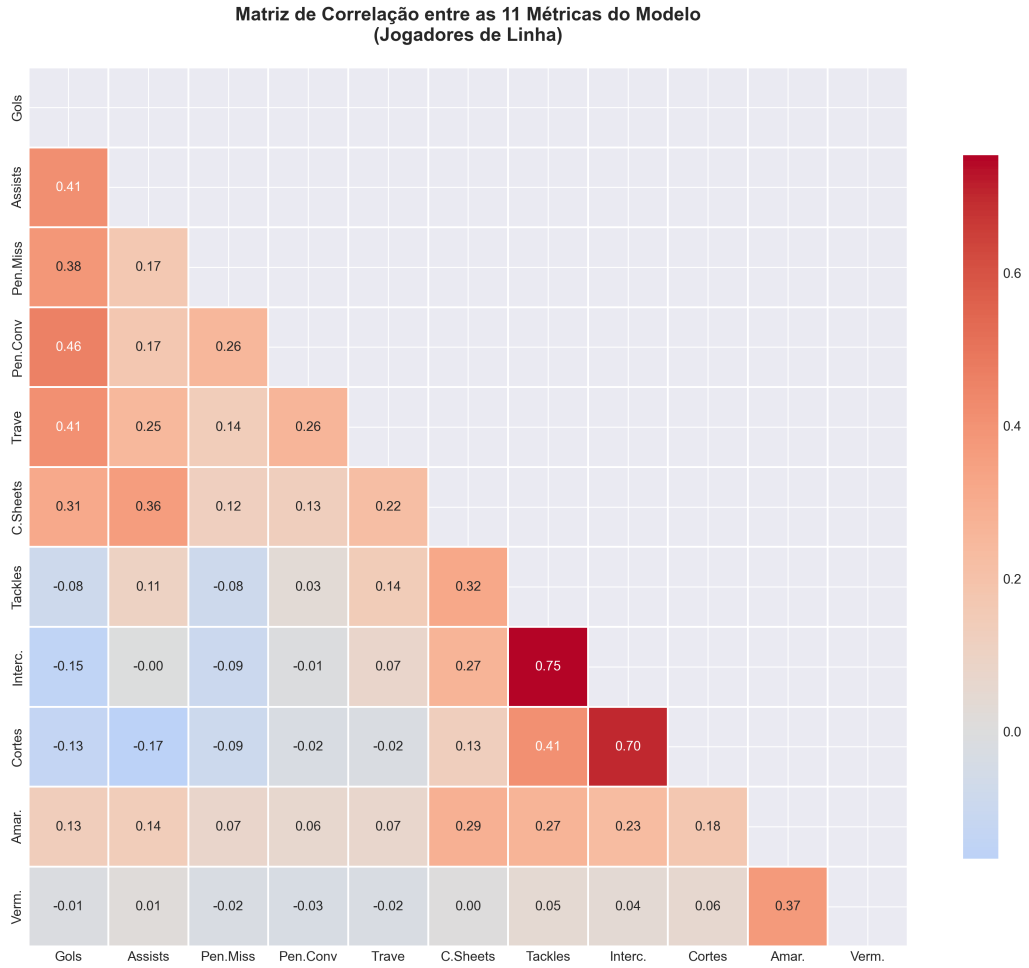

Figura 4.4 – Matriz de correlação entre as 11 métricas do modelo (jogadores de linha).

Fonte: Elaborado pelo autor.

Cluster defensivo: Desarmes e interceptações apresentam correlação forte (r=0,75), caracterizando perfil de "recuperador de bola". Interceptações-cortes também correlacionam fortemente (r=0,70).

Cluster ofensivo: Gols e assistências exibem correlação moderada (r=0,41), indicando perfil ofensivo completo mas com especialização significativa.

Especialização posicional: Correlações próximas a zero entre métricas ofensivas e defensivas (gols-desarmes: r=−0,08) confirmam funções mutuamente exclusivas. Clean sheets

apresenta correlações fracas com ações individuais (r=0,14 com desarmes), evidenciando natureza coletiva.

4.1.4 Persistência Temporal: Viabilidade de Predição

A Figura 4.5 apresenta a análise crítica de persistência de desempenho entre anos consecutivos.

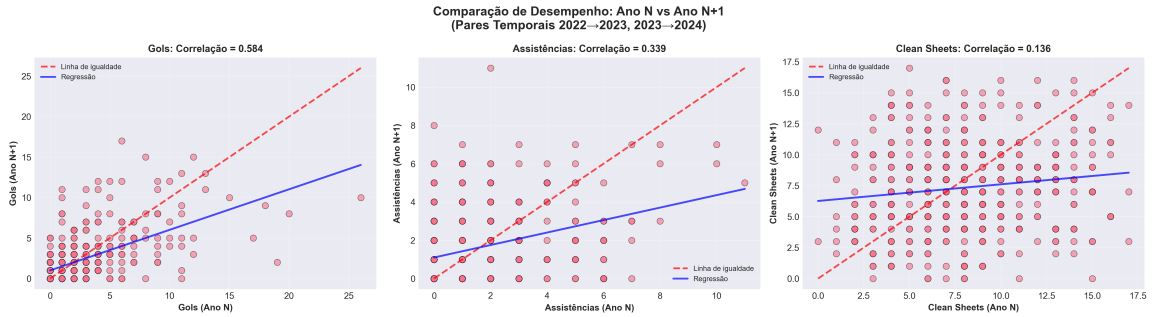

Figura 4.5 – Correlação entre desempenho em anos consecutivos (2022→2023, 2023→2024). Linha vermelha: igualdade (y=x). Linha azul: regressão linear.

Fonte: Elaborado pelo autor.

A análise de pares temporais quantifica a persistência de desempenho: jogadores com bom desempenho em um ano tendem a repetir no ano seguinte? A correlação de Pearson (r) mede a força dessa relação, enquanto o R² máximo teórico (r^2) indica o limite superior de previsibilidade, a porcentagem máxima de variação futura explicável apenas pelo desempenho passado, assumindo modelo perfeito. Por exemplo, r=0,584 para gols implica $R^2_{máx}$=0,341, significando que no máximo 34% da variação em gols futuros é previsível; os 66% restantes dependem de fatores não capturados (lesões, mudanças táticas, estocásticidade do jogo).

Gols (r=0,584): Melhor persistência temporal, com 58% de covariância compartilhada. A linha de regressão abaixo da linha de igualdade evidencia regressão à média: desempenhos excepcionais (25 gols) tendem a valores conservadores (14 gols). R² máximo teórico de 0,341.

Assistências (r=0,339): Persistência fraca-moderada (34% de covariância), com nuvem dispersa. A natureza colaborativa, dependência da finalização do companheiro limita previsibilidade. R² máximo teórico de 0,115.

Clean Sheets (r=0,136): Persistência muito fraca (14% de covariância), com dispersão extrema (jogadores com 5 clean sheets variaram de 2 a 17). Confirma natureza coletiva, não individual. R² máximo teórico de 0,018.

A Tabela 4.1 sintetiza os resultados.

4.1.5 Considerações Finais do Estudo das Variáveis

A análise exploratória estabelece expectativas realistas para modelagem:

Tabela 4.1 – Correlações temporais e expectativas de previsibilidade.

Métrica	r	R^2 máx. teórico	Natureza
Gols	0,584	0,341	Individual
Assistências	0,339	0,115	Semi-individual
Clean Sheets	0,136	0,018	Coletiva

1. **Filtro justificado:** Eliminação de 43,8% de registros instáveis, com ganho de 68% na média de participação;

2. **Distribuições heterogêneas:** Gols e assistências são *zero-inflated* (desafiam modelos lineares), enquanto métricas defensivas são mais equilibradas;

3. **Especialização confirmada:** Correlações próximas a zero entre métricas ofensivas-defensivas evidenciam funções mutuamente exclusivas;

4. **Limites de previsibilidade:** Gols (r=0,584) são moderadamente previsíveis, assistências (r=0,339) fracamente previsíveis, clean sheets (r=0,136) essencialmente imprevisíveis individualmente;

5. **Expectativa realista:** R^2 entre 0,10–0,35 para gols é razoável dado componente estocástico do futebol. Aproximadamente 40–60% da variação permanece imprevisível.

Com essas características estabelecidas, as próximas subseções apresentam os resultados dos três modelos implementados.

4.2 Resultados da Modelagem Preditiva

Após caracterizar o comportamento das variáveis preditivas e estabelecer limites de previsibilidade através da análise de pares temporais (Seção 4.1), esta seção apresenta os resultados obtidos com os três modelos implementados: *Random Forest*, *XGBoost* e Redes Neurais Artificiais (MLP).

A análise exploratória revelou desafios significativos que contextualizam os resultados apresentados a seguir: (1) distribuições *zero-inflated* em métricas ofensivas, com 35–40% dos jogadores apresentando valor zero, desafiando modelos de regressão tradicionais; (2) persistência temporal moderada para gols (r=0,584, $R^2_{máx}$=0,341) e fraca para assistências (r=0,339, $R^2_{máx}$=0,115), estabelecendo teto teórico de previsibilidade; (3) clean sheets com correlação muito fraca entre anos (r=0,136), confirmando natureza coletiva; e (4) tamanho amostral reduzido para goleiros (apenas 21 no treino), limitando capacidade de generalização.

Os resultados são organizados em quatro perspectivas: (1) Desempenho geral (Subseção 4.2.1), reportando métricas consolidadas no conjunto de validação; (2) Desempenho por

métrica individual (Subseção 4.2.2), identificando quais variáveis são melhor previstas e comparando com limites teóricos; (3) Comportamento de treinamento (Subseção 4.2.3), avaliando indicadores de *overfitting*; e (4) Comparação entre modelos (Subseção 4.2.4), sintetizando *trade-offs* de desempenho e robustez.

A interpretação dos resultados considera os limites identificados na análise exploratória: valores de R^2 modestos podem representar desempenho razoável dado o componente estocástico do futebol, onde aproximadamente 40–60% da variação permanece imprevisível mesmo em cenário teórico ideal.

4.2.1 Desempenho Geral dos Modelos

As Tabelas 4.2 e 4.3 apresentam as métricas consolidadas de desempenho dos três modelos no conjunto de validação. Os valores reportados correspondem às médias calculadas sobre todas as métricas previstas simultaneamente.

Tabela 4.2 – Métricas Globais de Desempenho – Jogadores de Linha (Validação).

Modelo	**MAE**	**RMSE**	**R^2**
Random Forest	3,06	6,95	$-0,253$
XGBoost	3,21	6,99	$-0,587$
MLP	3,04	6,94	$-0,240$

Tabela 4.3 – Métricas Globais de Desempenho – Goleiros (Validação).

Modelo	**MAE**	**RMSE**	**R^2**
Random Forest	16,31	25,15	$-0,992$
XGBoost	15,77	24,89	$-0,936$
MLP	16,91	25,56	$-1,164$

Os três modelos apresentaram valores de R^2 negativos na validação, indicando desempenho inferior à predição pela média simples. Esse resultado reflete a dificuldade intrínseca de predição de métricas esportivas, caracterizadas por alto componente estocástico e distribuições *zero-inflated* (como discutido na análise exploratória). Para jogadores de linha, os três modelos apresentaram MAE muito similares, variando entre 3,04 (MLP) e 3,21 (XGBoost), diferença de apenas 0,17 gols. Para goleiros, o MAE variou entre 15,77 (XGBoost) e 16,91 (MLP). Os erros para goleiros são 5–6 vezes maiores que para linha, refletindo tanto a maior magnitude das métricas de goleiros (gols sofridos, defesas totais) quanto o menor tamanho amostral (21 no treino vs. 290).

Em termos de R^2 global, o MLP apresentou melhor desempenho para jogadores de linha ($-0,240$), seguido pelo Random Forest ($-0,253$) e XGBoost ($-0,587$). Para goleiros, o Random Forest foi o melhor ($-0,992$), seguido pelo XGBoost ($-0,936$) e MLP ($-1,164$). A divergência nos valores de R^2 com variação de até 0,347 pontos para linha, indica que, embora os erros médios sejam similares, os modelos diferem na capacidade de capturar variabilidade das métricas.

Conforme estabelecido na análise de pares temporais (Seção 4.1), o R^2 máximo teoricamente alcançável para gols é 0,341, considerando apenas persistência temporal. Os valores negativos observados sugerem que os modelos não conseguiram capturar adequadamente os padrões preditivos, possivelmente devido à complexidade das distribuições *zero-inflated*, ao tamanho amostral limitado (especialmente para goleiros) e à alta estocásticidade inerente ao futebol. A análise por métrica individual (Subseção 4.2.2) identificará quais variáveis específicas apresentam melhor previsibilidade.

4.2.2 Desempenho por Métrica Individual

As Tabelas 4.4 e 4.5 apresentam os resultados detalhados para cada métrica prevista no conjunto de validação.

Tabela 4.4 – Desempenho por Métrica – Jogadores de Linha (Validação).

	MLP			**Random Forest**			**XGBoost**		
Métrica	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
goals_overall	1,70	2,35	0,181	1,80	2,37	0,168	1,91	2,40	0,149
assists_overall	1,19	1,55	0,101	1,23	1,58	0,071	1,77	2,05	$-0,570$
penalty_misses	0,12	0,26	0,082	0,11	0,26	0,068	0,43	0,46	$-1,902$
pen_scored_total	0,09	0,29	$-0,026$	0,06	0,29	$-0,050$	0,06	0,29	$-0,050$
hit_woodwork	0,19	0,50	$-0,174$	0,18	0,49	$-0,151$	0,18	0,49	$-0,151$
clean_sheets	2,71	3,29	0,123	2,76	3,38	0,073	3,09	3,82	$-0,187$
tackles_total	9,62	12,50	$-1,450$	9,60	12,49	$-1,445$	9,60	12,49	$-1,445$
interceptions	4,64	6,66	$-0,939$	4,64	6,66	$-0,941$	4,64	6,66	$-0,941$
clearances	10,67	17,39	$-0,603$	10,67	17,39	$-0,603$	10,67	17,39	$-0,603$
yellow_cards	2,22	2,79	0,057	2,27	2,85	0,020	2,42	3,10	$-0,167$
red_cards	0,34	0,46	0,004	0,34	0,46	0,003	0,53	0,58	$-0,589$

Tabela 4.5 – Desempenho por Métrica – Goleiros (Validação).

	MLP			**Random Forest**			**XGBoost**		
Métrica	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
clean_sheets	4,82	6,25	$-0,636$	3,88	4,98	$-0,040$	3,45	4,91	$-0,007$
conceded	16,02	18,86	$-0,412$	13,36	15,52	0,044	10,54	12,79	0,350
saves_total	26,40	33,51	$-1,636$	26,39	33,50	$-1,635$	26,39	33,50	$-1,635$
shots_faced	35,65	45,08	$-1,665$	35,70	45,13	$-1,671$	35,70	45,13	$-1,671$
inside_box_saves	14,58	18,44	$-1,670$	14,57	18,42	$-1,667$	14,57	18,42	$-1,667$
punches_total	3,98	5,65	$-0,963$	4,00	5,68	$-0,984$	4,00	5,68	$-0,984$

Para jogadores de linha, cinco métricas apresentaram R^2 positivo em pelo menos um modelo: goals_overall com R^2 máximo de 0,181 (MLP) e MAE de 1,70 gols, clean_sheets_overall com R^2 máximo de 0,123 (MLP), assists_overall com R^2 máximo de 0,101 (MLP), penalty_misses com R^2 máximo de 0,082 (MLP), e yellow_cards_overall com R^2 máximo de 0,057 (MLP). Para goleiros, apenas uma métrica apresentou R^2 positivo: conceded_overall com R^2 de 0,350 (XGBoost) e 0,044 (Random Forest). As demais métricas (tackles, interceptions, clearances para

linha; saves, shots_faced, inside_box_saves, punches para goleiros) apresentaram R² fortemente negativo em todos os modelos.

As cinco métricas com R² positivo (goals, clean_sheets, assists, penalty_misses, yellow_cards) correspondem a eventos cumulativos diretos ou ações individuais mensuráveis. Em contraste, as métricas com R² fortemente negativo são predominantemente ações defensivas: tackles ($R^2 \approx -1,45$), interceptions ($R^2 \approx -0,94$) e clearances ($R^2 \approx -0,60$). Nota-se que os três modelos apresentaram valores de R² praticamente idênticos para essas métricas defensivas, sugerindo que a dificuldade de predição não está relacionada à escolha do algoritmo.

Para a métrica goals_overall, o MLP obteve MAE de 1,70 gols e R² de 0,181, seguido pelo Random Forest (MAE 1,80, R² 0,168) e XGBoost (MAE 1,91, R² 0,149). A diferença de erro médio entre o melhor (MLP) e o pior (XGBoost) foi de apenas 0,21 gols, enquanto a diferença de R² foi de 0,032 pontos. Para goleiros, apenas a métrica conceded_overall apresentou R² positivo, exclusivamente nos modelos XGBoost (0,350) e Random Forest (0,044). Todas as métricas baseadas em volume de ações (saves_total, shots_faced, inside_box_saves) apresentaram R² fortemente negativo, com valores entre $-1,635$ e $-1,671$, indicando que todos os modelos falharam de forma similar nessas previsões.

4.2.3 Comportamento de Treinamento

As Figuras 4.6 e 4.7 apresentam a evolução do *loss* durante o treinamento do modelo MLP.

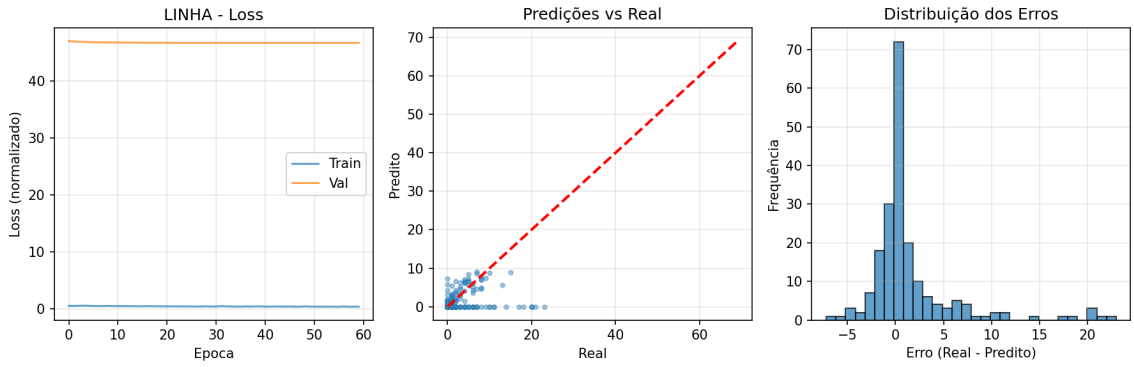

Figura 4.6 – Evolução do Loss durante Treinamento – Jogadores de Linha (MLP).

Para jogadores de linha, o *early stopping* foi acionado na época 26, mas o melhor modelo (menor *validation loss*) foi identificado na época 1. O *training loss* reduziu consistentemente até a época 10, estabilizando em 0,48. O *validation loss* iniciou em 46,17 e apresentou oscilações sem tendência clara de melhora. Para goleiros, o treinamento completou 100 épocas sem acionamento do *early stopping*. O *training loss* reduziu continuamente de 0,38 para 0,15. O *validation loss* reduziu modestamente de 586,87 para 581,91 (redução de menos de 1%).

A Tabela 4.6 apresenta a diferença entre R² de treino e validação.

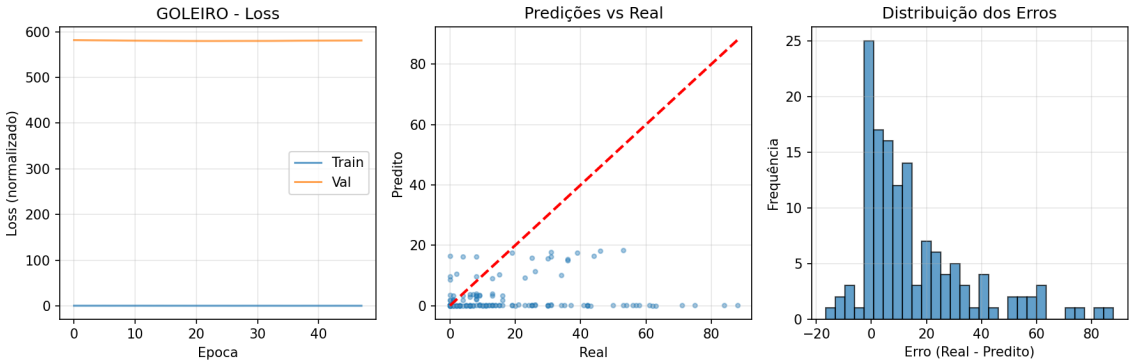

Figura 4.7 – Evolução do Loss durante Treinamento – Goleiros (MLP).

Tabela 4.6 – Diferença entre R² de Treino e Validação.

Modelo	Tipo	R² Treino	R² Val	Gap
Random Forest	Linha	0,636	$-0,253$	0,889
Random Forest	Goleiro	0,833	$-0,992$	1,825
XGBoost	Linha	0,988	$-0,587$	1,575
XGBoost	Goleiro	0,9999	$-0,936$	1,936
MLP	Linha	$-4,4M$	$-0,240$	N/A*
MLP	Goleiro	$-2,8M$	$-1,164$	N/A*

*MLP: R² treino negativo devido à normalização

O XGBoost apresenta o maior gap entre treino e validação, com diferenças de 1,575 (linha) e 1,936 (goleiro). Esse gap superior a 1,5 pontos indica que o modelo alcançou R² próximo a 1,0 no treino mas falhou na generalização. O Random Forest apresenta gap intermediário de 0,889 (linha) e 1,825 (goleiro), sendo o modelo mais equilibrado para jogadores de linha. O R² de treino negativo para o MLP é um artefato da normalização dos dados, não refletindo diretamente a qualidade do ajuste. A análise das curvas de *loss* (Figuras 4.6 e 4.7) fornece melhor indicação do comportamento: para linha, o *training loss* estabilizou em 0,48 enquanto o *validation loss* permaneceu em 46,17; para goleiros, os valores foram 0,15 (treino) e 581,91 (validação), uma diferença de aproximadamente 3.900 vezes.

Para goleiros, todos os modelos apresentaram gaps maiores que para linha. O Random Forest aumentou seu gap de 0,889 para 1,825 (+105%), e o XGBoost de 1,575 para 1,936 (+23%). Esse padrão reflete a maior dificuldade de generalização com apenas 21 amostras de treino.

4.2.4 Comparação entre Modelos

A Tabela 4.7 apresenta o desempenho dos três modelos para a métrica goals_overall.

O MLP apresentou melhor desempenho, seguido por *Random Forest* e *XGBoost*. A diferença entre MLP e *Random Forest* é de 0,013 pontos de R² (7,2% relativo). A Tabela 4.8 quantifica o número de métricas com R² positivo para cada modelo.

O *Random Forest* apresentou o maior número de métricas com R² positivo (6), seguido

Tabela 4.7 – Comparação de Desempenho – goals_overall (Validação).

Modelo	MAE	RMSE	R^2	Ranking
MLP	1,70	2,35	0,181	1º
Random Forest	1,80	2,37	0,168	2º
XGBoost	1,91	2,40	0,149	3º

Tabela 4.8 – Número de Métricas com R^2 Positivo por Modelo.

Modelo	Linha (de 11)	Goleiro (de 6)	Total (de 17)
MLP	5	0	5
Random Forest	5	1	6
XGBoost	1	1	2

pelo MLP (5) e *XGBoost* (2). A Tabela 4.9 resume as principais métricas de cada modelo.

Tabela 4.9 – Síntese Comparativa dos Modelos.

Característica	MLP	Random Forest	XGBoost
R^2 validação (linha)	$-0,240$	$-0,253$	$-0,587$
R^2 validação (goleiro)	$-1,164$	$-0,992$	$-0,936$
Melhor R^2 individual	0,181 (gols)	0,168 (gols)	0,350 (conceded)
Métricas R^2 positivo	5 de 17	6 de 17	2 de 17
Gap treino-val (linha)	N/A	0,889	1,575
Tempo médio treino	210s	0,15s	0,25s

Em termos de R^2 global, o MLP apresentou melhor desempenho na validação de linha ($-0,240$), enquanto o Random Forest foi melhor para goleiros ($-0,992$). O XGBoost apresentou o pior R^2 global em ambos os casos. O Random Forest teve o maior número de métricas individuais com R^2 positivo (6), enquanto o MLP teve o melhor R^2 individual absoluto (0,181 para gols). O XGBoost apresentou o maior gap entre treino e validação (1,575 para linha), enquanto o Random Forest foi o mais rápido em treinamento.

Os trade-offs identificados entre os modelos incluem: em termos de desempenho, o MLP obteve o melhor R^2 individual (0,181 para gols), com diferença de apenas 0,013 pontos em relação ao Random Forest (7,2% relativo); quanto a consistência versus pico de desempenho, o Random Forest apresentou 6 métricas com R^2 positivo contra 5 do MLP, indicando maior consistência em diferentes tipos de previsões, porém o MLP alcançou valores absolutos ligeiramente superiores nas métricas bem-sucedidas; em relação a generalização versus ajuste, o Random Forest apresentou o menor gap treino-validação para jogadores de linha (0,889), enquanto o XGBoost teve o maior (1,575), e apesar do ajuste quase perfeito no treino (R^2 = 0,988), o XGBoost teve o pior desempenho na validação; por tipo de jogador, para goleiros, o XGBoost foi o único modelo a obter R^2 positivo substancial em alguma métrica (conceded: 0,350), enquanto MLP e Random Forest falharam em todas exceto uma (Random Forest: conceded 0,044).

Considerando apenas as métricas com R^2 positivo, a média de R^2 foi de 0,109 para o MLP, 0,083 para o Random Forest e 0,250 para o XGBoost (considerando apenas suas 2 métricas

bem-sucedidas). Quando se considera todas as 17 métricas (incluindo as negativas), as médias caem para $-0,240$ (MLP), $-0,253$ (Random Forest) e $-0,587$ (XGBoost) para jogadores de linha.

5 Considerações Finais

Este capítulo apresenta as considerações finais do trabalho, destacando o trabalho realizado na preparação dos dados e implementação de modelos preditivos, os resultados alcançados, as limitações identificadas e as perspectivas para trabalhos futuros.

A Seção 5.1 resume os objetivos alcançados e os principais achados da modelagem preditiva, destacando a viabilidade parcial para métricas ofensivas e as limitações fundamentais para métricas defensivas. A Seção 5.2 discute as limitações do estudo em três dimensões: dados, metodologia e características inerentes ao problema. A Seção 5.3 apresenta direções para trabalhos futuros, incluindo expansão dos dados e desenvolvimento de sistemas de recomendação híbridos. Por fim, a Seção 5.4 sintetiza as principais lições aprendidas e destaca que, mais importante que alcançar métricas de desempenho artificialmente altas, é construir sistemas que comuniquem suas incertezas de forma honesta e científica.

5.1 Conclusão

Este trabalho teve como objetivo desenvolver e avaliar modelos de *machine learning* para predição de desempenho individual de jogadores de futebol, como etapa fundamental para viabilizar sistemas de recomendação. A proposta original previa duas etapas: (1) modelagem preditiva de métricas individuais futuras e (2) geração de recomendações fundamentadas nessas predições. Este trabalho concentrou-se na implementação e avaliação rigorosa da primeira etapa.

Objetivos alcançados: Os objetivos específicos foram plenamente cumpridos. Em relação à preparação de dados, foram coletados dados de 3 temporadas consecutivas (2022–2024) do Campeonato Brasileiro Série A, realizou-se organização e limpeza com redução de 277 para 66 variáveis organizadas em grupos temáticos e tratamento rigoroso de valores ausentes e inconsistências, e conduziu-se análise exploratória identificando padrões distribucionais, correlações entre variáveis e, crucialmente, persistência temporal de desempenho através de análise de pares temporais. Quanto à modelagem preditiva, criaram-se 571 pares temporais para jogadores de linha e 44 para goleiros com divisão temporal rigorosa (2022→2023 treino, 2023→2024 validação), implementaram-se três modelos distintos (*Random Forest*, *XGBoost* e Redes Neurais MLP), e realizou-se avaliação rigorosa com métricas consolidadas (MAE, RMSE, R²) globalmente e por métrica individual, incluindo análise de *overfitting* e generalização.

Principais achados: A experimentação com modelos de aprendizado de máquina revelou achados importantes sobre a previsibilidade do desempenho no futebol. Gols mostrou-se a métrica mais previsível, com MLP alcançando R^2=0,181 (18,1% da variação explicada) e erro médio de 1,7 gols. Esse resultado, embora modesto em termos absolutos, representa 53% do limite teórico

identificado na análise de pares temporais (R^2_{mx}=0,341), indicando que o modelo capturou parte significativa do sinal preditivo disponível nos dados. Outras métricas ofensivas apresentaram desempenho inferior: assistências com R²=0,101 (MLP) e clean sheets com R²=0,123 (MLP).

Métricas defensivas (desarmes, interceptações, cortes) e todas as métricas de goleiros apresentaram R² negativo, indicando desempenho inferior à predição pela média. As causas identificadas incluem dependência contextual (métricas defensivas dependem fortemente do adversário, sistema tático e função específica no esquema), insuficiência de dados (apenas 21 amostras de treino para goleiros, dramaticamente insuficiente para aprendizado de máquina), e natureza coletiva (análise de pares temporais revelou correlação r=0,136 para *clean sheets*, indicando que 98% da variação é imprevisível com R^2_{mx}=0,018, confirmando caráter coletivo).

Um achado crítico é que a limitação não é metodológica, mas inerente aos dados disponíveis. As três abordagens distintas (bagging, boosting, redes neurais) falharam similarmente, e múltiplas técnicas de regularização não eliminaram *overfitting*. A análise de pares temporais demonstrou que aproximadamente 40–60% da variação em métricas ofensivas permanece imprevisível mesmo em cenário teórico ideal, devido à alta estocásticidade do futebol. Na comparação entre modelos, o MLP apresentou melhor desempenho para gols (R²=0,181) mas é lento e com *overfitting* severo, o Random Forest mostrou-se mais equilibrado com 6 métricas com R² positivo e treinamento rápido, e o XGBoost apresentou *overfitting* crítico (R² treino=0,988, validação=$-0,587$). Para *datasets* de tamanho limitado, *Random Forest* apresentou o melhor equilíbrio entre desempenho, velocidade e robustez.

Contribuições. Este trabalho contribui para o campo de análise esportiva ao estabelecer limites realistas, documentando rigorosamente não apenas sucessos mas também limitações fundamentais da predição com dados básicos; comparar metodologias, avaliando três famílias algorítmicas sob condições idênticas e fornecendo conclusões robustas sobre adequação para contextos com dados limitados; e fornecer base estruturada, com *pipeline* de preparação e análise temporal reutilizável para trabalhos futuros.

5.2 Limitações do Estudo

As limitações deste estudo podem ser organizadas em três dimensões principais. Quanto aos dados, a janela temporal curta de apenas 3 temporadas limitou o número de pares temporais, as estatísticas básicas utilizadas não incluem métricas avançadas (*xG*, *xA*, contexto de oponentes), e a amostra pequena de 21 exemplos de treino para goleiros mostrou-se insuficiente para modelagem robusta. Em termos metodológicos, a premissa de estabilidade assume que padrões de desempenho são estáveis entre temporadas, e o uso de modelo único para jogadores de linha trata defensores, meio-campistas e atacantes conjuntamente sem distinção.

Existem também limitações inerentes ao próprio problema. A alta estocasticidade do futebol torna fatores como motivação, química de equipe e eventos aleatórios fundamentalmente

não capturáveis. A não-estacionariedade reflete que jogadores evoluem, envelhecem e mudam de contexto tático ao longo do tempo. Por fim, a interdependência com o time demonstra que o desempenho individual está fortemente acoplado à qualidade dos companheiros e ao sistema tático empregado.

5.3 Trabalhos Futuros

Diversas direções podem ser exploradas em trabalhos futuros. A expansão dos dados deve incluir coleta de 5–10 temporadas para aumentar pares temporais, inclusão da Série B para maior diversidade, integração de métricas avançadas (*xG*, *xA*, passes progressivos), e contextualização por força do oponente. O refinamento metodológico pode envolver desenvolvimento de modelos específicos por posição, criação de *ensemble* híbridos (MLP para métricas ofensivas, RF para outras), e aplicação de *transfer learning* utilizando dados de ligas europeias.

Abordagens alternativas incluem reformulação como problema de classificação (baixo/médio/alto desempenho) e utilização de *ranking* ao invés de predição de valores absolutos. O desenvolvimento de sistema de recomendação constitui o objetivo final: a base de dados preparada e os achados sobre limites de previsibilidade viabilizam o desenvolvimento futuro de um sistema híbrido que combine predição (para métricas ofensivas com desempenho razoável) com métodos de similaridade e *clustering* (para métricas defensivas e goleiros), reconhecendo explicitamente as limitações identificadas neste estudo.

5.4 Reflexões Finais

Este trabalho demonstrou que a predição de desempenho futuro no futebol utilizando dados estatísticos básicos enfrenta limitações fundamentais, mas não é totalmente inviável para métricas ofensivas. A análise de pares temporais estabeleceu que aproximadamente 40–60% da variação é fundamentalmente imprevisível, e os modelos implementados alcançaram desempenho próximo a esse limite teórico para gols.

A chave para avanços futuros está em: (1) reconhecer honestamente o que pode e não pode ser previsto; (2) focar esforços de ML nas métricas com sinal preditivo; (3) usar métodos complementares para aspectos imprevisíveis; e (4) investir em expansão de dados. Mais importante que alcançar R^2 artificialmente alto é construir sistemas que comuniquem suas incertezas de forma honesta e forneçam informações acionáveis, transparentes e fundamentadas cientificamente.

Referências

Analytics Vidhya. *How to Use Machine Learning in Sports Analytics?* 2025. Acesso em: 09 mar. 2026. Disponível em: <https://www.analyticsvidhya.com/blog/2025/07/machine-learning-in-sports/>.

BREIMAN, L. Random forests. *Machine Learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

BURKE, R. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, v. 12, n. 4, p. 331–370, 2002.

CARVALHO, M. de. *Construindo o saber: técnicas de metodologia científica*. [S.l.]: Papirus Editora, 1989. ISBN 9788530800710.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD*. [S.l.: s.n.], 2016. p. 785–794.

Confederação Brasileira de Futebol (CBF). *CBF - Confederação Brasileira de Futebol*. 2025. Disponível em: <https://www.cbf.com.br/>.

Doentes por Futebol. *Por que a análise de dados se tornou importante no futebol brasileiro?* 2022. Disponível em: <https://doentesporfutebol.com.br/por-que-a-analise-de-dados-se-tornou-importante-no-futebol-brasileiro/>.

FootyStats. *FootyStats: Football Statistics and Analytics Platform*. 2025. Disponível em: <https://footystats.org/>.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, Institute of Mathematical Statistics, v. 29, n. 5, p. 1189–1232, 2001.

HADAMA, P. D. Futebol brasileiro: cultura, história e sociedade. *Revista de Estudos Latino-Americanos*, 2015.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. ed. New York: Springer, 2009.

LIU, H.; CARLING, J. P.; FERNANDES, J. A. F.; LISBOA, T. P. Performance profiles of football players in the uefa champions league considering playing position and team quality. *International Journal of Performance Analysis in Sport*, v. 16, n. 2, p. 527–541, 2016.

MILLS, E. F. E. A.; DENG, Z.; ZHONG, Z.; LI, J. Data-driven prediction of soccer outcomes using enhanced machine and deep learning techniques. *Journal of Big Data*, Springer Open, v. 11, n. 1, p. 170, 2024. Disponível em: <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-024-01008-2>.

MOYA, D.; TIPANTUñA, C.; VILLA, G.; CALDERóN-HINOJOSA, X.; RIVADENEIRA, B.; ÁLVAREZ, R. Machine learning applied to professional football: Performance improvement and results prediction. *Machine Learning and Knowledge Extraction*, MDPI, v. 7, n. 3, p. 85, 2025. Disponível em: <https://www.mdpi.com/2504-4990/7/3/85>.

RAMPAZZO, L. *Metodologia científica*. [S.l.]: Edições Loyola, 2005. ISBN 9788515024988.

RICCI, F.; ROKACH, L.; SHAPIRA, B. *Recommender Systems Handbook*. 2nd. ed. [S.l.]: Springer, 2015.

ROMANO, A. *Player Scouting Recommendation System Using Similarity Methods and Generative AI*. Nápoles, Itália, 2023.