Universidade Federal da Paraíba
Centro de Ciências Sociais Aplicadas
Graduação em Ciência de Dados para Negócios

Estudo sobre a Evolução da Eficiência das Odds no Mercado de Apostas de Futebol

MATHEUS SANTOS DE OLIVEIRA FLOR

João Pessoa - PB
2026

MATHEUS SANTOS DE OLIVEIRA FLOR

Estudo sobre a Evolução da Eficiência das Odds no Mercado de Apostas de Futebol

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência de Dados para Negócios do Centro de Ciências Sociais Aplicadas da Universidade Federal da Paraíba (UFPB), como requisito para obtenção do grau de Bacharel em Ciência de Dados para Negócios.

Orientador: Jorge Henrique Norões Viana

João Pessoa - PB
2026

Catalogação na publicação
Seção de Catalogação e Classificação

F632e Flor, Matheus Santos de Oliveira.
Estudo sobre a evolução da eficiência das Odds no mercado de apostas de futebol / Matheus Santos de Oliveira Flor. - João Pessoa, 2026.
47 f. : il.

Orientação: Jorge Henrique Norões Viana.
TCC (Graduação) - UFPB/CCSA.

1. Análise de Dados esportivos. 2. Aprendizado de máquina no futebol. 3. Eficiências das Odds. 4. Sports Data Analysis. I. Viana, Jorge Henrique Norões. II. Título.

UFPB/CCSA CDU 004.6:005(043)

Elaborado por ANDRE DOMINGOS DA SILVA FERNANDES - CRB-15/00730

MATHEUS SANTOS DE OLIVEIRA FLOR

Estudo sobre a Evolução da Eficiência das Odds no Mercado de Apostas de Futebol

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência de Dados para Negócios do Centro de Ciências Sociais Aplicadas da Universidade Federal da Paraíba (UFPB), como requisito para obtenção do grau de Bacharel em Ciência de Dados para Negócios.

Trabalho Aprovado. João Pessoa - PB, 09 de abril de 2026:

Jorge Henrique Norões Viana
Orientador

Prof. Hilton Martins de Brito Ramalho
Examinador

Prof. Paulo Aguiar do Monte
Examinador

João Pessoa - PB
2026

Agradecimentos

Primeiramente, agradeço a Deus pela minha vida e aos meus pais pelo incentivo aos estudos desde a minha infância. Ao meu orientador, Jorge Henrique Norões Viana, expresso minha gratidão pelas reuniões semanais e pelas orientações que foram importantes para os ajustes e o desenvolvimento deste trabalho. Agradeço também à Universidade Federal da Paraíba pelo suporte acadêmico e financeiro, estruturais para a minha permanência e formação ao longo da graduação, e aos professores do corpo docente, cujas diferentes metodologias e níveis de exigência me ensinaram sobre adaptação, didática, entre outros valores. Por fim, agradeço aos amigos que conheci nesta universidade, pois tivemos várias trocas de experiências, aprendizado sobre como trabalhar melhor em grupo, contato com visões de vida diferentes e a oportunidade de evoluir ainda mais como pessoa.

Resumo

Este trabalho tem por objetivo analisar e comparar o grau de eficiência informacional dos mercados de apostas de futebol para os campeonatos da *Premier League* inglesa e do Campeonato Brasileiro (Série A), com o propósito de identificar a presença, a magnitude e as possíveis causas de ineficiências de precificação nas cotas (*odds*). A metodologia baseou-se na construção de uma base de dados por meio de extração automatizada (*web scraping*) e integração de variáveis climáticas. A modelagem preditiva, focada na classificação binária de resultados, comparou cinco algoritmos de aprendizado de máquina, otimizados via busca estocástica e validados pelo método intertemporal *walk-forward* para impedir o vazamento de dados. O teste das predições ocorreu em um ambiente simulado de *backtest* financeiro. Os resultados estatísticos demonstraram a superioridade da Regressão Logística com regularização L2 (acurácia de 60,17% e AUC de 0,6202 no conjunto de teste), que superou arquiteturas complexas baseadas em árvores de decisão. Na aplicação financeira direcionada por limiares de Valor Esperado (+EV), constatou-se uma assimetria de mercado: o algoritmo alcançou um Retorno sobre o Investimento (ROI) positivo de 0,78% no Campeonato Brasileiro, mas registrou prejuízo de -38,35% na *Premier League*. Conclui-se, amparado adicionalmente por simulações de Monte Carlo, que o mercado inglês valida a Hipótese do Mercado Eficiente em sua forma semiforte devido à sua alta liquidez, anulando vantagens estatísticas básicas, enquanto o cenário brasileiro, em fase de consolidação, ainda apresenta ineficiências marginais passíveis de exploração quantitativa.

Palavras-chave: Eficiências das Odds. Análise de Dados Esportivos. Apostas Esportivas. Aprendizado de Máquina no Futebol.

Abstract

This study aims to analyze and compare the degree of informational efficiency in football betting markets for the English *Premier League* and the Brazilian Championship (Série A), with the purpose of identifying the presence, magnitude, and possible causes of pricing inefficiencies in the *odds*. The methodology was based on building a database through automated extraction (*web scraping*) and the integration of weather variables. The predictive modeling, focused on binary outcome classification, compared five machine learning algorithms, optimized via stochastic search and validated by the intertemporal *walk-forward* method to prevent data leakage. The prediction testing occurred in a simulated financial *backtest* environment. Statistical results demonstrated the superiority of Logistic Regression with L2 regularization (60.17% accuracy and 0.6202 AUC in the test set), which outperformed complex architectures based on decision trees. In the financial application guided by Expected Value (+EV) thresholds, a market asymmetry was observed: the algorithm achieved a positive Return on Investment (ROI) of 0.78% in the Brazilian Championship, but recorded a loss of -38.35% in the *Premier League*. It is concluded, further supported by Monte Carlo simulations, that the English market validates the Efficient Market Hypothesis in its semi-strong form due to its high liquidity, nullifying basic statistical advantages, whereas the Brazilian scenario, in its consolidation phase, still presents marginal inefficiencies that can be quantitatively exploited.

Keywords: Odds Efficiency. Sports Data Analysis. Sports Betting. Machine Learning in Football.

Lista de tabelas

Lista de ilustrações

Sumário

1 Introdução

Este trabalho investiga a eficiência do mercado de apostas esportivas em partidas de futebol, focando na análise das cotas (ou *odds*). Tecnicamente, as *odds* são representações numéricas inversamente proporcionais à probabilidade estimada de um evento ocorrer, determinando o retorno financeiro potencial de um investimento. O problema preditivo é abordado de forma binária (vitória do mandante contra empate ou vitória do visitante). A premissa central desta pesquisa reside na comparação empírica entre dois ecossistemas estruturalmente distintos: a Premier League Inglesa e o Campeonato Brasileiro (Série A). A justificativa para essa escolha apoia-se nos critérios de microestrutura de mercado. Na literatura econômica, a maturidade de um mercado de apostas é atestada primariamente pelo seu volume de capital (liquidez) e pela velocidade de assimilação de novas informações. Trabalhos canônicos, como o de Croxson e Reade (2014), demonstram que a liga inglesa opera no ápice da eficiência: a injeção massiva de capital global e a atuação de sindicatos profissionais forçam as linhas de fechamento (*closing lines*) a refletirem instantaneamente as probabilidades reais, o que se traduz em dados com baixíssima margem de lucro (*juice*) para a casa e raras anomalias estatísticas. Em contrapartida, o mercado brasileiro atua como um cenário em consolidação, onde a menor liquidez relativa, a volatilidade do público e as nuances regionais sugerem a existência de ineficiências marginais e erros de precificação passíveis de exploração.

Esses achados dialogam diretamente com os limites preditivos mapeados pela academia. Historicamente, a acurácia de modelos quantitativos no futebol encontra um teto estatístico entre 54% e 57%, em virtude da natureza estocástica e de baixa pontuação do esporte. O desempenho de 60,17% alcançado nesta pesquisa posiciona o modelo linear no limite superior da previsibilidade para problemas binários. Ademais, a discrepância financeira entre as duas ligas corrobora a premissa de que mercados de altíssima liquidez ajustam suas linhas de fechamento com precisão quase perfeita, enquanto mercados locais ainda apresentam ineficiências passíveis de exploração.

A dificuldade de obter retornos consistentes nesse ecossistema reside no modelo de negócios das operadoras. A margem da casa (também denominada *vig* ou *juice*) atua como o pilar estrutural das plataformas de apostas. A operadora ajusta as cotas de tal forma que, ao receber um volume equilibrado de investimentos em todos os resultados possíveis de uma partida, obtenha lucro financeiro independentemente do desfecho esportivo. Em um cenário ideal com dois resultados equiprováveis, em vez de oferecer cotas justas de 2.00, a plataforma oferece valores em torno de 1.90. Essa redução sistemática garante a margem de lucro institucional e atua como a principal barreira matemática para o apostador.

Apesar da eficiência estrutural desse sistema, a dinâmica econômica e competitiva

de cada liga afeta o comportamento do público. As cotas nem sempre coincidem de forma idêntica com as probabilidades implícitas dos eventos puros, pois as operadoras também ajustam seus valores para equilibrar o capital injetado por apostadores enviesados, o que gera distorções temporárias.

No cenário brasileiro, a regulamentação do mercado, efetivada plenamente no final de 2024, atuou como o principal catalisador para o setor. A definição de regras claras trouxe segurança jurídica para as empresas e confiança para os usuários, resultando em um aumento expressivo no número de apostadores e no volume de depósitos. Análises de mercado focadas no comportamento do consumidor apontam a dominância absoluta do futebol, que responde por valores entre 75% e 85% de todo o volume de apostas gerado no país. O basquete, que historicamente ocupa o segundo lugar na preferência nacional, raramente ultrapassa a marca de 7% da participação.

A estratégia de marketing das operadoras evidencia esse fenômeno de concentração. Em 2025, todos os vinte clubes da Série A do Campeonato Brasileiro possuíam acordos de patrocínio com empresas do setor. A publicidade em transmissões esportivas e mídias digitais é massivamente direcionada ao futebol. Além disso, a profundidade de mercados disponíveis para uma única partida do torneio nacional, que engloba centenas de opções derivadas de estatísticas periféricas como escanteios e cartões, reflete o elevado engajamento do público. Diante desse volume financeiro, o presente estudo foca estritamente na evolução da eficiência das cotas no mercado de futebol.

Para sustentar a modelagem preditiva e testar as ineficiências matemáticas, buscou-se uma abordagem metodológica diferenciada na coleta de informações. A maioria dos trabalhos na área fundamenta-se em conjuntos de dados estáticos e pré-compilados, extraídos de repositórios como Kaggle ou plataformas genéricas. Embora possuam validade, esses arquivos oferecem um escopo limitado de variáveis preditivas. Esta pesquisa mitiga essa limitação ao aplicar a técnica de extração automatizada de dados (*web scraping*) na plataforma Flashscore, o que permitiu a coleta de métricas de jogo avançadas para a composição dos algoritmos de aprendizado de máquina.

Para estruturar a investigação desta hipótese, o presente trabalho está organizado em seis capítulos. O Capítulo 2 apresenta a Revisão da Literatura, estabelecendo os fundamentos teóricos que norteiam a pesquisa. O Capítulo 3 detalha a Metodologia, englobando os processos de coleta, a engenharia de características e os algoritmos de classificação selecionados. O Capítulo 4 expõe a Análise Exploratória dos Dados, detalhando as estatísticas descritivas, as tendências centrais e as medidas de dispersão das variáveis coletadas. O Capítulo 5 consolida a apresentação e a discussão empírica dos resultados obtidos nos testes financeiros. Por fim, o Capítulo 6 apresenta as Considerações Finais, sintetizando as descobertas e as limitações do estudo.

2 Revisão da Literatura

A literatura acadêmica frequentemente diferencia as apostas esportivas dos jogos de loteria, embora ambas as atividades sejam legalmente classificadas como apostas de quota fixa na legislação brasileira (BRASIL, 2023). A distinção teórica reside na relação entre aleatoriedade, informação e previsibilidade estatística. Na loteria, a probabilidade de um evento ocorrer é fixa e estocasticamente independente de sorteios anteriores. Conforme Mlodinow (2009), não existem variáveis observáveis que permitam ao indivíduo alterar a probabilidade matemática de acerto. A análise de resultados históricos não confere poder preditivo, uma vez que o sistema não possui memória.

Taleb (2021) aborda a dificuldade analítica na compreensão da aleatoriedade pura, argumentando que, em sistemas estritamente aleatórios, a habilidade do agente é um fator nulo. Em ambientes governados exclusivamente pelo acaso, variáveis como conhecimento ou experiência prévia não influenciam as probabilidades fundamentais. O autor defende que, nesses cenários, os dados passados não fornecem informações preditivas válidas sobre eventos futuros. Consequentemente, a loteria configura-se como um sistema de esperança matemática estruturalmente negativa, no qual a ação do participante restringe-se à decisão de alocação de capital.

Em contrapartida, o mercado de apostas esportivas opera como um mercado de informação complexo (SAUER, 1998; LEVITT, 2004), no qual as probabilidades são condicionais e dinâmicas. O resultado de uma partida de futebol não constitui um evento estocasticamente independente. Sua probabilidade é condicionada por múltiplas variáveis empíricas, tais como o desempenho histórico das equipes, métricas ofensivas e defensivas e a influência do mando de campo (DIXON; COLES, 1997). Além disso, fatores exógenos e táticos — como a condição física dos atletas, suspensões, condições climáticas e o contexto do confronto — exercem impacto sobre o desfecho. A interdependência dessas variáveis exige uma modelagem analítica que afasta a modalidade do puro acaso.

Para a estruturação quantitativa desse ecossistema, a literatura econômica adota o mercado de apostas esportivas como ambiente empírico para o teste da Hipótese do Mercado Eficiente (HME), formulada por Fama (1970) para o mercado financeiro. De acordo com a HME, em um mercado eficiente, os preços (no contexto esportivo, as *odds*) refletem de maneira completa e instantânea todas as informações disponíveis. Thaler e Ziemba (1988) investigaram anomalias preliminares nesses mercados e identificaram o fenômeno do viés favorito-zebra (*Favorite-Longshot Bias*), apontando assimetrias comportamentais na precificação inicial estabelecida pelas operadoras.

A eficiência do mercado, entretanto, apresenta variações correlacionadas à sua liquidez. Vlastakis, Dotsis e Markellos (2009) demonstraram que ligas com maior volume

financeiro ajustam suas ineficiências de precificação com rapidez, impulsionadas pela atuação de algoritmos de arbitragem. Em alinhamento a essa perspectiva, Angelini e De Angelis (2019) analisaram o mercado de apostas em futebol e observaram que competições consolidadas, como a English Premier League, exibem elevada eficiência informacional, o que restringe as oportunidades de extração sistemática de Valor Esperado Positivo (+EV).

Diante dessa eficiência informacional, a comunidade científica tem desenvolvido ferramentas preditivas avançadas. Dixon e Coles (1997) estabeleceram a base da modelagem quantitativa moderna no futebol ao adaptarem a distribuição de Poisson para a previsão de gols. Com a expansão do volume de dados (*Big Data*), as abordagens paramétricas tradicionais foram complementadas por técnicas de aprendizado de máquina. Constantinou, Fenton e Neil (2012) propuseram o modelo *pi-football*, fundamentado em Redes Bayesianas. Posteriormente, Baboota e Kaur (2019) avaliaram diferentes arquiteturas preditivas na liga inglesa e observaram que a acurácia global tende a estabilizar no intervalo entre 55% e 57%. Essa limitação preditiva é atribuída à natureza estocástica e de baixa pontuação do esporte (HUBÁČEK; ŠOUREK; ŽELEZNÝ, 2019), indicando que modelos matemáticos parcimoniosos podem apresentar melhor capacidade de generalização frente a arquiteturas complexas sujeitas ao sobreajuste (*overfitting*) causado pelo ruído estatístico.

Embora a literatura concentre suas análises na eficiência de mercados europeus de alta liquidez, a investigação do contexto brasileiro justifica-se pelo seu impacto socioeconômico e pela assimetria informacional em desenvolvimento. A pesquisa propõe-se a analisar as apostas esportivas sob a ótica quantitativa, demonstrando empiricamente o comportamento de um sistema de informações em consolidação.

O futebol é a modalidade esportiva de maior penetração no Brasil. Levantamentos indicam que aproximadamente 79% da população acompanha ou torce para uma equipe esportiva (CNN; ITATIAIA; QUAEST, 2023), o que o mantém na liderança isolada da preferência nacional (UOL, 2024). Esse engajamento converte-se diretamente em volume financeiro na indústria de apostas. Estimativas apontam que o setor movimentou dezenas de bilhões de reais em 2023 (PWC, 2023), impulsionado pela alta frequência de competições de abrangência nacional e internacional.

Do ponto de vista regulatório, o mercado operou inicialmente sob as diretrizes da Lei nº 13.756/2018 e alcançou sua formalização com a promulgação da Lei nº 14.790 (BRASIL, 2023). Essa estruturação jurídica conferiu segurança ao setor, atraiu investimentos e resultou em um crescimento expressivo na abertura de empresas operadoras (BNL DATA, 2023). Esse cenário de expansão acentuada, atrelado a um mercado cujas *odds* ainda não atingiram o grau de maturação das grandes ligas europeias, configura o ambiente empírico adequado para o teste quantitativo de ineficiências matemáticas abordado na presente investigação.

3 Metodologia

Esta seção apresenta os procedimentos metodológicos empregados na condução deste estudo, abrangendo as etapas de coleta, tratamento, enriquecimento e organização final dos dados. A arquitetura metodológica foi estruturada para assegurar a transparência e a replicabilidade da pesquisa, além de garantir a construção de um banco de dados robusto que serve como alicerce para as análises subsequentes.

A abordagem é fundamentalmente quantitativa, baseada na coleta e análise de um volume massivo de dados numéricos. Métricas como placares, estatísticas detalhadas de jogo (posse de bola, finalizações) e variáveis climáticas formam o núcleo da investigação. Essa estrutura permite a aplicação de técnicas estatísticas para identificar padrões, testar correlações e realizar análises comparativas entre os distintos contextos competitivos das ligas selecionadas, buscando a objetividade e a generalização dos achados.

O universo de interesse deste estudo compreende a totalidade das partidas de futebol profissional disputadas nas primeiras divisões do Brasil e da Inglaterra. Contudo, devido à inviabilidade técnica de analisar toda a extensão histórica dessas competições, definiu-se uma amostra representativa e viável. Trata-se de uma amostragem não probabilística e intencional, consistindo em um censo das partidas de duas das mais proeminentes ligas mundiais: o Campeonato Brasileiro (Série A) e a English Premier League. O recorte temporal definido abrange as temporadas de 2020 a 2025 para a competição brasileira e de 2019 a 2024 para a liga inglesa. A seleção destas competições específicas justifica-se por métricas contrastantes de engajamento e volume financeiro, detalhadas a seguir.

No contexto nacional, a escolha do Campeonato Brasileiro fundamenta-se na liderança global de consumo e acesso. A Similarweb, plataforma de inteligência de mercado e análise de tráfego web, atua como fonte credível para métricas de usuários únicos. Os relatórios do primeiro semestre de 2025 indicam que o Brasil ampliou a distância para o segundo colocado, mantendo-se como líder isolado em volume de tráfego em sites de apostas. As análises apontam que o acesso a essas plataformas no país superou a marca de 4 bilhões de visitas apenas nesse período. Eventos como as fases finais dos campeonatos estaduais, a Libertadores e o início do Brasileirão mantêm o engajamento em níveis elevados, consolidando o mercado local como um cenário de alta expansão.

De forma complementar ao volume de acessos brasileiro, a seleção da Premier League justifica-se pelo seu posto de maior mercado global em liquidez financeira. Relatórios de monitoramento elaborados por empresas de integridade de dados esportivos, como a Sportradar, apontam que a liga inglesa atrai o maior volume de capital do setor. As estimativas da indústria indicam que uma única partida da competição movimenta mais de 1 bilhão de euros globalmente. Esse massivo fluxo financeiro é o principal catalisador da

eficiência do mercado europeu. Com alto montante de capital injetado, as casas de apostas são forçadas a ajustar probabilidades com máxima precisão e a operar com margens de lucro reduzidas para equilibrar o risco. Portanto, a Premier League funciona como o padrão de eficiência na precificação de probabilidades, servindo de base comparativa ideal para o cenário brasileiro.

Por fim, a delimitação do recorte temporal para ambas as ligas fundamenta-se na viabilidade técnica da extração e na relevância do ciclo analisado. Observou-se que a plataforma Flashscore, fonte primária dos dados do estudo, consolidou uma estruturação mais rigorosa de estatísticas a partir de 2019 para a Inglaterra e 2020 para o Brasil. Em temporadas anteriores, o sistema apresentava lacunas significativas e ausência de variáveis vitais para a modelagem, o que comprometeria a integridade da base de dados. Dessa forma, o recorte temporal estabelecido cria uma linha de base pré-pandêmica segura, engloba o período de adaptação global e alcança as temporadas recentes, refletindo as novas dinâmicas e a estabilização do mercado de apostas.

Para garantir a transparência, o rigor científico e a reprodutibilidade exigidos na área de Ciência de Dados, a arquitetura metodológica, os *scripts* de extração (*web scraping*), as rotinas de engenharia de características e os modelos pré-treinados foram integralmente disponibilizados. O código-fonte completo pode ser acessado em repositório público através do endereço eletrônico: <https://github.com/Matheussantos25/TCC-efici-ncia-das-odds-no-futebol->.

O aprendizado de máquina é um subcampo da Inteligência Artificial (IA) voltado ao estudo de algoritmos que melhoram seu desempenho em determinadas tarefas à medida que adquirem experiência a partir dos dados, possibilitando identificar padrões e realizar previsões ou classificações sem depender de programação explícita (MITCHELL, 1997). A concepção de inteligência computacional fundamenta-se historicamente nas proposições de Alan Turing (1950), que idealizou a capacidade de uma máquina mimetizar o raciocínio humano por meio do aprendizado, superando a mera execução de algoritmos determinísticos baseados em regras estáticas. Contudo, a materialização teórica de Turing esbarrou, durante décadas, em severas limitações de processamento e armazenamento. Para dimensionar a diferença do salto tecnológico que superou essas barreiras, Narayana e Das (2024, p. 1) destacam que um smartphone contemporâneo possui um poder computacional 100.000 vezes superior ao do Apollo Guidance Computer (AGC), o sistema responsável por levar o homem à lua em 1969. Enquanto os computadores daquela época operavam com várias restrições arquitetônicas de memória e frequência com muitas limitações, os dispositivos e servidores atuais lidam rotineiramente com terabytes de informações em frações de segundo.

Essa democratização do poder de processamento, aliada à geração exponencial de informações na era digital, consolidou o paradigma do Big Data. No contexto esportivo moderno, cada partida gera um grande volumede variáveis estatísticas, táticas e financeiras. Com essa variedade nos dados isso viabiliza a aplicação de técnicas avançadas de aprendizado de máquina (machine learning), capazes de extrair padrões complexos e não-lineares que escapam à intuição humana ou à análise estatística tradicional.

Sabendo da importância do machine learning neste contexto, esta metodologia utilizará o aprendizado de máquina para desenvolver modelos preditivos capazes de estimar a probabilidade de vitória da equipe mandante em contraposição aos cenários de empate ou derrota (vitória do visitante), tratando o mercado de apostas sob a ótica de um problema de classificação binária. A premissa metodológica postula que a identificação sistemática de discrepâncias entre as probabilidades estimadas pelos algoritmos e aquelas implícitas no mercado permite inferir e explorar potenciais ineficiências financeiras (BUNKER; SUSNJAK, 2019).

A modelagem preditiva dos desfechos das partidas foi conduzida por meio da avaliação de cinco algoritmos com diferentes abordagens de aprendizado, partindo de um modelo linear paramétrico até métodos avançados de ensemble baseados em árvores de decisão, todos configurados para resolver um problema de classificação binária (Vitória do Mandante versus Empate/Derrota).

3.1 Regressão Logística

A Regressão Logística é o modelo de referência *(baseline)* clássico para problemas de classificação binária, sendo amplamente utilizada quando a interpretabilidade dos coeficientes e das odds resultantes é um requisito importante. Conforme explicam Hosmer, Lemeshow e Sturdivant (2013), o modelo opera transformando uma combinação linear de preditores em uma probabilidade por meio da função logística (curva sigmoide). Esta função restringe o resultado final ao intervalo entre zero e um, permitindo sua leitura direta como a probabilidade condicional de o time mandante vencer a partida:

$$P(Y = 1 \mid \mathbf{X}) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i X_i)}}$$

Nesta formulação, $P(Y = 1 \mid \mathbf{X})$ é a probabilidade estimada de ocorrer a vitória do mandante (classe positiva), dado o vetor $\mathbf{X}$ que contém as estatísticas da partida (como o diferencial de gols e finalizações). O termo β_0 é o intercepto, e β_i representa o peso (coeficiente) aprendido para cada variável X_i. A estimativa desses parâmetros ocorre maximizando a função de verossimilhança logarítmica, o que equivale a minimizar a função de erro conhecida como entropia cruzada binária (binary log-loss).Para resolver essa otimização, utilizam-se algoritmos iterativos numéricos, como o solver SAGA, que analisam a curvatura da função de erro para encontrar o ponto mínimo. No equilíbrio entre viés e variância, a regressão logística apresenta alto viés e baixa variância, traçando fronteiras de decisão estritamente lineares. Para evitar o subajuste em dados esportivos com alta multicolinearidade, aplicou-se a regularização L2 (Ridge) ou L1 (Lasso) via hiperparâmetro C, que restringe a magnitude dos coeficientes e impede que o modelo confie excessivamente em variáveis redundantes.

3.2 Random Forest

O algoritmo Random Forest, formalizado por Breiman (2001), representa um grande avanço não linear em relação aos modelos lineares. Ele opera combinando o resultado de múltiplas árvores de decisão, treinadas de forma independente em recortes aleatórios dos dados, gerando previsões mais robustas e generalizáveis. Para problemas de classificação binária, a decisão final do modelo é estabelecida por meio do voto majoritário (ou da média das probabilidades) de todas as árvores construídas:

$$\hat{y} = \arg\max_{c \in \{0,1\}} \sum_{b=1}^{B} I(T_b(\mathbf{x}) = c)$$

Nesta equação, $\hat{y}$ é a classe final prevista (1 para vitória, 0 para não vitória), e B é o número total de árvores no conjunto (ensemble). O termo $T_b(\mathbf{x})$ é a previsão individual da b-ésima árvore de decisão para os dados de entrada $\mathbf{x}$, e $I(\cdot)$ é a função indicadora,

que contabiliza um voto unitário sempre que a árvore prevê a classe c.A base teórica deste modelo é o princípio do bagging (bootstrap aggregating): criam-se várias amostras menores com reposição a partir do conjunto de treinamento, reduzindo drasticamente a variância final do estimador. Para garantir a descorrelação entre as árvores, Breiman (2001) introduziu o sorteio aleatório de variáveis (como max_features='sqrt') em cada nó. As divisões internas buscam separar os dados minimizando a impureza, quantificada pelo Índice de Gini, definido como $G = 1 - \sum_{i=1}^{2} p_i^2$, onde p_i é a proporção de amostras da classe vencedora ou perdedora naquele nó específico.

3.3 Gradient Boosting

O Gradient Boosting, introduzido por Friedman (2001), aborda a modelagem preditiva focando na redução rigorosa do viés por meio de uma construção sequencial e aditiva de árvores. Diferente do Random Forest, a ideia central do boosting é tratar o aprendizado como um problema de otimização em que cada nova árvore é construída especificamente para corrigir os erros (resíduos) deixados pelo conjunto das árvores anteriores. A previsão contínua final é dada pela soma:

$$F_M(\mathbf{x}) = F_0(\mathbf{x}) + \sum_{m=1}^{M} \nu \cdot h_m(\mathbf{x})$$

Nessa estrutura, $F_M(\mathbf{x})$ é o modelo final após M iterações, $F_0(\mathbf{x})$ é a estimativa inicial (o log-odds global da vitória do mandante), e $h_m(\mathbf{x})$ é a m-ésima árvore de decisão inserida no modelo. O parâmetro ν representa a taxa de aprendizado (learning rate), um fator de encolhimento essencial que pondera a contribuição de cada nova árvore, evitando que o algoritmo se ajuste rapidamente demais aos dados de treino.No contexto binário deste estudo, a saída contínua do modelo aditivo $F_M(\mathbf{x})$ é transformada na probabilidade final de vitória do mandante $\hat{p}(\mathbf{x})$ através da função sigmoide: $\hat{p}(\mathbf{x}) = \frac{1}{1+e^{-F_M(\mathbf{x})}}$. Por focar nas instâncias mais difíceis de prever, o Gradient Boosting pode modelar relações complexas entre diferenciais de ataque e defesa. Contudo, essa característica exige parametrização estrita, controlando a profundidade das árvores e subamostrando os dados (subsample) para conter o sobreajuste cronológico.

3.4 XGBoost

O XGBoost (eXtreme Gradient Boosting), criado por Chen e Guestrin (2016), é uma evolução metodológica e sistêmica do Gradient Boosting clássico. O algoritmo se destaca por incorporar diretamente o controle de complexidade estrutural durante o cálculo do ganho de informação para as ramificações. Sua principal inovação matemática

é a otimização de uma função objetivo rigorosamente regularizada:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)\right) + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

Nesta função objetivo $\mathcal{L}^{(t)}$ calculada no passo t, o primeiro termo representa o erro binary log-loss, onde l mede a diferença entre o alvo real $y_i \in \{0, 1\}$ e a soma da previsão anterior com o incremento da nova árvore $f_t(\mathbf{x}_i)$. O diferencial do modelo reside na penalidade estrutural: o hiperparâmetro γ atua exigindo uma redução mínima no erro para permitir a criação de novas folhas (sendo T o número total de folhas), funcionando como uma poda preventiva; enquanto λ aplica uma penalização L2 sobre os pesos contínuos das folhas (w_j).Para resolver essa equação com eficiência, Chen e Guestrin (2016) propõem o uso da expansão de Taylor de segunda ordem, analisando tanto o gradiente quanto a hessiana da função de perda. Isso permite calcular o ganho ótimo de cada corte analiticamente. Configurado com a diretiva objective='binary:logistic', o XGBoost demonstrou, no cenário esportivo brasileiro, forte capacidade de interagir métricas móveis recentes com as odds das casas de apostas, possuindo também suporte nativo ao tratamento eficiente de arrays esparsos.

3.5 LightGBM

O LightGBM (Light Gradient Boosting Machine), proposto por Ke et al. (2017), é uma estrutura desenhada para maximizar a escalabilidade do boosting sem perdas de acurácia. O algoritmo mitiga o gargalo computacional das avaliações de corte por meio do GOSS (Gradient-based One-Side Sampling), que foca o treinamento matemático nas instâncias com maiores erros de previsão, e do EFB (Exclusive Feature Bundling), que comprime variáveis.Para não enviesar a distribuição matemática ao subamostrar, o LightGBM utiliza uma fórmula específica de Ganho de Variância $\tilde{V}_j(d)$ ao avaliar o ponto de corte d na variável j:

$$\tilde{V}_j(d) = \frac{1}{n}\left(\frac{\left(\sum_{\mathbf{x}_i \in A_l} g_i + \frac{1-a}{b}\sum_{\mathbf{x}_i \in B_l} g_i\right)^2}{n_l^j(d)} + \frac{\left(\sum_{\mathbf{x}_i \in A_r} g_i + \frac{1-a}{b}\sum_{\mathbf{x}_i \in B_r} g_i\right)^2}{n_r^j(d)}\right)$$

Nesta formulação, o conjunto A retém os dados com grandes gradientes, e B é uma amostra reduzida dos dados já "resolvidos"pelo modelo (com pequenos gradientes). O termo $\frac{1-a}{b}$ é um amplificador que devolve o peso original aos dados de gradiente pequeno nas partições à esquerda (l) e à direita (r), onde g_i representa o erro. Com o objetivo configurado como binary_logloss, essa amostragem permitiu processar a engenharia de features do campeonato de forma extremamente veloz.Topologicamente, o LightGBM diverge por expandir suas árvores focado no melhor ganho por folha (leaf-wise), em vez de crescer nível por nível (level-wise). Essa expansão assimétrica reduz a entropia mais rápido, mas

requer forte controle arquitetônico. Por isso, a validação cruzada otimizou parâmetros cruciais como num_leaves e min_child_samples para impedir que o modelo memorizasse as nuances das partidas de treinamento em vez de generalizar os padrões de vitória.

3.5.1 Validação Cruzada e Otimização de Hiperparâmetros

A otimização de hiperparâmetros foi conduzida por meio de um processo de busca aleatória estruturada, aliada a uma validação cruzada temporal iterativa de três divisões. Esse delineamento cronológico é obrigatório no domínio de apostas esportivas para impedir o vazamento de dados do futuro para o passado (*data leakage*), garantindo que o algoritmo treine apenas com informações estritamente disponíveis antes da data do jogo predito.

Para assegurar uma exploração abrangente e com validade estatística no espaço de soluções, estipularam-se 100 candidatos aleatórios de configuração para cada modelo, totalizando 300 ajustes computacionais por algoritmo e 1.500 ajustes globais ao longo de todo o processo. As distribuições de busca incluíram variáveis uniformes e inteiras para calibrar a força de regularização (como o inverso da força de penalidade no modelo logístico e os termos de redução de perda e penalização nas estruturas de árvores), os limites de complexidade arquitetônica (como a profundidade máxima das árvores, variando de 3 a 15 níveis, e o limite de amostras mínimas por folha final) e as eficiências estocásticas (com taxas de aprendizado flutuando entre 0,001 e 0,055, além de frações de amostragem de colunas e observações). Todas as avaliações de otimização utilizaram como métrica de direcionamento a minimização da perda logística (*Log-Loss*).

Para quantificar de forma robusta o desempenho preditivo e a validade dos modelos na generalização em dados não vistos, foram adotadas duas métricas complementares de avaliação de classificação. A primeira é a **Acurácia (Accuracy)**, que representa a proporção bruta de previsões corretas estabelecidas pelo algoritmo. Matematicamente, a partir da simplificação dos Verdadeiros Positivos (TP), Verdadeiros Negativos (TN), Falsos Positivos (FP) e Falsos Negativos (FN), ela é definida por:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

No contexto deste estudo multinomial, ela reflete a taxa global de acerto categórico entre os desfechos das partidas. Embora seja uma métrica intuitiva e de interpretabilidade imediata, a acurácia isolada é insuficiente perante a complexidade e as incertezas estatísticas inerentes aos jogos de futebol. Por isso, adotou-se também a métrica **AUC (Area Under the Curve)**, que se refere à área matemática sob a curva ROC (*Receiver Operating Characteristic*). Ela transcende a decisão categórica simples e avalia a qualidade e a calibração das probabilidades geradas pelo modelo, traçando a relação contínua entre a Taxa de Verdadeiros Positivos (Sensibilidade) e a Taxa de Falsos Positivos (1 - Especificidade)

em infinitos limiares de corte. Uma AUC de 0,5 indica predições equivalentes ao puro acaso, enquanto 1,0 aponta para uma separação perfeita das classes.

No domínio do Aprendizado de Máquina, é imperativo estabelecer a distinção conceitual entre hiperparâmetros e parâmetros. Os hiperparâmetros são variáveis arquitetônicas definidas externamente antes do processo de treinamento, servindo como regras de controle do algoritmo. Por outro lado, os parâmetros são os valores internos que o modelo aprende durante o treinamento diretamente dos dados. A etapa de otimização conduzida neste estudo focou na calibração estocástica dos hiperparâmetros, visando maximizar a capacidade de generalização. A Tabela **??** consolida as configurações que minimizaram a função de entropia cruzada para cada arquitetura algorítmica após as iterações de busca.

3.5.2 Validação Cruzada e Otimização de Hiperparâmetros

A otimização de hiperparâmetros foi conduzida por meio de um processo de busca aleatória estruturada, aliada a uma validação cruzada temporal iterativa de três divisões. Esse delineamento cronológico é obrigatório no domínio de apostas esportivas para impedir o vazamento de dados do futuro para o passado (*data leakage*), garantindo que o algoritmo treine apenas com informações estritamente disponíveis antes da data do jogo predito.

Para assegurar uma exploração abrangente e com validade estatística no espaço de soluções, estipularam-se 100 candidatos aleatórios de configuração para cada modelo, totalizando 300 ajustes computacionais por algoritmo e 1.500 ajustes globais ao longo de todo o processo. A Tabela 1 detalha os valores mínimos, máximos e as opções categóricas que definiram a fronteira de pesquisa para cada algoritmo.

Tabela 1 – Espaço de busca de hiperparâmetros definido para a otimização

Modelo	**Hiperparâmetro**	**Espaço de Busca / Distribuição**
Regressão Logística	Força Inversa (C)	Uniforme (0.001 a 100.0)
	Penalidade	{L1, L2}
	Otimizador	{SAGA}
Random Forest	Número de Árvores	Inteiro (1000 a 3000)
	Profundidade Máxima	Inteiro (3 a 15)
	Amostras Mín. para Divisão	Inteiro (2 a 20)
	Amostras Mín. por Folha	Inteiro (1 a 15)
	Máximo de Variáveis	{*sqrt*, *log2*, None}
XGBoost	Número de Árvores	Inteiro (1500 a 4000)
	Taxa de Aprendizado	Uniforme (0.001 a 0.051)
	Profundidade Máxima	Inteiro (3 a 10)
	Fração de Subamostragem	Uniforme (0.6 a 1.0)
	Peso Mínimo na Folha	Inteiro (1 a 10)
	Redução Mínima de Perda (γ)	Uniforme (0.0 a 0.5)
LightGBM	Número de Árvores	Inteiro (1500 a 4000)
	Taxa de Aprendizado	Uniforme (0.001 a 0.051)
	Profundidade Máxima	Inteiro (3 a 12)
	Máximo de Folhas	Inteiro (15 a 127)
	Amostras Mín. por Folha	Inteiro (5 a 50)
	Regularização L1 (α)	Uniforme (0.0 a 1.0)
	Regularização L2 (λ)	Uniforme (0.0 a 1.0)
Gradient Boosting	Número de Árvores	Inteiro (1000 a 3000)
	Taxa de Aprendizado	Uniforme (0.005 a 0.055)
	Profundidade Máxima	Inteiro (3 a 10)
	Amostras Mín. para Divisão	Inteiro (2 a 20)
	Amostras Mín. por Folha	Inteiro (1 a 15)
	Máximo de Variáveis	{*sqrt*, *log2*, None}

Fonte: Elaboração própria (2026).

No domínio do Aprendizado de Máquina, é imperativo estabelecer a distinção conceitual entre hiperparâmetros e parâmetros. Os hiperparâmetros são variáveis arquitetônicas definidas externamente antes do processo de treinamento, servindo como regras de controle do algoritmo. Por outro lado, os parâmetros são os valores internos que o modelo aprende durante o treinamento diretamente dos dados. O objetivo da busca aleatória neste estudo foi encontrar a combinação ideal de hiperparâmetros que orientasse o aprendizado dos parâmetros internos de forma a minimizar a função de erro Log-Loss.

Para quantificar a eficácia das configurações selecionadas, o desempenho preditivo foi avaliado no conjunto de validação temporal isolado utilizando duas métricas complemen-

tares. A primeira é a **Acurácia**, que representa a proporção bruta de previsões corretas (vitória do mandante ou empate/derrota). A segunda é a **AUC** (Área sob a Curva ROC), que transcende a decisão categórica e avalia a qualidade da calibração das probabilidades geradas, penalizando modelos com excesso de confiança incorreta.

3.5.3 Engenharia de Características e Controle de Vazamento

A modelagem preditiva de eventos esportivos exige que os dados históricos brutos sejam transformados em indicadores que reflitam o momento das equipes antes do início da partida. Modelos de aprendizado de máquina não conseguem inferir dinâmicas temporais complexas apenas observando o registro estático de um jogo passado. Assim, é necessário construir preditores retrospectivos que resumam o desempenho acumulado de cada equipe. Neste trabalho, a engenharia de características foi estruturada a partir do cálculo de médias móveis globais e da derivação de diferenciais de desempenho.

Para capturar o momento de cada equipe, foram calculadas as médias móveis de quatro métricas principais: Gols Feitos, Gols Sofridos, Finalizações Feitas e Finalizações Sofridas. Com o intuito de analisar tendências de curtíssimo e curto prazo, adotaram-se duas janelas temporais (w): 3 e 5 partidas.

Um aspecto metodológico crucial abordado na literatura é a prevenção do vazamento de dados (*data leakage*). Se o desempenho da partida atual fosse incluído no cálculo da média, o modelo teria acesso a informações do futuro durante o treinamento, o que invalidaria sua capacidade de generalização. Para garantir a integridade temporal, aplicou-se um deslocamento (*shift*) de uma unidade de tempo. Assim, a média móvel de uma métrica M para a equipe k na partida atual t, considerando uma janela de w jogos, é formalmente definida pela seguinte equação:

$$\bar{M}_{k,t}^{(w)} = \frac{1}{w} \sum_{i=1}^{w} M_{k,t-i}$$

Onde:

- $\bar{M}_{k,t}^{(w)}$ é a média da métrica calculada para a partida t.
- $M_{k,t-i}$ representa o valor real da métrica registrada nos i-ésimos jogos estritamente anteriores à partida t.

Esta transformação foi aplicada independentemente do mando de campo (empilhamento de dados), para garantir que a média refletisse a força global da equipe, seja jogando em casa ou fora, nas últimas w rodadas.

3.6 Diferenciais de Desempenho

Embora as médias móveis forneçam um panorama do momento isolado de cada time, algoritmos baseados em árvores de decisão beneficiam-se substancialmente de características relativas, que comparam diretamente as duas entidades em confronto. Para comparar a diferença de habilidade entre equipe Mandante e a equipe Visitante, foram criados indicadores para cada janela temporal de 3 e 5 jogos.

Esses indicadores medem o saldo de poder ofensivo, defensivo e de volume de jogo. A formulação matemática para a métrica de ataque, por exemplo, é dada pela Equação 3.1:

$$\Delta\text{Ataque}_t^{(w)} = \bar{G}\text{Feitos}_{\text{Mandante},t}^{(w)} - \bar{G}\text{Feitos}_{\text{Visitante},t}^{(w)} \tag{3.1}$$

Seguindo a mesma lógica matemática, foram construídas as seguintes variáveis de confronto:

- Diferencial Ofensivo (Diff_Ataque): Subtração entre a média de gols feitos do mandante e a média de gols feitos do visitante. Valores positivos indicam superioridade ofensiva recente do time da casa.
- Diferencial Defensivo (Diff_Defesa): Subtração entre a média de gols sofridos do mandante e a média de gols sofridos do visitante. Valores menores (ou negativos) indicam que a defesa do mandante tem se mostrado mais sólida que a do adversário.
- Diferencial de Volume (Diff_Volume): Subtração entre as médias de finalizações feitas. Atua como um proxy (variável de aproximação) para a pressão e o controle ofensivo que a equipe costuma exercer sobre seus oponentes.

O alvo de predição do algoritmo (target) foi estabelecido a partir do resultado determinístico do confronto. A variável resposta (Y_t) foi codificada como um problema binário, assumindo valor 1 (Vitória do Mandante) se os Gols do Mandante superassem os Gols do Visitante na partida t, e valor 0 nos demais cenários (Empate ou Vitória do Visitante). Após a criação destas matrizes de características, as linhas contendo valores nulos (ocorridos naturalmente nas primeiras rodadas do campeonato devido à ausência de histórico suficiente) foram removidas da base de treinamento, assegurando a consistência estatística da modelagem.

3.7 Validação dos dados

A análise de eventos esportivos lida com séries de dados em que a ordem dos acontecimentos é importante, e ignorar essa sequência pode levar ao problema do data leakage

(vazamento de dados). Esse fenômeno ocorre quando informações futuras, que não estariam disponíveis em um cenário real de previsão, são utilizadas durante a fase de treinamento, resultando em avaliações de desempenho excessivamente otimistas (KAUFMAN et al., 2012). Por exemplo, se nos dados de treino for incluída uma coluna chamada placar final, o modelo aprenderá a prever resultados com base em informações futuras, obtendo 100% de acurácia de forma irrealista.

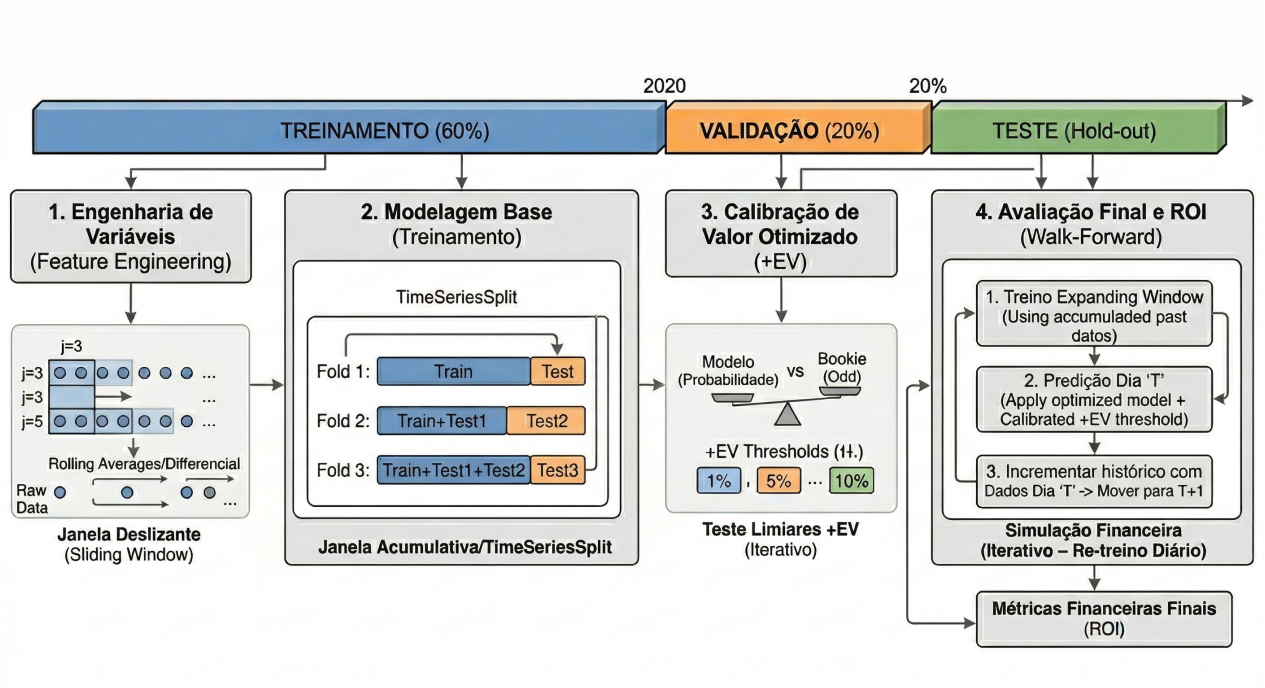
Para mitigar esse risco e capturar a dinâmica do esporte de forma adequada, a arquitetura metodológica adotada fez uso concomitante de duas abordagens de janelas temporais, aplicadas em etapas distintas. Inicialmente, na fase de engenharia de variáveis (*feature engineering*), utilizou-se a técnica de janelas deslizantes (*sliding windows*) para o cálculo das métricas de desempenho recente das equipes. Essa escolha fundamenta-se na premissa de que o futebol possui natureza dinâmica, sendo o recorte de momento (*momentum*) mais preditivo do que o histórico distante.Em contrapartida, para o treinamento e avaliação dos algoritmos, a técnica escolhida foi a validação cruzada intertemporal (*walk-forward*), operando sob a lógica de janelas acumulativas (*expanding windows*). Nela, o modelo é treinado em um conjunto de dados históricos e validado no período subsequente, repetindo-se o processo iterativamente. Diferente da janela deslizante, a janela acumulativa cresce continuamente para frente no tempo, incorporando novas informações sem descartar o histórico base, simulando de forma fiel como o modelo operaria em cenários reais para reconhecer padrões estatísticos de longo prazo (Bergmeir; Hyndman, 2018).

Para assegurar a validade empírica da modelagem e simular com fidelidade o cenário de apostas, a amostra total foi particionada de maneira estritamente cronológica, sem embaralhamento aleatório, dividida em três subconjuntos sequenciais: Treinamento (60%), Validação (20%) e Teste (20%).A proporção inicial de 60% dos dados históricos mais antigos compôs o conjunto de Treinamento. Foi exclusivamente nesta partição que os algoritmos de aprendizado de máquina (Regressão Logística e os modelos baseados em árvores) foram ajustados. Para otimizar os hiperparâmetros e mitigar o sobreajuste cronológico dentro deste bloco, aplicou-se a validação cruzada temporal (*TimeSeriesSplit* com 3 *folds*), garantindo que as otimizações iterativas também respeitassem a expansão acumulativa e a seta do tempo.A transição da avaliação estatística para a avaliação financeira exigiu o uso do conjunto de Validação (20% dos dados intermediários). Modelos preditivos esportivos não geram ordens de aposta automáticas; eles geram probabilidades. A decisão de apostar ocorre quando a probabilidade do modelo supera a probabilidade implícita da casa de apostas, gerando um Valor Esperado Positivo ($+EV$).

O conjunto de Validação atuou como um ambiente isolado para calibrar a agressividade do sistema, testando iterativamente múltiplos limiares de corte percentual de EV para identificar qual exigência de margem de valor maximizava o lucro simulado sem expor o capital a riscos desnecessários.Por fim, os 20% dos dados mais recentes constituíram o conjunto de Teste (*hold-out*). Esta partição foi mantida em isolamento absoluto (cega)

durante todas as etapas de modelagem e calibração de limiares. O desempenho financeiro final da estratégia preditiva, consubstanciado na métrica de Retorno sobre Investimento (ROI), foi apurado única e exclusivamente neste subconjunto. Durante este teste final, a rotina de *walk-forward* foi aplicada: a cada novo dia cronológico avaliado, os dados testados eram acrescidos à janela acumulativa de treinamento, atestando a real capacidade do sistema em generalizar seus lucros para eventos futuros não vistos e de se adaptar continuamente ao mercado.

Figura 1 – Validação dos Dados



Fonte: Elaboração Própria.

3.8 Dados

A coleta de dados foi realizada em três fases distintas e complementares, utilizando diferentes fontes e técnicas para construir um banco de dados abrangente.

O primeiro passo consistiu em obter os identificadores únicos de cada partida. Para isso, foi desenvolvido um script em Python utilizando a biblioteca Selenium com WebDriver Manager. O processo possui o seguinte fluxo: O script acessava as páginas de resultados de cada temporada das ligas selecionadas no site Flashscore, depois automatizava a navegação, clicando repetidamente no botão "Mostrar mais jogos"para carregar todo o histórico de partidas da página, para cada partida listada, o identificador único (ID) foi extraído do código-fonte HTML, Ao final, foi gerado um arquivo (match_ids_todas_ligas.csv) contendo uma lista consolidada e sem duplicatas de todos os IDs das partidas da amostra. Essa ideia foi necessária, pois para grande volume de dados fica inviável buscar url por url.

Com a posse da lista contendo os identificadores únicos de cada partida, deu-se início a uma nova fase do projeto, executando um segundo script em Python. Esta nova ferramenta, igualmente fundamentada nas robustas bibliotecas Selenium e BeautifulSoup, foi desenvolvida com o propósito de aprofundar a análise, nas estatísticas detalhadas de cada um dos jogos previamente mapeados. O processo é feito como um loop iterativo, no qual o script percorre, um a um, cada ID presente na lista. Para cada identificador, uma URL específica é dinamicamente construída, apontando diretamente para a página de estatísticas da partida correspondente no portal Flashscore.

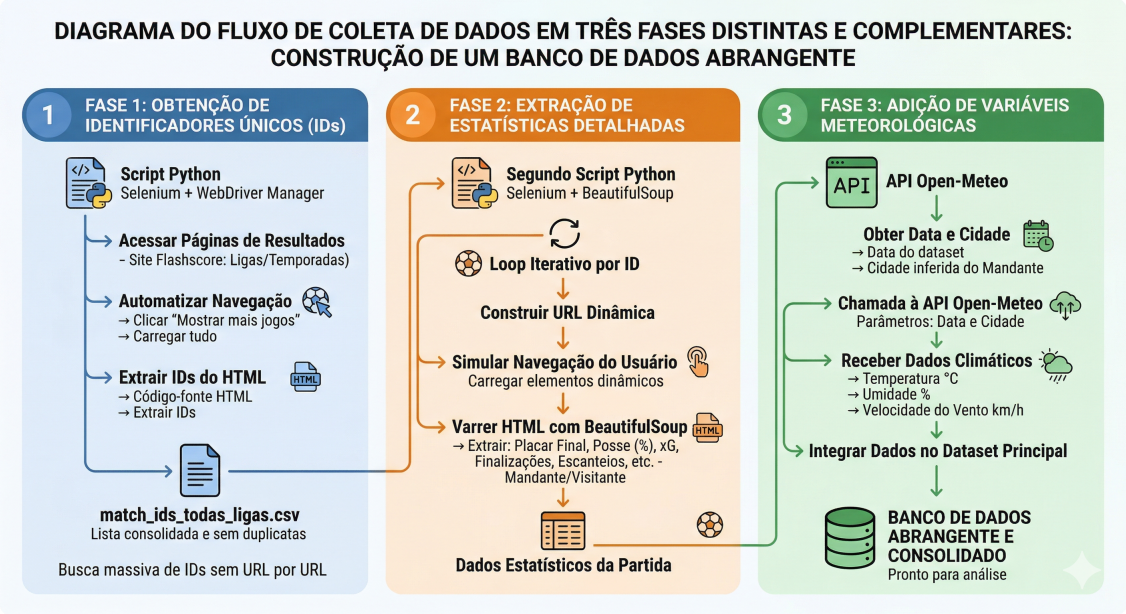
Essa abordagem garantiu que nenhuma partida fosse omitida. Uma vez na página de destino, o script entrava em ação, simulando a navegação de um usuário para garantir o carregamento completo de todos os elementos dinâmicos da página. Em seguida, com o auxílio do BeautifulSoup, era realizada uma varredura precisa do código HTML, extraindo um vasto e detalhado conjunto de variáveis de desempenho. Métricas cruciais como o placar final, os percentuais de posse de bola, gols esperados (xG), o número total de finalizações, a quantidade de escanteios e diversas outras estatísticas importantes foram coletadas, tanto para a equipe mandante tanto para a visitante. Ao final de cada iteração, os dados recém-extraídos eram organizados e estruturados, sendo então armazenados de forma temporária em uma estrutura de dados apropriada. Esta etapa de armazenamento provisório foi fundamental para consolidar as informações de maneira ordenada, preparando o terreno para fase de análise dos dados.

A terceira etapa foi buscar mais variáveis além dos dados que tem no site do flashscore e para alcançar esse objetivo, optou-se pela utilização da API Open-Meteo, uma solução robusta, de código aberto e gratuita, que se mostrou adequada para as necessidades do projeto, considerando o contexto acadêmico. Para cada partida registrada no dataset, havia data e a cidade do confronto, esta última sendo inferida a partir da identificação do time mandante, assimforam extraídas e utilizadas como parâmetros essenciais para

realizar uma chamada à API. Essa requisição programática permitiu consultar o banco de dados meteorológicos históricos da plataforma.

Como resultado de cada chamada bem-sucedida, a API retornava um conjunto detalhado de dados climáticos correspondentes à data e hora exatas em que cada jogo ocorreu. Dentre as informações disponíveis, foram selecionadas e integradas ao dataset principal as seguintes variáveis: a temperatura, medida em graus Celsius (°C); a umidade relativa do ar, expressa em porcentagem (%); e a velocidade do vento, registrada em quilômetros por hora (km/h). Essa integração adicionou uma nova dimensão de análise, permitindo investigar a potencial influência das condições meteorológicas no desempenho das equipes e nos resultados das partidas.

Figura 2 – Fluxo de Coleta de Dados



Fonte: Elaboração Própria.

4 Análise Exploratória dos Dados

Nesta seção, realiza-se a análise exploratória dos dados referentes aos resultados e aos gols das partidas do Brasileirão Série A e da Premier League nas temporadas analisadas. O objetivo é compreender o comportamento, as tendências centrais e a dispersão dos dados por meio de estatísticas descritivas. A seguir, detalham-se as medidas utilizadas para a construção das tabelas e suas respectivas formulações matemáticas, apresentadas de forma sequencial.

4.1 Frequência Relativa

A Frequência Relativa foi inicialmente utilizada para analisar a proporção de vitórias dos mandantes, empates e vitórias dos visitantes. No contexto do futebol, essa medida permite identificar padrões de resultados, como a vantagem de jogar em casa (fator mandante).

A frequência relativa (f_i) de um evento é calculada pela razão entre o número de vezes que o evento de interesse ocorreu (n_i) e o número total de partidas analisadas (N), expressa em percentual:

$$f_i = \left(\frac{n_i}{N}\right) \times 100 \tag{4.1}$$

Tabela 2 – Frequência de Resultados por liga e temporada

Liga: Brasileirão Série A			
Temporada	**% Vitória Mandante**	**% Empate**	**% Vitória Visitante**
2020	45,0%	28,4%	26,6%
2021	45,8%	29,7%	24,5%
2022	44,2%	28,4%	27,4%
2023	46,8%	25,8%	27,4%
2024	47,4%	26,6%	26,1%
2025	50,3%	26,1%	23,7%
Liga: Premier League			
Temporada	**% Vitória Mandante**	**% Empate**	**%Vitória Visitante**
2019	45,3%	24,2%	30,5%
2020	37,9%	21,8%	40,3%
2021	42,9%	23,2%	33,9%
2022	48,4%	22,9%	28,7%
2023	46,1%	21,6%	32,4%
2024	40,8%	24,5%	34,7%

Fonte: Elaboração Própria a partir de dados do Flashscore (2026).

A leitura da Tabela 2 evidencia uma diferença estrutural no fator mando de campo entre as competições. O Campeonato Brasileiro demonstra uma notável estabilidade defensiva e dependência do fator casa, com a taxa de vitória do mandante oscilando em uma faixa estreita e ascendente, culminando em 50,3% na temporada de 2025. A liga brasileira também apresenta uma elevada taxa de empates, frequentemente próxima à marca de 28%. Em contrapartida, a Premier League exibe maior volatilidade nos resultados, com o fator casa sendo menos determinante em temporadas específicas, como a de 2020, na qual a taxa de vitórias dos visitantes atingiu 40,3%. O campeonato inglês também registra uma proporção de empates sistematicamente inferior à do cenário nacional, indicando um estilo de jogo com menor propensão à neutralização tática prolongada.

A mesma lógica da frequência relativa é aplicada para avaliar a ocorrência de cenários específicos de gols, evidenciando a propensão ofensiva de cada liga. A Tabela 3 apresenta a proporção de partidas com pelo menos dois ou três gols e jogos onde ambas as equipes marcam.

Tabela 3 – Frequência de Gols e Ambas Marcam

Liga: Brasileirão Série A			
Temporada	**% Pelo menos 2 gols**	**% Pelo menos 3 gols**	**% Ambas Marcam**
2020	73,9%	46,1%	52,6%
2021	64,5%	39,5%	44,7%
2022	68,2%	43,4%	49,7%
2023	71,3%	45,0%	48,2%
2024	72,9%	47,4%	53,2%
2025	72,1%	46,1%	47,4%
Liga: Premier League			
Temporada	**% Pelo menos 2 gols**	**% Pelo menos 3 gols**	**% Ambas Marcam**
2019	80,0%	52,1%	51,1%
2020	73,4%	50,0%	48,9%
2021	77,1%	53,9%	50,0%
2022	75,5%	52,6%	51,6%
2023	86,1%	64,7%	61,6%
2024	81,3%	56,6%	57,4%

Fonte: Elaboração Própria a partir de dados do Flashscore (2026).

Os dados da Tabela 3 quantificam o perfil ofensivo mais pronunciado da liga inglesa. Enquanto o Campeonato Brasileiro mantém a frequência de jogos com três ou mais gols estagnada na faixa de 40% a 47%, a Premier League supera a marca de 50% na quase totalidade do recorte temporal, atingindo um pico de 64,7% na temporada 2023. A métrica de "Ambas Marcam"(*Both Teams to Score*) reforça esse cenário, mantendo-se consistentemente superior no campeonato europeu. Esses indicadores apontam que o futebol praticado na Inglaterra gera ambientes táticos mais abertos e propícios a transições ofensivas, refletindo diretamente na precificação das casas de apostas para os mercados de totais de gols (*Over/Under*).

4.2 Medidas de Tendência Central: Mediana e Moda

Para entender o comportamento padrão dos gols marcados, optou-se pela utilização da Mediana e da Moda. A distribuição de gols em partidas de futebol geralmente é assimétrica, com muitos jogos de poucos gols e poucos jogos com muitos gols, o que torna a média aritmética suscetível a distorções por valores extremos.

A **Mediana** (Tabela 4) representa o valor central do conjunto de gols marcados quando organizados em ordem crescente. Ela indica que 50% das partidas tiveram um número de gols igual ou inferior a esse valor. Sendo X o conjunto ordenado de gols e n o número total de observações, a mediana (Md) é dada por:

Se n for ímpar:

$$Md = X_{\frac{n+1}{2}} \tag{4.2}$$

Se n for par:

$$Md = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} \tag{4.3}$$

Tabela 4 – Mediana de gols por liga e temporada

Liga: Brasileirão Série A			
Temporada	**Mediana Mandante**	**Mediana Visitante**	**Mediana Total**
2020	1,000	1,000	2,000
2021	1,000	1,000	2,000
2022	1,000	1,000	2,000
2023	1,000	1,000	2,000
2024	1,000	1,000	2,000
2025	1,000	1,000	2,000
Liga: Premier League			
Temporada	**Mediana Mandante**	**Mediana Visitante**	**Mediana Total**
2019	1,000	1,000	3,000
2020	1,000	1,000	2,500
2021	1,000	1,000	3,000
2022	1,000	1,000	3,000
2023	2,000	1,000	3,000
2024	1,000	1,000	3,000

Fonte: Elaboração Própria a partir de dados do Flashscore (2026).

A **Moda** (Tabela 5) representa o valor que ocorre com maior frequência no conjunto de dados. Na análise, a moda indica o placar ou o número de gols mais comum por equipe ou por partida. Matematicamente, a moda (Mo) é o valor x_k tal que sua frequência absoluta $f(x_k)$ seja máxima:

$$Mo = \underset{x_k}{\operatorname{argmax}}\, f(x_k) \tag{4.4}$$

Tabela 5 – Moda de gols por liga e temporada

Liga: Brasileirão Série A			
Temporada	**Moda Mandante**	**Moda Visitante**	**Moda Total**
2020	1	1	2
2021	1	0	2
2022	1	1	1
2023	1	1	2
2024	1	1	3
2025	1	0	2
Liga: Premier League			
Temporada	**Moda Mandante**	**Moda Visitante**	**Moda Total**
2019	1	1	2
2020	1	1	2
2021	1	1	2
2022	1	0	2
2023	1	1	2
2024	1	1	2

Fonte: Elaboração Própria a partir de dados do Flashscore (2026).

A análise conjunta da mediana e da moda revela a expectativa matemática primária do apostador em cada competição. No Brasil, a mediana de gols totais permanece inalterada em 2 ao longo de todo o período, suportada por uma moda que varia frequentemente entre 1 e 2 gols por partida. O dado mais revelador do Campeonato Brasileiro encontra-se na moda do visitante, que registrou valor zero nas temporadas de 2021 e 2025. Isso comprova que o resultado mais frequente para uma equipe que atua fora de casa no Brasil é não marcar nenhum gol. Por outro lado, a Premier League sustenta uma mediana total consolidada de 3 gols na maioria das temporadas, atestando uma distribuição estatística intrinsecamente deslocada para placares mais elásticos.

4.3 Medidas de Dispersão: Máximos, Mínimos e Amplitude

Para compreender a variação extrema na quantidade de gols, analisam-se os valores máximos, mínimos e a amplitude. O valor **Máximo** (Tabela 6) representa o maior número de gols ocorridos em uma temporada. O valor máximo é expresso por $X_{max} = \max(X)$.

Como o valor mínimo de gols possível em uma partida é estatisticamente atestado como zero em ambas as ligas para todas as temporadas, a **Amplitude** (Tabela **??**) neste conjunto de dados torna-se equivalente ao valor máximo absoluto. A amplitude (A) é calculada por $A = X_{max} - X_{min}$. Valores elevados indicam a ocorrência de partidas com disparidade técnica aguda.

Tabela 6 – Máximo e Amplitude de gols por liga e temporada

Liga: Brasileirão Série A			
Temporada	**Máx/Amp Mandante**	**Máx/Amp Visitante**	**Máx/Amp Total**
2020	5	5	8
2021	5	5	7
2022	6	4	8
2023	7	6	10
2024	5	6	8
2025	8	6	8
Liga: Premier League			
Temporada	**Máx/Amp Mandante**	**Máx/Amp Visitante**	**Máx/Amp Total**
2019	8	9	9
2020	9	7	9
2021	7	6	9
2022	9	6	9
2023	6	8	8
2024	7	6	9

Fonte: Elaboração Própria a partir de dados do Flashscore (2026).

A observação dos extremos na Tabela 6 quantifica a presença de disparidades técnicas em campo. A liga inglesa demonstra capacidade rotineira de produzir placares extremos, registrando marcas de 9 gols totais em quatro das seis temporadas analisadas, inclusive com visitantes atingindo tetos de 9 gols feitos em uma única partida (temporada 2019). No Campeonato Brasileiro, o limite usual de gols totais estabelece-se na marca de 8, com o campeonato de 2023 configurando um *outlier* estatístico ao registrar uma partida com 10 gols. Essa métrica sugere que a elite inglesa, impulsionada pela discrepância financeira entre os clubes do topo e da base da tabela, proporciona um terreno fértil para goleadas elásticas, enquanto o cenário brasileiro mantém os confrontos limitados a escores mais contidos.

4.4 Medida de Dispersão: Desvio Padrão

O **Desvio Padrão** (Tabela 7) quantifica o grau de dispersão ou variabilidade da quantidade de gols em torno da média aritmética. Um desvio padrão baixo indica que a maioria dos jogos teve um número de gols próximo à média da liga (refletindo um maior equilíbrio nos placares), enquanto um desvio padrão alto sugere uma maior irregularidade e volatilidade de gols ao longo das rodadas.

Sendo os dados referentes a uma amostra de uma temporada específica, a fórmula do desvio padrão amostral (s) é dada por:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{4.5}$$

Onde n é o número de partidas, x_i é a quantidade de gols na partida i, e $\bar{x}$ é a média de gols da temporada.

Tabela 7 – Desvio Padrão de gols por liga e temporada

Liga: Brasileirão Série A			
Temporada	**Desvio Mandante**	**Desvio Visitante**	**Desvio Total**
2020	1,102	1,060	1,543
2021	1,079	0,991	1,498
2022	1,184	0,888	1,465
2023	1,190	1,091	1,633
2024	1,085	0,986	1,422
2025	1,259	1,029	1,547
Liga: Premier League			
Temporada	**Desvio Mandante**	**Desvio Visitante**	**Desvio Total**
2019	1,249	1,200	1,512
2020	1,320	1,258	1,756
2021	1,327	1,259	1,626
2022	1,420	1,184	1,791
2023	1,366	1,278	1,657
2024	1,278	1,190	1,617

Fonte: Elaboração Própria a partir de dados do Flashscore (2026).

A análise da dispersão confirma estatisticamente as inferências feitas a partir das amplitudes. A Premier League exibe desvios padrões consistentemente mais altos em todas as frentes (mandante, visitante e total), atingindo índices superiores a 1,7em gols totais nas temporadas de 2020 e 2022. Essa alta variabilidade matemática demonstra que prever a quantidade exata de gols em uma partida no campeonato inglês é um problema estocástico dotado de considerável ruído. Em contrapartida, o Campeonato Brasileiro apresenta desvios predominantemente contidos na faixa de 1,4 a 1,5 para gols totais e frequentemente inferiores a 1 para a produção ofensiva dos visitantes. Isso atesta que o cenário nacional, por ser mais nivelado tecnicamente, converge a maioria absoluta dos seus jogos para o entorno do placar médio, com baixa ocorrência de anomalias estatísticas na forma de goleadas atípicas.

4.5 Considerações Finais da Análise Exploratória

A compilação e a interpretação das métricas descritivas estabelecem um panorama estrutural contrastante entre os dois ecossistemas estudados, evidenciando que o futebol

não possui uma distribuição estatística universal, mas sim características moldadas pela natureza de cada competição.

O Campeonato Brasileiro configura-se matematicamente como um torneio de forte solidez defensiva, considerável nivelamento técnico e alta dependência da vantagem de jogar em casa. O fato de a moda de gols do visitante atingir o valor zero em múltiplas temporadas, aliado ao baixo desvio padrão de gols totais e à proporção elevada de empates, indica que as partidas no Brasil tendem a ser disputadas sob o viés do controle territorial e da redução de riscos. O mandante detém uma vantagem estatística clara, mas raramente essa vantagem se converte em placares elásticos, resultando em margens de vitória estreitas e jogos de baixa pontuação agregada.

Por outro lado, a Premier League reflete um perfil voltado à alta produtividade ofensiva e à assimetria técnica. Os indicadores apontam para medianas de gols estendidas para a marca de três tentos por partida, desvios padrões elevados e frequências sistematicamente maiores de eventos extremos (goleadas) e cenários nos quais ambas as equipes marcam. A flutuação nas taxas de vitória dos mandantes demonstra que o poder ofensivo global e o investimento na qualidade dos elencos superam a influência da localização geográfica da partida, permitindo aos times visitantes imporem seus esquemas táticos com maior liberdade do que no cenário sul-americano.

Do ponto de vista metodológico para a etapa subsequente deste trabalho, essas descobertas validam a premissa de que algoritmos de aprendizado de máquina não podem ser alimentados de forma genérica. As probabilidades basais (*priors*) necessárias para a modelagem são inerentemente distintas. A precificação estrita de uma probabilidade de vitória ou da ocorrência de múltiplos gols exigirá que as árvores de decisão e os coeficientes logísticos capturem a contenção tática do campeonato brasileiro de forma independente do viés ofensivo do mercado inglês.

5 Resultados

Este capítulo apresenta os resultados empíricos da modelagem preditiva e a subsequente simulação financeira das estratégias nos mercados do Campeonato Brasileiro e da Premier League. Inicialmente, expõe-se o desempenho comparativo dos algoritmos de aprendizado de máquina submetidos à etapa de otimização e validação, o que fundamentou a seleção do modelo principal. Na sequência, detalha-se a plataforma analítica desenvolvida para o teste cego financeiro e avalia-se a eficiência do mercado frente às predições geradas.

A Tabela 8 unifica as configurações ótimas encontradas pela busca estocástica e seus respectivos indicadores de desempenho estatístico no conjunto de validação.

Tabela 8 – Hiperparâmetros Otimizados e Avaliação de Desempenho no Conjunto de Validação

Modelo	**Configuração Otimizada**	**Log-Loss (Treino)**	**Acurácia (Valid.)**	**AUC (Valid.)**
Regressão Logística	C = 0,0789 Penalidade = L2 Otimizador = SAGA	-0,6739	60,17%	0,6202
Random Forest	Árvores = 2768 Profundidade = 3 Mín. Divisão = 14 Mín. Folha = 9 Variáveis = *sqrt*	-0,6739	59,53%	0,6140
XGBoost	Árvores = 1975 Taxa Aprend. = 0,0017 Profundidade = 9 γ = 0,0669 Subamost. = 92,52%	-0,6965	59,10%	0,6105
LightGBM	Árvores = 1873 Taxa Aprend. = 0,0033 Profundidade = 9 Folhas = 79 α=0,4804 | λ=0,9853	-0,7125	56,53%	0,5911
Gradient Boosting	Árvores = 1223 Taxa Aprend. = 0,0063 Profundidade = 3 Mín. Divisão = 19 Variáveis = *log2*	-0,7029	58,46%	0,6185

Fonte: Elaboração própria (2026).

A análise das configurações consolidadas revela que a alta estocasticidade do esporte exigiu forte controle sobre a complexidade para que os modelos pudessem aprender seus parâmetros de forma saudável. Na Regressão Logística, a severa penalidade L2 encontrada foi fundamental para forçar o algoritmo a encolher os coeficientes de variáveis esportivas correlacionadas, evitando o sobreajuste cronológico. Nos modelos baseados em árvores, notou-se a convergência unânime para taxas de aprendizado microscópicas aliadas a

truncamentos estruturais algorítmicos profundos, como o limite estrito de três níveis de profundidade na Random Forest.

Em uma quebra da expectativa teórica de que maior complexidade algorítmica necessariamente resulta em melhor capacidade preditiva, a avaliação identificou o modelo paramétrico linear como a arquitetura com o maior poder de generalização. Ao ultrapassar a barreira teórica do esporte e liderar tanto na Acurácia global (60,17%) quanto na calibração rigorosa das probabilidades via métrica AUC (0,6202), evidencia-se que a dinâmica subjacente das variáveis esportivas modeladas neste estudo expressa-se, primariamente, de forma linear e aditiva.

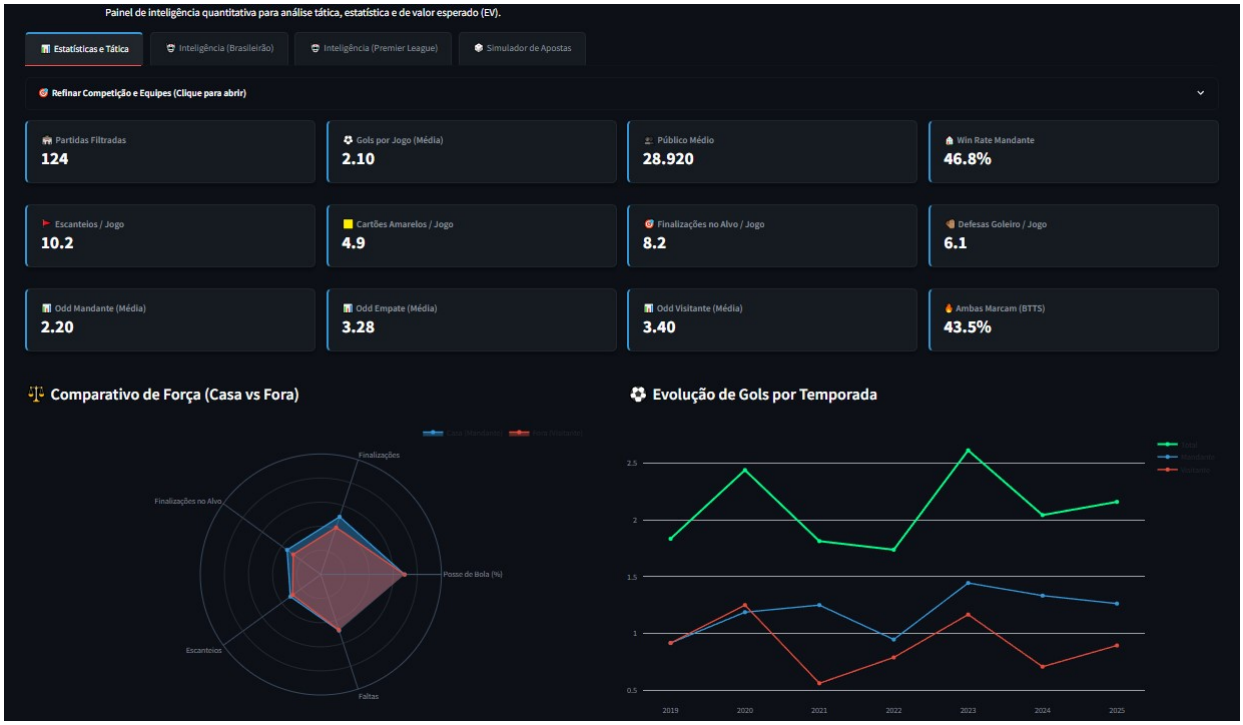
Com a Regressão Logística estabelecida como o preditor mais eficiente, a transição da avaliação puramente estatística para a simulação financeira exigiu a construção de um ambiente de testes rigoroso. Para operacionalizar a extração de resultados e validar empiricamente as estratégias de apostas, desenvolveu-se uma aplicação interativa focada em inteligência quantitativa esportiva, acessível publicamente por meio do endereço eletrônico <https://oddseficientes.streamlit.app>. O sistema foi programado na linguagem Python, utilizando a biblioteca de código aberto Streamlit. A adoção desta tecnologia justifica-se pela eficiência na transposição de rotinas de aprendizado de máquina e manipulação de conjuntos de dados para interfaces gráficas dinâmicas.

A visualização de dados foi estruturada utilizando as bibliotecas Plotly, Matplotlib e Seaborn. Enquanto o Matplotlib e o Seaborn foram empregados para a renderização estática de distribuições táticas, o Plotly assumiu a prototipação dos gráficos interativos, como os radares de força comparativa das equipes e as curvas financeiras de Retorno sobre Investimento (ROI).

A arquitetura do painel interativo (*dashboard*) foi projetada para atuar como um simulador dinâmico de *backtest*. O sistema importa o banco de dados histórico das partidas e carrega o *pipeline* do modelo vencedor pré-treinado (Regressão Logística, serializado via biblioteca *Joblib*). Por meio de componentes de interface, o usuário pode interagir com o conjunto de dados aplicando filtros multidimensionais. Essa funcionalidade permite isolar o desempenho por competições (Campeonato Brasileiro e Premier League), recortes temporais e faixas de cotações, estabelecendo limites de *odds* mínimas e máximas para os desfechos de mandante, empate e visitante.

Foi por intermédio desta plataforma que os resultados empíricos foram monitorados. A interface facilitou a extração instantânea de Indicadores-Chave de Desempenho (KPIs), a verificação comparativa da eficiência das casas de apostas e o acompanhamento contínuo da exposição ao risco (*drawdown*) gerada pelas entradas baseadas no Valor Esperado (+EV) ótimo do algoritmo. A Figura 3 ilustra a interface de estatísticas descritivas da aplicação, demonstrando a sumarização dos dados gerais e a capacidade de análise tática por meio de gráficos de radar e linhas de evolução temporal.

Figura 3 – Painel interativo de análise tática e estatística exploratória



Fonte: Elaboração Própria (2026).

5.1 Desempenho no Campeonato Brasileiro (Série A)

Para o cenário nacional, o algoritmo estabeleceu um limiar de corte restrito, exigindo um EV mínimo de 22,0% para legitimar uma aposta. Sob esta métrica, o modelo identificou 85 oportunidades (entradas) ao longo do período de teste. A curva de capital resultante demonstrou alta volatilidade inicial, atingindo um rebaixamento severo (*drawdown*) próximo à 40ª aposta, seguido de uma aguda e consistente recuperação que impulsionou o saldo para o território positivo nas rodadas finais.

Apesar das oscilações, o sistema encerrou o período em superávit, contabilizando um Retorno sobre o Investimento (ROI) final de 0,78%, equivalente a um lucro líquido aproximado de 0,66 unidades financeiras. Para a compreensão exata desta métrica, é fundamental detalhar a mecânica de gestão de banca utilizada na simulação. O modelo operou sob a premissa de *Flat Betting* (gestão de unidade fixa), estipulando que cada aposta validada pelo algoritmo representou o custo exato de 1 unidade de capital. Conforme a lógica do sistema, o lucro de uma aposta vencedora é calculado subtraindo-se o valor apostado da cotação original (Odd - 1), enquanto uma aposta perdedora resulta na dedução integral de 1 unidade.

Consequentemente, o montante total investido (*turnover*) no campeonato foi diretamente proporcional ao número de entradas (85 apostas), totalizando um denominador de 85 unidades financeiras. A extração do percentual de eficiência aplica a razão entre o

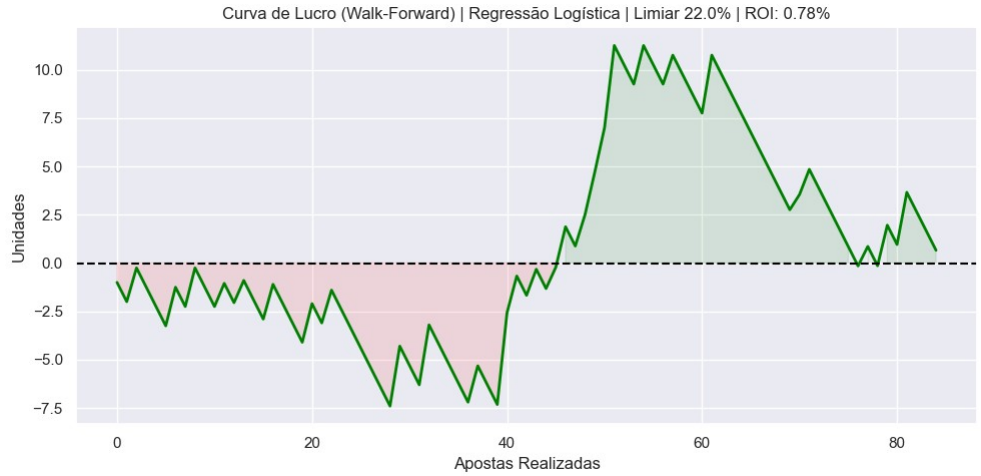
saldo líquido acumulado e o custo total da operação:

$$\text{ROI} = \left(\frac{\text{Lucro Líquido}}{\text{Total Investido}} \right) \times 100$$

$$\text{ROI} = \left(\frac{0,66}{85} \right) \times 100 \approx 0,7764\%$$

Com o arredondamento matemático padrão do algoritmo computacional para duas casas decimais, consolida-se o índice final de 0,78%.

Figura 4 – Curva de lucro acumulado no Campeonato Brasileiro via Regressão Logística



Fonte: Elaboração Própria (2026).

Alcançar um ROI positivo, ainda que modesto, significa superar matematicamente a margem de lucro estrutural embutida pelas próprias casas de apostas (*juice* ou *vig*), o que é um feito notável na literatura de *Sports Analytics*. Isso sugere que, no Campeonato Brasileiro, as probabilidades extraídas pelo modelo a partir das médias móveis de finalizações e gols conseguiram identificar ineficiências marginais na precificação, encontrando uma leve vantagem matemática (*edge*) ao longo da simulação de teste cego intertemporal.

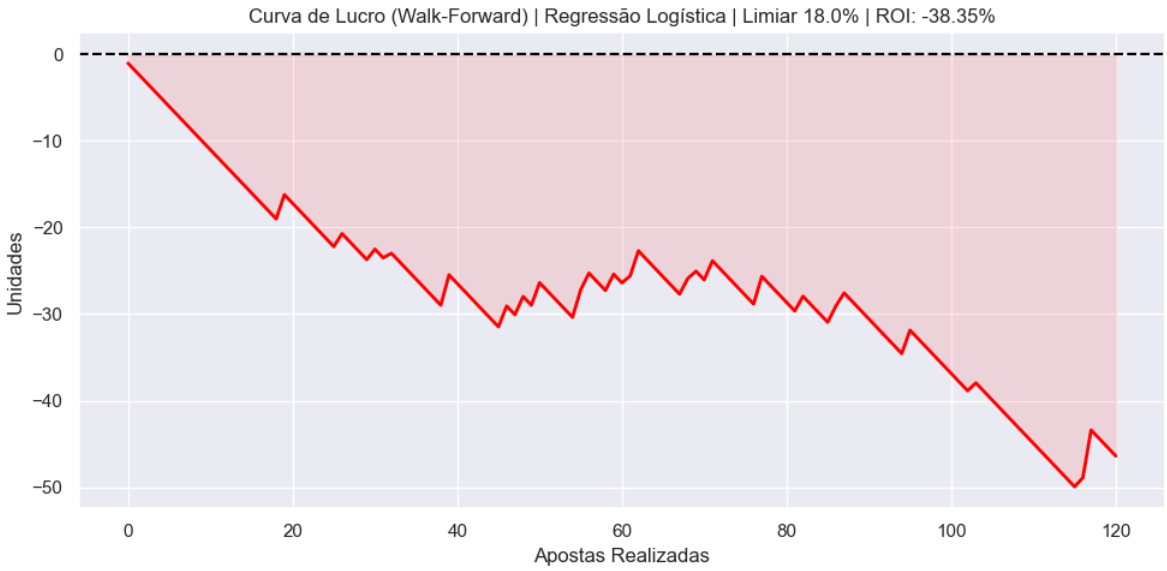
5.2 Desempenho na Premier League

A avaliação estendida ao campeonato inglês evidenciou um cenário preditivo notavelmente mais hostil. Otimizado para um limiar de EV de 18,0%, o sistema demonstrou maior agressividade, registrando 121 entradas no conjunto de teste. Contudo, a trajetória financeira (curva de lucro acumulado) apresentou um declínio agudo, quase linear e desprovido de recuperações significativas.

O resultado final na Premier League culminou em um colapso preditivo, registrando um ROI de -38,35% e uma perda acumulada de -50,90 unidades financeiras. A severidade

deste rebaixamento atesta que as cotas consideradas "de valor"pelo modelo eram, na realidade, armadilhas probabilísticas (*value traps*).

Figura 5 – Curva de lucro acumulado na Premier League via Regressão Logística



Fonte: Elaboração Própria (2026).

5.3 Discussão sobre a Eficiência de Mercado

A diferença entre as métricas da etapa de validação (na qual os modelos apresentaram calibração via AUC) e o resultado financeiro negativo no conjunto de teste suporta empiricamente a teoria de Fama (1970) sobre a Hipótese do Mercado Eficiente (HME).

A *Premier League*, por ser um dos mercados esportivos de maior liquidez financeira do mundo, incorpora em suas *odds* as informações estatísticas públicas, como poder ofensivo, volume de jogo e histórico recente. O modelo de Regressão Logística, alimentado por variáveis descritivas, revelou-se insuficiente para superar a precificação desse mercado. O resultado negativo indica que as casas de apostas calibram suas probabilidades com modelos dinâmicos que vão além dos recortes amostrais de gols e finalizações. Em suma, o mercado apresenta eficiência no longo prazo, e as divergências probabilísticas identificadas pela regressão representam vieses do algoritmo, e não falhas de precificação das operadoras.

Para ilustrar a eficiência do mercado e o impacto estocástico da variância nos resultados de curto prazo, a arquitetura do painel foi equipada com um módulo de Simulação de Monte Carlo. Essa ferramenta projeta múltiplos cenários de apostas independentes, isolando o fator sorte em relação à habilidade preditiva. A Figura 6 apresenta um teste estocástico composto por 10 usuários simulados, cada um executando uma sequência de 50 apostas aleatórias, sem a aplicação de filtros analíticos.

Figura 6 – Simulação de Monte Carlo ilustrando a variância e a margem das casas de apostas

Fonte: Elaboração Própria (2026).

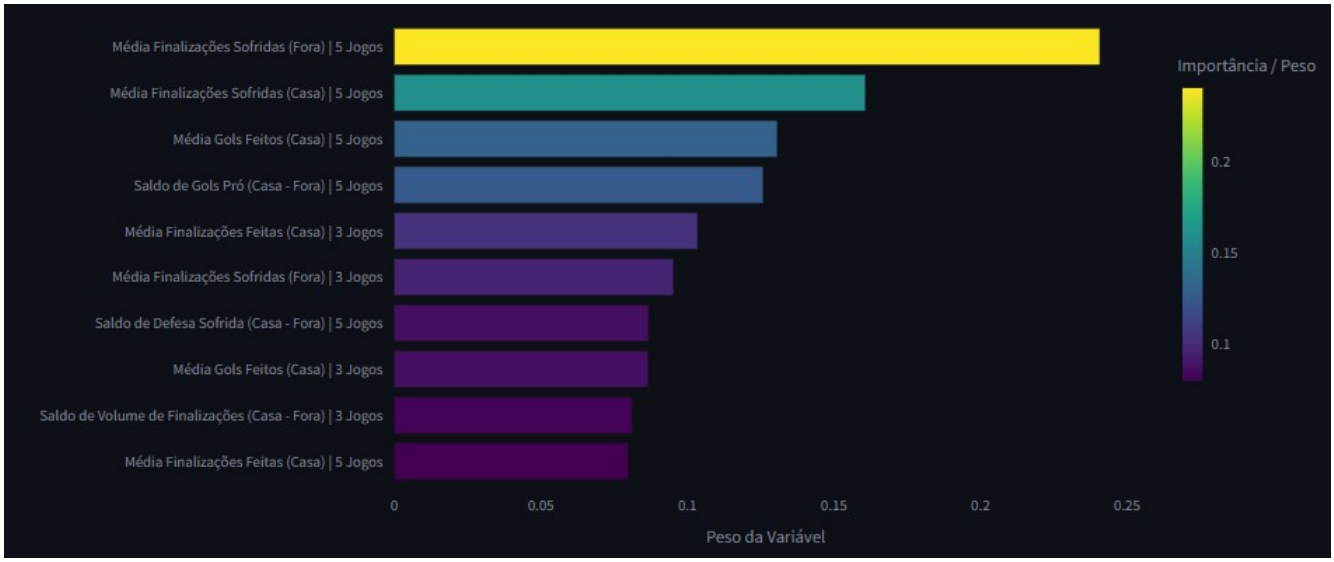
Os resultados globais extraídos desta simulação fornecem base empírica complementar à HME. A média do Retorno sobre o Investimento (ROI Final) dos usuários convergiu para -5,82%. Este indicador negativo reflete matematicamente a margem de lucro estrutural (*juice* ou *vig*) embutida pelas operadoras nas linhas de *odds*. Em um mercado eficiente, estratégias preditivas que não possuam vantagem matemática tendem a convergir para esse patamar de prejuízo.

Adicionalmente, o gráfico das trajetórias de saldo expõe o efeito da variância. Mesmo submetidos a uma esperança matemática negativa (-EV), 40,0% dos usuários simulados encerraram o ciclo de 50 apostas com saldo positivo, com o melhor desempenho isolado atingindo um ROI de +13,42%. Esse fenômeno demonstra a limitação das análises de curto prazo no futebol, pois oscilações estocásticas criam trajetórias vitoriosas provisórias que podem gerar viés de sobrevivência e falsa percepção de habilidade. Contudo, conforme a Lei dos Grandes Números atua, como observado no desempenho do modelo de *Machine Learning* ao longo das apostas na *Premier League*, a eficiência do mercado tende a conduzir o saldo final para o prejuízo esperado matematicamente.

5.4 Análise de Importância das Variáveis

Para compreender os fatores estatísticos determinantes nas predições do algoritmo otimizado, realizou-se a extração da importância das variáveis (*feature importance*). Devido à estrutura paramétrica da Regressão Logística, a metodologia consistiu na avaliação do valor absoluto dos coeficientes matemáticos atribuídos a cada preditor na equação de probabilidade. Para garantir a comparabilidade entre métricas de naturezas distintas, os dados foram previamente padronizados por meio da técnica *RobustScaler*. A Figura 7 apresenta as dez características de maior peso no processo de decisão do modelo.

Figura 7 – Importância das variáveis baseada no valor absoluto dos coeficientes logísticos



Fonte: Elaboração Própria (2026).

A leitura do ranqueamento evidencia uma predominância de métricas focadas em volume de jogo e vulnerabilidade defensiva. A variável de maior impacto absoluto no modelo foi a média de finalizações sofridas pela equipe visitante nos últimos cinco jogos, apresentando um coeficiente de peso de 0,2408. Na segunda posição, identificou-se a mesma métrica aplicada à equipe mandante, com peso de 0,1608. Esse comportamento indica que a consistência do sistema defensivo atua como um preditor de resultados mais seguro do que variáveis puramente ofensivas, tendo em vista que a média de gols feitos pelo mandante ocupou apenas a terceira colocação (0,1307).

Outro padrão relevante extraído do gráfico refere-se à amplitude das janelas de desempenho. As características baseadas em recortes temporais de cinco jogos dominam as quatro primeiras posições do ranqueamento, superando a influência das médias restritas a três jogos. Essa configuração sugere que a manutenção do nível técnico das equipes no médio prazo oferece um sinal estatístico mais limpo para o modelo preditivo, diluindo o ruído causado por eventuais oscilações de forma imediatas.

6 Considerações Finais

O presente trabalho teve como objetivo investigar a eficiência do mercado de apostas esportivas em partidas de futebol, comparando a precisão preditiva de algoritmos de aprendizado de máquina com as cotas (*odds*) estabelecidas pelas casas de apostas. Por meio de extração de dados e integração de variáveis climáticas, construiu-se um banco de dados com informações do Campeonato Brasileiro (Série A) e da *English Premier League*. Para a análise prática, desenvolveu-se uma aplicação interativa que permitiu a simulação financeira (*backtest*) das estratégias preditivas.

Na modelagem estatística, os resultados indicaram um desempenho da Regressão Logística superior ao de arquiteturas de *ensemble*, como o XGBoost e o LightGBM. A validação demonstrou que, devido à estocasticidade do esporte e à alta correlação entre as variáveis descritivas utilizadas (como médias de finalizações e saldo de gols), fronteiras de decisão lineares associadas à severa regularização L2 apresentaram maior capacidade de generalização. O modelo paramétrico obteve melhor ajuste aos dados não vistos, mitigando o sobreajuste cronológico comumente observado nos algoritmos baseados em árvores de decisão.

Na etapa de aplicação financeira, a submissão das previsões da Regressão Logística ao teste cego intertemporal (*out-of-sample*) revelou uma dicotomia acentuada entre as competições avaliadas. No Campeonato Brasileiro, a estratégia alcançou um Retorno sobre Investimento (ROI) positivo de 0,78% ao longo de 85 entradas, superando matematicamente a margem de lucro estrutural (*juice* ou *vig*) das operadoras. Em contrapartida, na *Premier League*, o sistema registrou um ROI negativo de -38,35% em 121 entradas, evidenciando que as cotas classificadas pelo algoritmo com valor esperado positivo (+EV) resultaram em perdas severas e contínuas.

Esse contraste de resultados oferece uma perspectiva empírica valiosa sobre a Hipótese do Mercado Eficiente (HME) de Fama (1970). A perda financeira expressiva observada na liga inglesa corrobora a premissa de que mercados com altíssimo volume financeiro e liquidez global operam em um nível de eficiência quase perfeito, incorporando instantaneamente todas as variáveis estatísticas de domínio público nas linhas de fechamento. Por outro lado, o lucro marginal obtido no cenário nacional sugere que o mercado brasileiro, por estar em uma fase de expansão estrutural e apresentar dinâmicas de imprevisibilidade mais acentuadas, ainda abriga pequenas ineficiências de precificação passíveis de exploração por modelos quantitativos rigorosos.

Conclui-se que as variáveis históricas descritivas são insuficientes para gerar lucros sistemáticos em mercados consolidados de eficiência máxima, mas ainda apresentam valor preditivo em competições de eficiência intermediária. Como recomendação para

trabalhos futuros, sugere-se a expansão da engenharia de características com a inclusão de métricas táticas avançadas, como informações de posicionamento espacial (*tracking data*), probabilidade de gols baseada na qualidade da finalização (*Expected Goals* - xG), escalação confirmada dos atletas e dados de carga fisiológica do elenco. A exploração de fontes de informação assimétricas pode ser o caminho adequado para elevar a robustez dos algoritmos perante a precificação dinâmica das casas de apostas.

Referências

BERGMEIR, C.; HYNDMAN, R. J. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, Amsterdam, v. 120, p. 70–83, 2018.

BNL DATA. *Setor de apostas esportivas mais que dobrou no Brasil em 2023*. 2024. Disponível em: <https://bnldata.com.br/setor-de-apostas-esportivas-mais-que-dobrou-no-brasil-em-2023/>. Acesso em: 28 mar. 2026.

BRASIL. *Lei nº 14.790, de 29 de dezembro de 2023. Dispõe sobre a modalidade lotérica aposta de quota fixa e altera as Leis nºs 5.768, de 20 de dezembro de 1971, e 13.756, de 12 de dezembro de 2018*. 2023. Diário Oficial da União: seção 1, Brasília, DF, p. 1.

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

BUNKER, R.; SUSNJAK, T. The application of machine learning techniques for predicting results in team sport: A review. *Applied Sciences*, v. 9, n. 21, p. 1–23, 2019.

CNN BRASIL. *O Maior Raio-X do Torcedor: 79% dos brasileiros têm time; jovens são mais fanáticos*. 2023. Disponível em: <https://www.cnnbrasil.com.br/esportes/outros-esportes/o-maior-raio-x-do-torcedor-79-dos-brasileiros-tem-time-jovens-sao-mais-fanaticos/>. Acesso em: 28 mar. 2026.

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

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. *Applied Logistic Regression*. 3. ed. Hoboken: Wiley, 2013.

JOVEM PAN. *Estudo: apostas esportivas movimentaram entre R$ 60 e R$ 100 bilhões em 2023*. 2024. Disponível em: <https://jovempan.com.br/noticias/brasil/estudo-apostas-esportivas-movimentaram-entre-60-e-100-bilhoes-em-2023.html>. Acesso em: 28 mar. 2026.

KAUFMAN, S. et al. Leakage in data mining: formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, New York, v. 6, n. 4, p. 15:1–15:21, 2012.

MITCHELL, T. M. *Machine Learning*. New York: McGraw-Hill, 1997.

MKT ESPORTIVO. *Mercado de apostas esportivas no Brasil cresce com Mundial de Clubes da FIFA 2025*. 2025. Disponível em: <https://www.mktesportivo.com/2025/07/mercado-de-apostas-esportivas-no-brasil-cresce-com-mundial-de-clubes-da-fifa-2025/>. Acesso em: 28 mar. 2026.

NARAYANA, N. G. S.; DAS, A. *Apparate: Evading Memory Hierarchy with GodSpeed Wireless-on-Chip*. 2024.

PWC. *O impacto das apostas esportivas no consumo*. 2023. Disponível em: <https://www.strategyand.pwc.com/br/pt/relatorios/o-impacto-das-apostas-esportivas-no-consumo.html>. Acesso em: 28 mar. 2026.

TALEB, N. N. *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets*. 2. ed. New York: Random House, 2004.

TURING, A. Computing machinery and intelligence. *Mind*, v. 59, n. 236, p. 433–460, 1950.

UOL. *Pesquisa: Vôlei e F1 são esportes mais acompanhados no Brasil após futebol*. 2024. Disponível em: <https://www.uol.com.br/esporte/futebol/ultimas-noticias/2024/05/14/pesquisa-volei-e-f1-sao-esportes-mais-acompanhados-no-brasil-apos-futebol.htm>. Acesso em: 28 mar. 2026.

Emitido em 09/04/2026

DOCUMENTO Nº 1/2026 - CCDN (11.00.52.09)
(Nº do Documento: 1)

(Nº do Protocolo: NÃO PROTOCOLADO)

(Assinado digitalmente em 22/05/2026 12:10)
ANDREA ALVES BORBA
ASSISTENTE EM ADMINISTRACAO
1517808

Para verificar a autenticidade deste documento entre em https://sipac.ufpb.br/documentos/ informando seu número: **1**, ano: **2026**, documento (espécie): **DOCUMENTO**, data de emissão: **22/05/2026** e o código de verificação: **b9cd75abb3**