UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
BACHARELADO EM ESTATÍSTICA

Danilo Silva Carvalho de Oliveira

Comparação de modelos preditivos para campeonatos de futebol: Uma análise de seis ligas mundiais (2003-2023)

Goiânia
2025

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC no 1240/2014, sem ressarcimento dos direitos autorais, de acordo com a Lei no 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do Trabalho de Conclusão de Curso de Graduação (TCCG)

Nome(s) completo(s) do(a)(s) autor(a)(es)(as): Danilo Silva Carvalho de Oliveira.

Título do trabalho: Comparação de Modelos Preditivos para Campeonatos de Futebol: Uma Análise de Seis Ligas Mundiais (2003-2023).

2. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador) Concorda com a liberação total do documento [x] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante: a) consulta ao(à)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. O documento não será disponibilizado durante o período de embargo.

Casos de embargo:

- Solicitação de registro de patente;

- Submissão de artigo em revista científica;

- Publicação como capítulo de livro.

Obs.: Este termo deve ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Valdivino Vargas Junior**, **Professor do Magistério Superior**, em 28/11/2025, às 19:05, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Danilo Silva Carvalho De Oliveira**, **Discente**, em 07/12/2025, às 11:17, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5822505** e o código CRC **A396475F**.

Referência: Processo nº 23070.059750/2025-26 SEI nº 5822505

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
BACHARELADO EM ESTATÍSTICA

Danilo Silva Carvalho de Oliveira

Comparação de modelos preditivos para campeonatos de futebol: Uma análise de seis ligas mundiais (2003-2023)

Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Estatística da Universidade Federal de Goiás para aprovação no componente curricular TCC, como parte das exigências para a obtenção do título de bacharel em Estatística. **Orientador**: Valdivino Vargas Junior

Goiânia

2025

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Oliveira, Danilo Silva Carvalho de
Comparação de modelos preditivos para campeonatos de futebol [manuscrito] : Uma análise de seis ligas mundiais (2003-2023) / Danilo Silva Carvalho de Oliveira. - 2025.
55 f.

Orientador: Prof. Dr. Valdivino Vargas Junior.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Instituto de Matemática e Estatística (IME), Estatística, Goiânia, 2025.
Bibliografia. Apêndice.
Inclui gráfico, tabelas, lista de figuras, lista de tabelas.

1. Distribuição de Poisson. 2. Simulação de Monte Carlo. 3. Suavização exponencial. 4. Validação cruzada temporal. 5. Análise estatística esportiva. I. Junior, Valdivino Vargas, orient. II. Título.

CDU 519.22

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO

Aos vinte e cinco dias do mês de novembro do ano de 2025 iniciou-se a sessão pública de defesa do Trabalho de Conclusão de Curso (TCC) intitulado “Comparação de Modelos Preditivos para Campeonatos de Futebol: Uma Análise de Seis Ligas Mundiais (2003-2023)”, de autoria de Danilo Silva Carvalho de Oliveira, do curso de Estatística, do Instituto de Matemática e Estatística da UFG. Os trabalhos foram instalados pelo Prof. Dr. Valdivino Vargas Junior com a participação dos demais membros da Banca Examinadora: David Henriques da Matta e (IME/UFG), Mario Ernesto Piscoya Diaz (IME/UFG). Após a apresentação, a banca examinadora realizou a arguição do estudante. Posteriormente, de forma reservada, a Banca Examinadora atribuiu a nota final de 9,6, tendo sido o TCC considerado aprovado.

Proclamados os resultados, os trabalhos foram encerrados e, para constar, lavrou-se a presente ata que segue assinada pelos Membros da Banca Examinadora.

Documento assinado eletronicamente por **Valdivino Vargas Junior**, **Professor do Magistério Superior**, em 27/11/2025, às 10:04, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **David Henriques Da Matta**, **Professor do Magistério Superior**, em 27/11/2025, às 11:55, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Mario Ernesto Piscoya Diaz**, **Professor do Magistério Superior**, em 27/11/2025, às 21:47, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5794731** e o código CRC **F3D5CF68**.

Referência: Processo nº 23070.059750/2025-26 SEI nº 5794731

Agradecimentos

Agradeço primeiramente ao meu orientador, Professor Valdivino Vargas Junior, pela paciência, orientação precisa e confiança depositada neste trabalho. Suas contribuições foram fundamentais para transformar uma ideia inicial em um projeto concreto e academicamente robusto.

À Universidade Federal de Goiás e ao Instituto de Matemática e Estatística, pela formação de excelência e por proporcionar o ambiente acadêmico necessário para o desenvolvimento deste trabalho.

Aos professores do curso de Bacharelado em Estatística, que ao longo desses anos compartilharam não apenas conhecimento técnico, mas também inspiração e paixão pela ciência.

À minha namorada Laura, por todo o amor, apoio incondicional e paciência infinita durante as incontáveis horas dedicadas a este trabalho. Obrigado por compreender os finais de semana perdidos, as conversas interrompidas por "só mais uma simulação", e por fingir interesse quando eu explicava, pela décima vez, a diferença entre o modelo de Poisson e o modelo híbrido. Sua presença tornou esta jornada mais leve e significativa.

Aos meus cachorros, que perderam inúmeras horas de sono ao meu lado durante as madrugadas de programação e escrita, e que certamente não aguentam mais ouvir as palavras "simulação"e "futebol"na mesma frase. Obrigado pela companhia silenciosa e pelos olhares de incompreensão que, paradoxalmente, me mantiveram focado.

Aos meus pais e familiares, pelo apoio constante e por acreditarem na importância da educação e do conhecimento.

Aos colegas de curso, pelas discussões enriquecedoras, pelo apoio mútuo e pela amizade construída ao longo desta jornada acadêmica.

Por fim, agradeço a todos que, direta ou indiretamente, contribuíram para a realização deste trabalho e para minha formação como estatístico.

“Ano passado eu morri, mas esse ano eu não morro”
(Belchior, Sujeito de Sorte, 1976)

Resumo

Este trabalho compara sistematicamente três abordagens de modelagem preditiva para campeonatos de futebol: (i) modelo puramente quantitativo baseado em Poisson, (ii) modelo puramente categórico baseado em perfis de desempenho, e (iii) modelo híbrido que combina ambas as abordagens. A análise foi aplicada a seis das principais ligas mundiais (Brasil, Inglaterra, Espanha, Itália, Alemanha e França) ao longo de 21 temporadas (2003 a 2023), totalizando 126 temporadas-campeonato e 46.503 partidas. Utilizando validação cruzada temporal com 5 *folds*, o parâmetro de memória h foi otimizado independentemente para cada modelo, liga e temporada. Os resultados revelaram melhor desempenho do modelo de Poisson Puro para o conjunto de dados analisado, que apresentou as melhores métricas em todas as seis ligas (MAE global: $3{,}37$ versus $3{,}43$ dos modelos categóricos) e maior acurácia categórica ($82{,}5\%$ de acerto de campeões versus $79{,}4\%$), validando empiricamente a adequação da distribuição de Poisson para modelar gols no futebol dentro do escopo desta pesquisa. Todas as combinações apresentaram alta inércia de desempenho ($h > 0{,}75$), indicando que o histórico acumulado é mais informativo que a forma recente. Identificaram-se diferenças sistemáticas entre ligas: Alemanha mais previsível (MAE = $2{,}96$), Inglaterra menos previsível (MAE = $3{,}71$). A análise de agrupamento hierárquico identificou três grupos distintos de ligas com dinâmicas similares, demonstrando que inércia e previsibilidade são dimensões parcialmente independentes. Como aplicação prática, embora o modelo de Poisson Puro tenha apresentado métricas superiores, o modelo Híbrido foi utilizado para gerar previsões dos Campeonatos Brasileiros Série A e B de 2025 por sua completude metodológica em gerar placares realistas enquanto captura as dinâmicas de resultado (V/E/D). A aplicação indicou uma disputa equilibrada pelo título da Série A entre Flamengo (46%) e Palmeiras (43%) e o favoritismo do Coritiba na Série B (72% de título), além de detalhar as 35 combinações de acesso possíveis nesta última.

Palavras-chave: Distribuição de Poisson. Simulação de Monte Carlo. Suavização exponencial. Validação cruzada temporal. Análise estatística esportiva.

Abstract

This study systematically compares three predictive modeling approaches for football championships: (i) a purely quantitative model based on Poisson distribution, (ii) a purely categorical model based on performance profiles, and (iii) a hybrid model combining both approaches. The analysis was applied to six major world leagues (Brazil, England, Spain, Italy, Germany, and France) over 21 seasons (2003 to 2023), totaling 126 season-championships and 46,503 matches. Using temporal cross-validation with 5 *folds*, the memory parameter h was optimized independently for each model, league, and season. The results revealed better performance of the Pure Poisson model for the analyzed dataset, which presented the best metrics across all six leagues (global MAE: 3.37 versus 3.43 for categorical models) and higher categorical accuracy (82.5% champion prediction versus 79.4%), empirically validating the adequacy of the Poisson distribution for modeling goals in football within the scope of this research. All combinations showed high performance inertia ($h > 0.75$), indicating that accumulated history is more informative than recent form. Systematic differences between leagues were identified: Germany more predictable (MAE = 2.96), England less predictable (MAE = 3.71). Hierarchical clustering analysis identified three distinct groups of leagues with similar dynamics, demonstrating that inertia and predictability are partially independent dimensions. As a practical application, although the Pure Poisson model presented superior metrics, the Hybrid model was used to generate predictions for the 2025 Brazilian Championships Series A and B due to its methodological completeness in generating realistic scores while capturing match outcome dynamics (W/D/L). The application indicated a balanced title dispute in Series A between Flamengo (46%) and Palmeiras (43%) and Coritiba's favoritism in Series B (72% title probability), in addition to detailing the 35 possible promotion combinations in the latter.

Keywords: Poisson distribution. Monte Carlo simulation. Exponential smoothing. Temporal cross-validation. Sports analytics.

Lista de figuras

Lista de tabelas

Sumário

Introdução

O futebol é um dos fenômenos esportivos mais complexos de modelar estatisticamente. Sua natureza fluida, a baixa frequência de gols e a alta influência de fatores aleatórios tornam cada partida um evento único, mas não completamente imprevisível. É neste contexto, entre o caos aparente e os padrões subjacentes, que este trabalho se insere.

A análise quantitativa no esporte, conhecida como *sports analytics*, revolucionou a forma como compreendemos e prevemos desempenhos esportivos. No futebol, esta revolução começou com o trabalho pioneiro de Maher (Maher, 1982), que demonstrou que a ocorrência de gols pode ser modelada através de distribuições de probabilidade, especificamente a distribuição de Poisson. Desde então, diversos pesquisadores refinaram esta abordagem. Dixon e Coles (Dixon; Coles, 1997) introduziram o conceito de parâmetros dinâmicos, permitindo que o modelo capture mudanças no desempenho das equipes ao longo de uma temporada. Trabalhos mais recentes, como os de Ramos, Lemos e Batista (Ramos; Lemos; Batista, 2019) e Kuhnert e Possato (Kuhnert; Possato, 2023), aplicaram e adaptaram estas técnicas a diferentes contextos competitivos, demonstrando sua versatilidade e robustez.

Apesar desses avanços significativos, uma questão fundamental permanece em aberto: qual abordagem de modelagem é mais eficaz para prever resultados no futebol? Modelos puramente quantitativos, baseados em distribuições de Poisson para gols, ou modelos categóricos, que trabalham diretamente com probabilidades de vitória, empate e derrota? Seria possível combinar ambas as abordagens em um modelo híbrido que capture as vantagens de cada uma? Adicionalmente, as dinâmicas competitivas seriam universais, ou cada liga possui características próprias que afetam tanto a previsibilidade quanto o desempenho relativo de diferentes modelos? Estas questões motivam a necessidade de uma análise comparativa sistemática que avalie rigorosamente diferentes abordagens em múltiplos contextos competitivos.

Neste contexto, o presente trabalho tem como objetivo principal comparar sistematicamente três abordagens de modelagem preditiva para campeonatos de futebol: modelo puramente quantitativo baseado em Poisson, modelo puramente categórico baseado em perfis de desempenho, e modelo híbrido que combina ambas as abordagens. Esta comparação é realizada sobre seis das principais ligas mundiais (Brasil, Inglaterra, Espanha, Itália, Alemanha e França) ao longo de 21 temporadas (2003 a 2023), totalizando 126 temporadas-campeonato e 46.503 partidas analisadas.

Para alcançar este objetivo, o trabalho desenvolve e implementa as três variantes de modelo, todas utilizando suavização exponencial para atualização dinâmica de parâmetros. O parâmetro de memória h, que determina o peso dado ao histórico acumulado versus resultados recentes, é otimizado para cada modelo, liga e temporada através de validação cruzada temporal. O

desempenho preditivo dos três modelos é então comparado em termos de erro médio absoluto (MAE), erro quadrático médio (RMSE) e acurácia categórica (identificação de campeões, top 4 e rebaixados), permitindo identificar em quais contextos cada abordagem se destaca. As diferenças nas dinâmicas competitivas entre as seis ligas são analisadas, caracterizando-as em termos de inércia de desempenho e previsibilidade. Por fim, o modelo híbrido é aplicado para gerar previsões das classificações finais dos Campeonatos Brasileiros Série A e Série B de 2025, demonstrando a aplicabilidade prática da metodologia desenvolvida.

A principal contribuição deste trabalho é fornecer evidências empíricas sobre a eficácia relativa de diferentes abordagens de modelagem no futebol, em um escopo sem precedentes na literatura. A análise sistemática de 126 temporadas-campeonato permite não apenas identificar qual modelo é superior, mas também compreender por que e em quais contextos cada abordagem se destaca. Os resultados revelam que, embora o modelo de Poisson seja consistentemente superior, as diferenças são modestas, e que características específicas de cada liga (equilíbrio competitivo, frequência de empates, vantagem de mando de campo) influenciam significativamente tanto a previsibilidade quanto o desempenho relativo dos modelos.

Este trabalho está organizado da seguinte forma: o Capítulo 1 apresenta a revisão bibliográfica, cobrindo os fundamentos teóricos de probabilidade, teoremas limite, análise de agrupamentos e os principais modelos de previsão no futebol. O Capítulo 2 detalha a metodologia empregada, incluindo coleta de dados, arquitetura dos três modelos, métricas de avaliação e processo de validação cruzada temporal. O Capítulo 3 apresenta e discute os resultados da otimização de parâmetros, comparação entre modelos, análise de agrupamento hierárquico das ligas e aplicação prática para previsão de 2025. Por fim, a Conclusão sintetiza os principais achados, discute limitações e aponta direções para trabalhos futuros.

1 Revisão Bibliográfica

Este capítulo apresenta na Seção 1.1 os fundamentos teóricos e matemáticos do trabalho. Na Seção 1.2, discutimos os teoremas limite que fundamentam as simulações de Monte Carlo. A Seção 1.3 introduz a técnica de análise de aglomerados. Por fim, a Seção 1.4 revisa os principais modelos existentes para previsão de resultados no futebol.

1.1 Modelo Matemático para um Experimento (Modelo Probabilístico)

A teoria moderna da probabilidade, que confere rigor matemático ao estudo de fenômenos aleatórios, é fundamentada na abordagem axiomática proposta por Andrey Kolmogorov na década de 1930 (Kolmogorov, 1956). A estrutura central dessa teoria é o espaço de probabilidade.

1.1.1 Espaço Amostral e Eventos

Definição 1.1.1 *Seja Ω o conjunto de todos os resultados possíveis de um experimento aleatório, denominado espaço amostral. Todo subconjunto $A \subseteq \Omega$ é chamado de evento. Em particular, Ω é o evento certo e o conjunto vazio $\emptyset$ é o evento impossível.*

1.1.2 σ-álgebra

Definição 1.1.2 *Uma coleção $\mathcal{F}$ de subconjuntos de Ω é uma σ-álgebra se satisfaz as seguintes propriedades:*

1. $\Omega \in \mathcal{F}$.
2. *Se $A \in \mathcal{F}$, então seu complemento $A^c \in \mathcal{F}$.*
3. *Se $A_1, A_2, \ldots$ é uma sequência contável de eventos em $\mathcal{F}$, então a união $\cup_{i=1}^{\infty} A_i$ também pertence a $\mathcal{F}$.*

O par $(\Omega, \mathcal{F})$ é chamado de espaço mensurável (Kolmogorov, 1956; Ross, 2014).

1.1.3 Medida de Probabilidade

Definição 1.1.3 *Uma medida de probabilidade P é uma função $P : \mathcal{F} \to [0, 1]$ que satisfaz os três Axiomas de Kolmogorov:*

1. ***Não-negatividade:*** *Para qualquer evento* $A \in \mathcal{F}$, $P(A) \geq 0$.

2. ***Normalização:*** $P(\Omega) = 1$.

3. ***Aditividade Contável:*** *Para qualquer sequência de eventos* $A_1, A_2, \ldots$ *em* $\mathcal{F}$ *que sejam mutuamente exclusivos (disjuntos),*

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i).$$

1.1.4 Espaço de Probabilidade

Definição 1.1.4 *Um espaço de probabilidade é uma tríade* $(\Omega, \mathcal{F}, P)$*, onde* Ω *é um espaço amostral,* $\mathcal{F}$ *é uma* σ*-álgebra de subconjuntos de* Ω*, e* P *é uma medida de probabilidade sobre* $\mathcal{F}$*.*

Exemplo 1.1.1 (Aplicação ao Futebol) *Considere uma temporada do Campeonato Brasileiro com 380 partidas. Os componentes são:*

- ***Espaço Amostral (*** Ω***):*** *O conjunto de todos os resultados possíveis das 380 partidas, onde cada partida pode ter 3 resultados (vitória casa, empate, vitória visitante). Logo,* $|\Omega| = 3^{380} \approx 2{,}02 \times 10^{181}$.

- ***σ-álgebra (*** $\mathcal{F}$***):*** *O conjunto de todos os subconjuntos de* Ω *que representam eventos de interesse (e.g., "Flamengo campeão", "Corinthians rebaixado").*

- ***Medida de Probabilidade (*** P***):*** *Atribui probabilidades a cada evento, respeitando os axiomas de Kolmogorov.*

1.2 Convergência e Simulação: Teoremas Limite

Conforme ilustrado no exemplo da Seção 1.1, o espaço amostral de uma temporada completa de futebol é astronomicamente grande ($|\Omega| \approx 10^{181}$). Esta magnitude torna computacionalmente intratável o cálculo analítico direto de probabilidades para eventos complexos, como a chance de um time ser campeão, pois é impossível enumerar todos os cenários possíveis.

Assim, para analisar modelos tão complexos, recorre-se a teoremas limite e métodos de simulação (como o Monte Carlo), que formam a ponte entre a teoria e a prática computacional. Os teoremas a seguir fornecem a garantia matemática de que, ao simular um número suficiente de temporadas, os resultados médios convergem para o valor esperado real.

1.2.1 Lei Fraca dos Grandes Números

A Lei Fraca dos Grandes Números afirma que, para uma sequência de variáveis aleatórias $X_1, X_2, \ldots, X_n$ independentes e identicamente distribuídas (i.i.d.), com valor esperado finito $E(X_i) = \mu$ e variância σ^2, a média amostral $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converge em probabilidade para a média populacional μ.

Matematicamente, para qualquer número positivo ϵ, por menor que seja, temos:

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \epsilon) = 1.$$

1.2.2 Lei Forte dos Grandes Números

A Lei Forte dos Grandes Números estabelece que, sob as mesmas condições de independência e distribuição idêntica com valor esperado finito $E(X_i) = \mu$, a média amostral $\bar{X}_n$ converge quase certamente para a média populacional μ.

A convergência quase certa é uma condição mais forte que a convergência em probabilidade e é expressa como:

$$P\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1.$$

Isso implica que a probabilidade do evento em que a média amostral não converge para a média populacional é zero. Em outras palavras, com probabilidade 1, a média das observações irá convergir para o valor esperado à medida que o número de observações tende ao infinito.

1.2.3 Método de Monte Carlo (MMC)

O Método de Monte Carlo é uma classe de algoritmos computacionais que utiliza a amostragem aleatória massiva para obter resultados numéricos. Fundamentado na Lei Forte dos Grandes Números, o método consiste em:

1. Definir um modelo do sistema com variáveis de entrada regidas por distribuições de probabilidade.

2. Gerar um grande número de cenários (realizações) por meio de amostragem aleatória dessas variáveis.

3. Agregar os resultados de todos os cenários para estimar a quantidade de interesse, como uma média, uma probabilidade ou uma distribuição de resultados.

A Lei Forte dos Grandes Números garante que, à medida que o número de simulações aumenta, a estimativa do MMC converge para o valor verdadeiro.

1.2.4 Teorema Central do Limite (TCL)

Teorema 1.2.1 *O Teorema Central do Limite afirma que, sob condições gerais, a distribuição da soma (ou média) de um grande número de variáveis aleatórias independentes e identicamente distribuídas se aproxima de uma distribuição Normal, independentemente da distribuição original das variáveis. Se* $S_n = X_1 + \ldots + X_n$*, com* $E[X_i] = \mu$ *finita e Var*$(X_i) = \sigma^2$ *finita e não nula, então a variável padronizada* Z_n*:*

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1) \text{ quando } n \to \infty$$

O TCL é crucial pois permite quantificar a incerteza da estimativa de Monte Carlo, possibilitando a construção de intervalos de confiança.

1.3 Análise de Agrupamentos

A Análise de Agrupamentos (*cluster analysis*) compreende um conjunto de métodos de aprendizagem não supervisionada cujo objetivo é particionar um conjunto de observações em subgrupos (aglomerados ou *clusters*) de forma que observações dentro de um mesmo grupo possuam alta similaridade entre si, enquanto observações em grupos distintos possuam baixa similaridade (Hastie; Tibshirani; Friedman, 2009). O método não requer conhecimento prévio sobre os rótulos das classes, descobrindo estruturas inerentes aos dados.

Formalmente, dado um conjunto de dados $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ onde $\mathbf{x}_i \in \mathbb{R}^p$, o objetivo é encontrar uma partição $C = \{C_1, \ldots, C_k\}$ que otimize um critério pré-definido de homogeneidade intra-grupo e heterogeneidade inter-grupos.

1.3.1 Métodos Hierárquicos

Os métodos hierárquicos constroem uma hierarquia de agrupamentos, representada por um dendrograma, que mostra as relações de similaridade entre observações e grupos em diferentes níveis de agregação. Existem duas abordagens principais:

- **Aglomerativa (bottom-up):** Inicia com cada observação como um grupo individual e progressivamente une os grupos mais similares até que todos estejam em um único grupo.
- **Divisiva (top-down):** Inicia com todas as observações em um único grupo e progressivamente divide em grupos menores.

O método aglomerativo é mais comum e foi utilizado neste trabalho. O algoritmo básico consiste em:

1. Calcular a matriz de distâncias D entre todas as observações;

2. Unir os dois grupos mais próximos em um novo grupo;
3. Recalcular as distâncias do novo grupo aos demais;
4. Repetir até que reste apenas um grupo.

1.3.2 Métodos de Ligação

A escolha de como calcular a distância entre grupos (método de ligação) afeta significativamente o resultado. Os principais métodos incluem:

- **Ligação simples (*single linkage*):** Distância mínima entre quaisquer dois pontos dos grupos;
- **Ligação completa (*complete linkage*):** Distância máxima entre quaisquer dois pontos dos grupos;
- **Ligação média (*average linkage*):** Média das distâncias entre todos os pares de pontos;
- **Método de Ward:** Minimiza a variância intra-grupo ao unir clusters.

1.3.3 Método de Ward

O método de Ward (Jr, 1963) é particularmente eficaz para identificar grupos compactos e bem separados. Em cada etapa, une-se o par de grupos que resulta no menor aumento da soma total de quadrados intra-grupos (WCSS):

$$\text{WCSS} = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 . \tag{1.1}$$

onde $\boldsymbol{\mu}_i$ é o centroide do grupo C_i. Este critério tende a produzir grupos de tamanhos similares e é robusto a outliers.

A Análise de Agrupamentos é amplamente utilizada em análise exploratória de dados para descoberta de padrões, segmentação de mercado, taxonomia biológica e, no contexto deste trabalho, para identificar grupos de ligas de futebol com dinâmicas competitivas similares.

1.4 Modelagem Preditiva no Futebol: Uma Revisão

Esta seção revisa os trabalhos fundamentais que estabelecem a base para a modelagem de resultados no futebol, partindo dos modelos estocásticos clássicos e avançando para as abordagens dinâmicas e computacionais que formam o estado da arte.

1.4.1 A Distribuição de Poisson como Geradora de Gols

A distribuição de Poisson é uma distribuição de probabilidade discreta que modela o número de eventos que ocorrem em um intervalo fixo de tempo ou espaço, quando esses eventos ocorrem com uma taxa média constante e independentemente do tempo desde o último evento. A probabilidade de observar exatamente k eventos é dada por:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots \tag{1.2}$$

onde $\lambda > 0$ é o parâmetro que representa a taxa média de ocorrência dos eventos.

Esta distribuição é particularmente adequada para modelar gols no futebol porque:

- Gols são eventos relativamente raros durante os 90 minutos de jogo;
- A probabilidade de marcar um gol em qualquer instante é aproximadamente constante.

O trabalho pioneiro de Maher (Maher, 1982) foi o primeiro a formalizar e validar empiricamente o uso da distribuição de Poisson para modelar resultados no futebol. Seu modelo assume que os gols marcados pelos times mandante e visitante seguem distribuições de Poisson independentes:

$$X_{\text{casa}} \sim \text{Poisson}(\lambda_{\text{casa}}), \tag{1.3}$$

$$X_{\text{visitante}} \sim \text{Poisson}(\lambda_{\text{visitante}}). \tag{1.4}$$

As taxas esperadas de gols são determinadas pelas forças de ataque (α) e defesa (β) das equipes, além do fator casa (γ):

$$\log(\lambda_{\text{casa}}) = \mu + \alpha_i - \beta_j + \gamma, \tag{1.5}$$

$$\log(\lambda_{\text{visitante}}) = \mu + \alpha_j - \beta_i. \tag{1.6}$$

A principal limitação deste modelo é sua natureza estática: os parâmetros de força das equipes permanecem constantes durante toda a temporada, ignorando variações de desempenho ao longo do tempo.

1.4.2 Modelo Dinâmico de Dixon e Coles

Dixon e Coles (Dixon; Coles, 1997) aprimoraram o modelo de Maher abordando duas limitações principais:

1.4.2.1 Correção para Placares Baixos

Os autores observaram que placares como 0-0, 1-0, 0-1 e 1-1 eram sub-representados pelo modelo de Poisson independente. Introduziram um parâmetro de correção ρ que ajusta especificamente as probabilidades desses placares, mantendo os demais inalterados.

1.4.2.2 Ponderação Temporal

A principal inovação foi tornar o modelo dinâmico através de ponderação exponencial. Para refletir a natureza dinâmica do desempenho, construíram uma "pseudoverossimilhança" onde resultados recentes recebem maior peso na estimação dos parâmetros:

$$\chi(t) = \exp(-\xi \cdot t), \tag{1.7}$$

onde t é o tempo decorrido desde a partida e ξ controla a taxa de "esquecimento" dos resultados antigos. Esta função exponencial permite que todos os resultados anteriores sejam incluídos no cálculo, mas os dados históricos mais antigos são desvalorizados exponencialmente.

1.4.3 Fundamentos da Suavização Exponencial

A Suavização Exponencial, formalizada por Holt (1957) e Winters (1960), é a técnica fundamental que permite aos modelos dinâmicos incorporarem variação temporal. A forma básica atualiza estimativas através de uma média ponderada:

$$\hat{X}_{t+1} = \alpha \cdot X_t + (1 - \alpha) \cdot \hat{X}_t, \tag{1.8}$$

onde $\alpha \in [0,1]$ é o parâmetro de suavização. Equivalentemente, usando o parâmetro de memória $h = 1 - \alpha$:

$$\hat{X}_{t+1} = h \cdot \hat{X}_t + (1 - h) \cdot X_t. \tag{1.9}$$

O parâmetro controla o equilíbrio entre estabilidade e reatividade:

- h próximo de 0 (memória curta): modelo reativo a resultados recentes;
- h próximo de 1 (memória longa): modelo considera histórico mais amplo.

1.4.4 Outras Abordagens na Literatura

Paralelamente aos modelos de Poisson, outras linhas de pesquisa exploraram abordagens categóricas e métodos alternativos de atualização dinâmica. No contexto brasileiro, destacam-se:

1.4.4.1 Modelo UFMG

Lima et al. (Lima *et al.*, 2010) desenvolveram no Departamento de Matemática da UFMG um modelo que trabalha diretamente com probabilidades de vitória, empate e derrota, sem modelar explicitamente os gols.

O modelo mantém para cada time dois vetores de probabilidades:

- $P_C = (PV_C, PE_C, PD_C)$ – perfil como mandante;
- $P_F = (PV_F, PE_F, PD_F)$ – perfil como visitante.

Na primeira rodada, todos os times são considerados igualmente fortes. Os vetores são então atualizados progressivamente após cada partida, considerando o rendimento do adversário – vitórias contra times fortes têm maior peso que vitórias contra times fracos. A principal contribuição do modelo é sua capacidade de, ao focar em probabilidades categóricas, capturar dinâmicas de resultado (especialmente empates) de forma mais eficaz do que os modelos de Poisson puros.

1.4.4.2 Método PROFMAT

Ramos, Lemos e Batista (Ramos; Lemos; Batista, 2019) propuseram uma metodologia no âmbito do Mestrado Profissional em Matemática (PROFMAT/UFSJ) que utiliza Poisson com uma estrutura de vetores específica.

Cada time possui dois vetores de médias de gols:

- Mandante: (GFM, GSM) – gols feitos e sofridos em casa;
- Visitante: (GFV, GSV) – gols feitos e sofridos fora.

Para prever um jogo entre time A (casa) e time B (fora), calculam-se as médias:

$$\lambda_A = \frac{GFM_A + GSV_B}{2}, \tag{1.10}$$

$$\lambda_B = \frac{GFV_B + GSM_A}{2}. \tag{1.11}$$

Os vetores são atualizados iterativamente. Na rodada r, a nova média incorpora todos os jogos anteriores:

$$GFM_{\text{novo}} = \frac{GFM_{\text{inicial}} + \sum_{i=1}^{r} \text{gols feitos em casa}}{r+1}. \tag{1.12}$$

Notavelmente, esta atualização representa uma média móvel simples, onde todos os jogos passados têm peso idêntico. Esta abordagem difere fundamentalmente da Suavização

Exponencial (discutida na Seção 1.4.3), que aplica pesos dinâmicos para priorizar a informação recente.

Embora a abordagem utilize simulações Monte Carlo para estimar a classificação final, os autores reportam desafios significativos na modelagem de empates, uma limitação comum de modelos baseados em Poisson independente que não utilizam correções (como a de Dixon e Coles (Dixon; Coles, 1997)).

1.4.5 Aplicações Recentes no Contexto Brasileiro

Kuhnert e Possato (Kuhnert; Possato, 2023) implementaram uma versão do modelo de Poisson com atualização dinâmica usando Média Móvel Exponencial. Através de *backtest*, encontraram parâmetros ótimos para o futebol brasileiro: correção Rho ($\rho = 0{,}07$) para ajustar previsões de empates e período média móvel exponencial (MME) de 17,4. Este alto fator de retenção (aproximadamente 89% do valor anterior) sugere que o campeonato brasileiro tem maior inércia competitiva comparado a ligas europeias, onde atualizações mais rápidas são necessárias.

Estes trabalhos demonstram que, embora os modelos de Poisson sejam eficazes para prever tendências gerais e resultados de longo prazo no futebol brasileiro, ainda existem desafios significativos na previsão de resultados específicos, particularmente empates e vitórias visitantes.

De fato, a revisão da literatura no contexto brasileiro revela uma dicotomia clara nas abordagens. De um lado, modelos puramente categóricos, como o da UFMG (Lima *et al.*, 2010), demonstram superioridade na captura de padrões de resultado, especialmente empates. Contudo, sua limitação é não gerar placares específicos.

Do outro lado, modelos baseados na distribuição de Poisson, como o Método PROFMAT (Ramos; Lemos; Batista, 2019), geram placares, mas, como revisado, reportam desafios significativos na modelagem de empates.

Este trabalho se posiciona exatamente nesta lacuna. Ao propor e comparar três arquiteturas (Poisson Puro, Perfil Categórico e Híbrido), o objetivo é investigar o desempenho relativo de cada uma. Será avaliado como o modelo híbrido, que combina a determinação categórica do resultado (Etapa 1) com a geração de placar condicional via Poisson (Etapa 2), se posiciona em relação às vantagens e desvantagens de cada abordagem pura: a precisão probabilística do modelo categórico e a capacidade de geração de placares do modelo de Poisson.

2 Metodologia

Este capítulo detalha o *framework* metodológico completo empregado para desenvolver, validar e aplicar os modelos preditivos. A abordagem foi estruturada de forma sequencial: a Seção 2.1 descreve a coleta e preparação dos dados; a Seção 2.2 apresenta a arquitetura dos três modelos de simulação desenvolvidos; a Seção 2.3 define as métricas de avaliação utilizadas; a Seção 2.4 detalha o processo de validação cruzada temporal e otimização de parâmetros – o núcleo metodológico deste trabalho; a Seção 2.5 descreve a análise comparativa entre ligas; a Seção 2.7 explica a aplicação prática do modelo; e a Seção 2.8 discute as limitações metodológicas e considerações éticas.

2.1 Coleta e Preparação dos Dados

Os dados históricos necessários para o desenvolvimento do modelo foram obtidos do portal Transfermarkt (<www.transfermarkt.com>), uma das mais completas bases de dados sobre futebol mundial. O período de análise compreendeu 21 temporadas consecutivas, de 2003 a 2023, abrangendo seis das principais ligas de futebol: Campeonato Brasileiro (Série A), Premier League (Inglaterra), La Liga (Espanha), Serie A (Itália), Bundesliga (Alemanha) e Ligue 1 (França).

2.1.1 Processo de Extração

Foi desenvolvido um *web scraper* automatizado em linguagem R (R Core Team, 2018) para realizar a coleta sistemática dos dados. O *script* de extração foi projetado para navegar pelas páginas de cada campeonato e temporada, extraindo informações estruturadas de cada partida realizada.

Os dados de cada partida foram estruturados nas seguintes variáveis-chave, coletadas via *web scraper*:

Tabela 1 – Estrutura do conjunto de dados por partida.

Variável	**Descrição**
`season`	Ano de início da temporada (e.g., 2003, 2004, ...).
`date`	Data exata da realização da partida (para ordenação cronológica).
`home`	Nome da equipe mandante.
`away`	Nome da equipe visitante.
`home_goal`	Número de gols marcados pela equipe mandante.
`away_goal`	Número de gols marcados pela equipe visitante.

Fonte: Elaborado pelo autor

2.1.2 Preparação e Limpeza dos Dados

Os dados coletados passaram por um processo de limpeza e padronização para garantir a consistência e qualidade das análises subsequentes. As principais etapas incluíram:

1. **Validação de integridade:** Verificação de valores ausentes, duplicatas e inconsistências nos placares;
2. **Padronização de nomes:** Unificação dos nomes das equipes ao longo das temporadas, considerando mudanças de nomenclatura;
3. **Ordenação cronológica:** Estruturação rigorosa dos dados pela data real de realização da partida. Esta etapa é de importância crítica para a integridade da série temporal. É comum que partidas sejam adiadas por diversos motivos, mas permaneçam listadas nas bases de dados em suas rodadas originais. A ordenação pela data efetiva garante que o processo de atualização sequencial dos parâmetros (Equação 2.1) reflita a ordem real em que os eventos ocorreram, tratando jogos adiados apenas em rodadas futuras, quando foram de fato disputados.
4. **Cálculo de rodadas:** Após a ordenação cronológica, foi determinado o número da rodada para cada partida, com base no formato de cada campeonato. Esta etapa é necessária para a divisão dos folds na validação cruzada e para a definição do ponto de corte da previsão final.

2.1.3 Análise descritiva do conjunto de dados

O conjunto de dados final compreende 126 observações de temporada-campeonato (6 ligas $\times$ 21 temporadas), totalizando 46.503 partidas analisadas. A Tabela 2 apresenta as estatísticas descritivas agregadas de cada liga ao longo do período estudado.

Tabela 2 – Estatísticas agregadas por campeonato (2003–2023).

Campeonato	Temp.	Jogos Totais	Gols Casa	Gols Fora	Gols Total	Vit. Casa (%)	Vit. Fora (%)	Emp. (%)	Média Times
Brasileiro	21	8.406	1,54	1,03	2,57	49,7	23,9	26,4	20,6
Alemão	21	6.426	1,66	1,27	2,93	45,7	29,6	24,8	18,0
Espanhol	21	7.980	1,53	1,13	2,66	47,0	28,1	25,0	20,0
Francês	21	7.879	1,42	1,04	2,46	45,0	27,0	28,0	20,0
Inglês	21	7.980	1,53	1,16	2,69	46,0	29,3	24,7	20,0
Italiano	21	7.832	1,51	1,17	2,68	44,9	28,3	26,8	19,8

Fonte: Elaborado pelo autor

Na Tabela 2, as colunas Gols Casa, Gols Fora e Gols Total expressam as médias de gols por partida. Já os valores não inteiros na média de times para o Brasil (20,6) e para a Itália

(19,8) decorrem de variações no formato das competições durante o período: o campeonato brasileiro contou com 24 times em 2003-2004 e 22 em 2005, enquanto o italiano teve 18 times na temporada 2003-2004.

2.2 Arquitetura dos Modelos Preditivos

Este trabalho desenvolveu e comparou três arquiteturas de modelo distintas, todas compartilhando o mesmo mecanismo de atualização de parâmetros via suavização exponencial, mas diferindo fundamentalmente na forma como o resultado da partida é gerado. Esta abordagem comparativa permite avaliar a contribuição relativa de diferentes componentes do modelo.

Uma característica distintiva de todos os modelos é a não utilização de um parâmetro de mando de campo global único, como proposto originalmente por Maher (Maher, 1982). No modelo clássico de Maher (Equações 1.5 e 1.6), o parâmetro γ representa o efeito médio de jogar em casa, aplicado uniformemente a todas as equipes.

Neste trabalho, o efeito do mando de campo foi modelado individualmente para cada equipe, por meio de parâmetros de desempenho distintos para quando jogam em casa (mandante) e fora de casa (visitante). Esta escolha metodológica, inspirada na abordagem do modelo UFMG (Lima *et al.*, 2010), reconhece que a vantagem de jogar em casa é uma característica intrínseca e heterogênea de cada time, não uma constante universal do campeonato. Assim, o modelo pode capturar, por exemplo, que o Flamengo tem maior diferença de desempenho entre casa e fora do que outros times, sem impor esta estrutura *a priori*.

2.2.1 Formulação Matemática dos Parâmetros Dinâmicos

O desempenho de cada time i no campeonato foi rastreado por um conjunto de seis parâmetros dinâmicos, que são atualizados sequencialmente após cada partida disputada. Estes parâmetros capturam tanto aspectos categóricos (probabilidades de resultado) quanto quantitativos (capacidade de marcar e sofrer gols).

2.2.1.1 Perfis de Desempenho (Componente Categórico)

Para modelar a probabilidade de cada resultado possível (Vitória, Empate ou Derrota), foram definidos dois vetores de perfil categórico para cada time i:

- $\mathbf{PM}_i = [p_v^m, p_e^m, p_d^m]$: perfil de desempenho como mandante;
- $\mathbf{PV}_i = [p_v^v, p_e^v, p_d^v]$: perfil de desempenho como visitante.

Onde p_v, p_e e p_d representam as probabilidades estimadas de vitória, empate e derrota naquela condição, respectivamente, satisfazendo a restrição $p_v + p_e + p_d = 1$.

2.2.1.2 Parâmetros de Poisson (Componente Quantitativo)

Para modelar a capacidade de marcar e sofrer gols, foram definidos quatro parâmetros de taxa (forças) para cada time i, representando as médias esperadas de gols em cada condição:

- λ_i^m : força de **ataque** como mandante (taxa esperada de gols marcados em casa);
- μ_i^m : força de **defesa** como mandante (taxa esperada de gols sofridos em casa);
- λ_i^v : força de **ataque** como visitante (taxa esperada de gols marcados fora);
- μ_i^v : força de **defesa** como visitante (taxa esperada de gols sofridos fora).

2.2.1.3 Atualização via Suavização Exponencial

Todos os seis parâmetros foram atualizados após cada jogo t usando Suavização Exponencial Simples (SES), conforme descrito na Seção 1.4.3 do Capítulo 1. A atualização é controlada por um único parâmetro de suavização $h \in [0,1]$ e segue a fórmula:

$$\text{Parâmetro}_{novo} = h \cdot \text{Parâmetro}_{anterior} + (1-h) \cdot \text{Observação}_t. \tag{2.1}$$

Onde h representa o fator de memória (o peso dado ao histórico acumulado) e $(1-h)$ representa o fator de aprendizado (o peso dado à nova observação). Um valor de h próximo de 1 (memória longa) torna o modelo mais conservador e estável, privilegiando o histórico acumulado. Um valor de h próximo de 0 (memória curta) torna o modelo mais reativo e adaptativo aos resultados recentes.

Neste trabalho, o parâmetro h é interpretado como um índice de persistência de desempenho: valores altos de h indicam que o desempenho passado tende a se manter no futuro (alta persistência), enquanto valores baixos indicam que o desempenho é mais volátil e sujeito a mudanças rápidas (baixa persistência). Esta interpretação permite caracterizar as dinâmicas competitivas de cada liga em termos de quão estável ou volátil é o desempenho das equipes ao longo de uma temporada.

Para os parâmetros de Poisson, a observação no tempo t corresponde aos gols efetivamente marcados ou sofridos. Para os perfis categóricos, a observação é um vetor indicador: $(1,0,0)$ para vitória, $(0,1,0)$ para empate, e $(0,0,1)$ para derrota.

2.2.2 Definição dos Três Modelos de Simulação

Três processos distintos de simulação de jogo foram implementados e sistematicamente comparados:

1. **Modelo 1 (Poisson Puro):** Um modelo puramente quantitativo baseado na distribuição de Poisson. Para uma partida entre time A (mandante) e time B (visitante), as taxas de gol esperadas são calculadas pela média aritmética das forças de ataque e defesa dos oponentes:

$$\lambda_A = \frac{\lambda_A^m + \mu_B^v}{2}, \tag{2.2}$$

$$\lambda_B = \frac{\lambda_B^v + \mu_A^m}{2}. \tag{2.3}$$

Foi escolhida uma formulação de média aritmética (modelo aditivo) pela sua simplicidade e interpretabilidade direta, onde a taxa de gols esperada representa um equilíbrio exato entre a força de ataque de uma equipe e a força de defesa da outra, diferindo das abordagens multiplicativas mais clássicas (Maher, 1982).

O placar final (g_A, g_B) é então determinado pela amostragem de duas distribuições de Poisson independentes:

$$g_A \sim \text{Poisson}(\lambda_A), \tag{2.4}$$

$$g_B \sim \text{Poisson}(\lambda_B). \tag{2.5}$$

2. **Modelo 2 (Perfil Categórico):** Um modelo puramente categórico que trabalha diretamente com probabilidades de resultado. As probabilidades de vitória mandante, empate e vitória visitante para a partida são calculadas pela média dos perfis correspondentes:

$$P(\text{Vitória A}) = \frac{p_v^m(A) + p_d^v(B)}{2}, \tag{2.6}$$

$$P(\text{Empate}) = \frac{p_e^m(A) + p_e^v(B)}{2}, \tag{2.7}$$

$$P(\text{Vitória B}) = \frac{p_d^m(A) + p_v^v(B)}{2}. \tag{2.8}$$

Um resultado é então sorteado desta distribuição de probabilidade combinada através de uma variável aleatória uniforme $U \sim \text{Uniforme}(0{,}1)$. O resultado é determinado por:

$$\text{Resultado} = \begin{cases} \text{Vitória A} & \text{se } 0 \leq U < P(\text{Vitória A}) \\ \text{Empate} & \text{se } P(\text{Vitória A}) \leq U < P(\text{Vitória A}) + P(\text{Empate}) \\ \text{Vitória B} & \text{se } P(\text{Vitória A}) + P(\text{Empate}) \leq U \leq 1 \end{cases} \tag{2.9}$$

Por exemplo, se $P(\text{Vitória A}) = 0{,}30$, $P(\text{Empate}) = 0{,}50$ e $P(\text{Vitória B}) = 0{,}20$, e sortearmos $U = 0{,}72$, então o resultado seria empate, pois $0{,}30 \leq 0{,}72 < 0{,}80$.

O placar específico não é modelado neste modelo; assume-se um placar genérico (e.g., 1-0 para vitória, 1-1 para empate, 0-1 para derrota).

3. **Modelo 3 (Híbrido):** Um modelo condicional de duas etapas que combina as abordagens anteriores:

   - **Etapa 1 (Classificação):** O resultado categórico (Vitória A / Empate / Vitória B) é determinado exatamente como no Modelo 2 (Perfil Categórico), usando os perfis de desempenho das equipes.
   - **Etapa 2 (Geração de Placar):** O placar específico é gerado a partir de uma distribuição de Poisson condicional. As taxas de gol (λ_A, λ_B), calculadas como no Modelo 1 (Equações 2.2 e 2.3), são usadas para definir uma distribuição conjunta de placares, que é então condicionada ao resultado sorteado na Etapa 1:

$$P(g_A, g_B \mid \text{resultado}) \propto P(g_A \mid \lambda_A) \times P(g_B \mid \lambda_B) \times \mathbb{I}(\text{resultado}) \tag{2.10}$$

   Onde $\mathbb{I}(\text{resultado})$ é uma função indicadora que vale 1 se o placar (g_A, g_B) é consistente com o resultado da Etapa 1, e 0 caso contrário. Por exemplo, se o resultado sorteado foi "Vitória A", então $\mathbb{I} = 1$ se e somente se $g_A > g_B$.
   Um placar (g_A, g_B) é então amostrado da distribuição de probabilidade condicional normalizada.

2.3 Métricas de Avaliação e Função Objetivo

Para avaliar a acurácia das simulações de temporada completa, foram utilizadas duas métricas de erro complementares, ambas baseadas na pontuação final das equipes. A escolha de métricas baseadas em pontuação (e não em placares individuais) reflete o objetivo prático do modelo: prever a classificação final do campeonato.

Definição 2.3.1 (Erro Médio Absoluto (MAE)) *O MAE foi a função objetivo primária utilizada para a otimização do parâmetro h. Ele é definido como:*

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|, \tag{2.11}$$

onde n é o número de equipes no campeonato, y_i é a pontuação final real da equipe i ao término da temporada, e $\hat{y}_i$ é a pontuação final média prevista nas k simulações de Monte Carlo.

O MAE tem interpretação direta: representa o erro médio de previsão de pontos por equipe. Por exemplo, MAE = 2,5 significa que, em média, o modelo erra a pontuação final de cada time em 2,5 pontos.

Definição 2.3.2 (Erro Quadrático Médio (RMSE)) *O RMSE foi calculado como métrica complementar para análise de sensibilidade a erros grandes:*

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{2.12}$$

O RMSE penaliza mais fortemente erros grandes devido à operação de quadrado.

2.4 Processo de Otimização

2.4.1 Simulação de Temporada via Monte Carlo

O processo de previsão do modelo, seja para validação ou avaliação final, baseia-se em uma simulação iterativa da temporada.

Os parâmetros de cada time são inicializados de forma distinta dependendo do componente. Para os perfis categóricos, todos os times iniciam com perfis uniformes $\mathbf{PM}_i = \mathbf{PV}_i = [1/3, 1/3, 1/3]$. Esta inicialização reflete a ausência de informação *a priori* sobre o desempenho das equipes. Já os parâmetros de Poisson são inicializados com os valores observados na primeira partida de cada condição: λ_i^m e μ_i^m são definidos pelos gols marcados e sofridos no primeiro jogo como mandante, enquanto λ_i^v e μ_i^v são definidos pela primeira partida como visitante. Por exemplo, se o time i vence seu primeiro jogo em casa por 2 a 1, então $\lambda_i^m = 2$ e $\mu_i^m = 1$. Se posteriormente empata 1 a 1 fora de casa, então $\lambda_i^v = 1$ e $\mu_i^v = 1$. Embora esta inicialização pontual com base em um único jogo possa introduzir volatilidade inicial, o processo de suavização exponencial (Equação 2.1) converge rapidamente os parâmetros para valores mais estáveis após as primeiras rodadas.

Após a inicialização, todos os parâmetros são atualizados sequencialmente usando o procedimento de suavização exponencial (Equação 2.1) conforme os jogos reais até o ponto de corte são processados. A partir deste ponto, o modelo entra em modo de simulação:

1. Para cada jogo futuro entre os times A e B, o resultado é simulado usando um dos três modelos (Seção 2.2.2).
2. Os pontos são atribuídos (3 para o vencedor, 1 para cada time em caso de empate, 0 para o perdedor).
3. Os parâmetros de ambos os times são atualizados com o resultado simulado (via Equação 2.1) antes de simular a próxima partida.

Ao final da simulação de cada temporada, são contabilizados não apenas os pontos totais, mas também as estatísticas secundárias necessárias para a classificação oficial: o número total de vitórias, o saldo de gols e os gols pró.

A classificação final de cada temporada simulada é então determinada aplicando-se os critérios de desempate (Pontos, depois Vitórias, depois Saldo de Gols, depois Gols Pró). Este passo é crucial, pois garante que a determinação de posições, seja precisa, refletindo as regras do campeonato em vez de um sorteio aleatório em casos de empate de pontos.

Este processo completo é repetido k vezes. O resultado da simulação Monte Carlo é, portanto, um conjunto de k tabelas de classificação finais, que são usadas para calcular as probabilidades de cada time terminar em cada posição específica.

As seções seguintes detalham como este processo de simulação foi aplicado para a otimização de h e para a avaliação final, utilizando diferentes valores de k.

2.4.2 Validação Cruzada Temporal para Otimização de h

A determinação do parâmetro ótimo de memória, $h_{\text{ótimo}}$, foi realizada via validação cruzada temporal (Tashman, 2000; Bergmeir; Benítez, 2012).

O processo seguiu uma abordagem de janela expansível para cada temporada, com $k_{cv} = 5$ folds. Na primeira iteração, o modelo foi treinado com os dados do Fold 1 (rodadas 1 a R_1). Em seguida, os resultados do Fold 2 (rodadas R_1+1 a R_2) foram previstos aplicando-se o processo de simulação Monte Carlo descrito na Seção 2.4.1 com $k = 2.000$ iterações. Na segunda iteração, o modelo foi retreinado com os dados dos Folds 1 e 2 (rodadas 1 a R_2) e testado para prever o Fold 3, novamente com $k = 2.000$ simulações. O processo continuou expandindo progressivamente a janela de treinamento.

Este procedimento foi executado para cada valor de $h \in \{0{,}00, 0{,}05, \ldots, 1{,}00\}$. O valor $h_{\text{ótimo}}$ foi então selecionado como aquele que apresentou o menor MAE médio através de todos os $k_{cv} - 1$ folds de validação:

$$h_{\text{ótimo}} = \arg\min_{h \in \mathcal{H}} \left\{ \frac{1}{k_{cv} - 1} \sum_{j=1}^{k_{cv}-1} \text{MAE}_j(h) \right\}. \tag{2.13}$$

onde $\mathcal{H}$ é o conjunto de valores de h testados e $\text{MAE}_j(h)$ é o erro médio absoluto no fold j, calculado comparando a pontuação real com a **pontuação média prevista** nas 2.000 simulações daquele fold. Este procedimento foi aplicado independentemente para cada uma das 126 combinações de temporada-campeonato.

2.4.3 Avaliação Final do Modelo

Após a determinação do $h_{\text{ótimo}}$ para cada combinação (conforme Seção 2.4.2), foi realizada a avaliação final do desempenho preditivo. Para esta avaliação, o ponto de corte de treinamento foi fixado em aproximadamente 75% da temporada.

As rodadas restantes foram então simuladas aplicando-se o processo de simulação de temporada descrito na Seção 2.4.1, desta vez com $k = 10.000$ iterações.

Com base nestas 10.000 simulações, foram calculadas:

- A pontuação final média prevista para cada equipe;
- As probabilidades de cada equipe terminar em faixas específicas da tabela.

As métricas de erro (MAE e RMSE) reportadas nos resultados (Capítulo 3) referem-se a esta avaliação final, comparando as pontuações reais finais com as pontuações médias previstas nas 10.000 simulações.

2.5 Análise Comparativa entre Ligas

Uma vez estabelecido o protocolo de otimização, a mesma metodologia foi replicada de forma independente e sistemática para os seis campeonatos analisados. Ao aplicar um *framework* metodológico idêntico a todas as ligas (mesmos modelos, mesmas métricas, mesmo processo de validação cruzada), garantiu-se que as diferenças observadas nos resultados refletissem genuinamente as dinâmicas intrínsecas de cada competição, e não artefatos metodológicos.

Para cada liga, os três modelos (Poisson Puro, Perfil Categórico e Híbrido) foram otimizados independentemente. Serão então calculadas e comparadas as seguintes métricas agregadas ao longo das 21 temporadas:

- **Inércia de Desempenho:** O valor médio do parâmetro h ótimo de cada modelo. Isso permite caracterizar a inércia de desempenho típica da liga (se ela depende mais do histórico ou da forma recente) e comparar as dinâmicas capturadas por cada abordagem (categórica vs. quantitativa).
- **Previsibilidade da Liga:** O MAE e RMSE médios, calculados sobre as 21 temporadas. Para cada temporada individual, é gerado um único valor de MAE e RMSE (conforme a metodologia da Seção 2.4.3). Os valores "médios"que serão comparados referem-se à média aritmética destes 21 valores sazonais, permitindo avaliar a previsibilidade geral de cada liga.

Para facilitar a visualização e identificar objetivamente padrões de similaridade entre as ligas, será aplicada uma análise de agrupamento hierárquico (discutida na Seção 1.3). As características do modelo Híbrido (h ótimo médio e MAE médio) serão utilizadas para agrupar os campeonatos, permitindo identificar grupos de ligas com dinâmicas de inércia e previsibilidade similares.

2.6 Análise de Acurácia Categórica

Além da acurácia de pontuação medida pelo MAE e RMSE, este trabalho propõe uma avaliação da utilidade prática dos modelos. Para cada uma das 126 combinações de temporada-campeonato, será avaliada a capacidade do modelo em identificar corretamente as equipes que terminam em posições-chave da tabela.

Esta análise de acurácia categórica focará em três áreas principais:

- **Campeão:** Verificação se a equipe com a maior probabilidade de título nas simulações foi de fato a campeã real.
- **Top 4 (G4):** Avaliação da taxa de acerto completo (identificação de todas as 4 equipes) e parcial (identificação de 2 ou mais equipes).
- **Rebaixados (Rebx):** Avaliação da taxa de acerto completo e parcial para as equipes na zona de rebaixamento, conforme o formato de cada liga.

Esta avaliação será realizada de forma agregada para os três modelos e também detalhada por liga, permitindo identificar se diferentes modelos se destacam em prever faixas específicas da tabela.

2.7 Geração da Previsão Final

Como etapa final e aplicação prática do estudo, a metodologia será utilizada para gerar previsões para os Campeonatos Brasileiros Série A e Série B da temporada de 2025. Diferentemente da análise histórica (onde h é otimizado retrospectivamente com dados completos), a previsão de uma temporada em andamento requer uma abordagem prospectiva.

Para esta aplicação, o parâmetro h será calibrado aplicando-se o procedimento de validação cruzada temporal (Seção 2.4.2) utilizando apenas os dados disponíveis da própria temporada de 2025 (e.g., jogos realizados até a rodada 30 na Série A ou 35 na Série B).

Embora a análise de desempenho (Seção 3.1) indique uma vantagem para o modelo Poisson Puro, o modelo Híbrido será selecionado para esta aplicação prática, dada a sua completude metodológica em gerar placares realistas ao mesmo tempo em que captura as dinâmicas de resultado (V/E/D), preenchendo a lacuna identificada na Seção 1.4.5.

Após a determinação do $h_{\text{ótimo}}$ prospectivo, será aplicado o processo de simulação Monte Carlo (Seção 2.4.1) com $k = 10.000$ iterações para estimar as probabilidades finais de classificação (Campeão, G4, Rebaixamento, etc.).

Adicionalmente às probabilidades de classificação gerais, o *framework* permite análises de cenários combinatórios específicos. Como estudo de caso detalhado para a Série B de 2025,

foi realizada uma análise das combinações de acesso. Identificou-se que, no ponto de corte da previsão (rodada 36), 7 equipes ainda possuíam chances matemáticas de acesso para as 4 vagas disponíveis. Isso gera $C(7,4) = 35$ combinações possíveis para o G4. As $k = 500.000$ simulações foram então processadas para determinar a probabilidade de ocorrência de cada uma destas 35 combinações específicas, detalhando os cenários mais prováveis para a definição do acesso.

2.8 Limitações

2.8.1 Limitações Metodológicas

Este estudo reconhece as seguintes limitações metodológicas inerentes à abordagem adotada:

1. **Independência condicional entre jogos:** O modelo assume que, dados os parâmetros de força das equipes, os resultados dos jogos são independentes. Na realidade, podem existir dependências (e.g., sequências de vitórias que aumentam a confiança do time, ou lesões de jogadores-chave que afetam múltiplos jogos).

2. **Estacionariedade intra-temporada:** O modelo assume que o processo gerador de resultados é estável dentro de uma temporada, sendo ajustado apenas gradualmente pelo parâmetro h. Mudanças abruptas (e.g., troca de técnico, janela de transferências) não são explicitamente modeladas.

3. **Fatores externos não observados:** O modelo não considera explicitamente fatores como lesões de jogadores, transferências, mudanças táticas, motivação diferencial (e.g., times já rebaixados jogando contra candidatos ao título), ou condições climáticas extremas. Assume-se que o impacto desses eventos será capturado ao longo do tempo pela atualização dos parâmetros de desempenho, o que pode não ocorrer de forma imediata.

4. **Simplificação de empates:** No modelo híbrido, empates são tratados como uma categoria única na Etapa 1, sem distinção entre diferentes tipos de empate (0-0 defensivo vs. 3-3 ofensivo). Embora a Etapa 2 gere placares específicos, a probabilidade inicial de empate não diferencia esses cenários.

5. **Distribuição de Poisson:** A suposição de que gols seguem uma distribuição de Poisson, embora amplamente validada na literatura, pode não capturar perfeitamente todas as nuances (e.g., correlação entre gols do mandante e visitante em placares baixos, conforme identificado por Dixon e Coles (Dixon; Coles, 1997)).

2.8.2 Considerações Éticas

Este trabalho tem fins exclusivamente acadêmicos e científicos. Os modelos e previsões desenvolvidos visam contribuir para o entendimento estatístico de padrões em competições esportivas e para o avanço do campo de *sports analytics*.

Os resultados não devem ser interpretados como aconselhamento para apostas esportivas. O futebol é um esporte com alta componente aleatória, e mesmo modelos bem calibrados possuem margens de erro significativas. Qualquer uso dos resultados deste trabalho para fins de apostas é de inteira responsabilidade do usuário e não é endossado pelo autor.

Adicionalmente, reconhece-se que previsões estatísticas, quando divulgadas publicamente, podem influenciar percepções e expectativas sobre equipes e jogadores. Este trabalho adota uma postura de transparência metodológica, explicitando claramente as limitações e incertezas inerentes às previsões.

3 Resultados e Discussão

Este capítulo apresenta e discute os resultados obtidos pela aplicação da metodologia descrita no Capítulo 2. A Seção 3.1 apresenta os resultados centrais da otimização, incluindo a comparação de desempenho (MAE/RMSE) dos três modelos e a análise de agrupamento das ligas. A Seção 3.2 foca na acurácia categórica (previsão de campeão, G4 e rebaixados). Por fim, a Seção 3.3.1 apresenta a aplicação prática do modelo para previsão dos Campeonatos Brasileiros de 2025.

3.1 Otimização e Comparação dos Modelos

Conforme descrito na Seção 2.4.2, o parâmetro de memória h foi otimizado independentemente para cada uma das 126 combinações de temporada-campeonato (6 ligas $\times$ 21 temporadas) e para cada um dos três modelos desenvolvidos, utilizando validação cruzada temporal com 5 *folds*. Para cada combinação, testou-se 21 valores de h no intervalo $[0{,}00, 1{,}00]$ com incrementos de 0,05, selecionando o valor que minimizava o MAE médio.

Após a identificação do h ótimo via validação cruzada, cada modelo foi avaliado simulando 10.000 vezes as rodadas finais da temporada. A Tabela 3 apresenta os resultados agregados por campeonato e modelo.

Tabela 3 – Resultados agregados por campeonato e modelo (2003–2023).

Campeonato	**Nº Temp.**	**h Ótimo (Média)**			**MAE (Média)**			**RMSE (Média)**		
		Híbrido	**Perfil**	**Poisson**	**Híbrido**	**Perfil**	**Poisson**	**Híbrido**	**Perfil**	**Poisson**
Alemão	21	0,771	0,760	0,769	3,04	3,06	2,96	3,83	3,85	3,72
Brasileiro	21	0,833	0,836	0,752	3,33	3,33	3,28	4,03	4,03	3,99
Espanhol	21	0,802	0,817	0,802	3,38	3,36	3,33	4,24	4,22	4,15
Francês	21	0,826	0,817	0,750	3,48	3,50	3,54	4,39	4,39	4,41
Inglês	21	0,783	0,798	0,814	3,84	3,82	3,71	4,66	4,63	4,53
Italiano	21	0,757	0,786	0,783	3,51	3,49	3,43	4,29	4,28	4,19

Fonte: Elaborado pelo autor.

3.1.1 Comparação entre Modelos

A Tabela 3 revela que o modelo de Poisson Puro apresentou o melhor desempenho em todas as seis ligas, com MAE médio global de 3,37 pontos (versus 3,43 e 3,44 dos modelos categóricos). Este resultado valida a adequação da distribuição de Poisson para modelar gols no futebol (Maher, 1982; Dixon; Coles, 1997). As vantagens do Poisson variam entre ligas, sendo mais pronunciadas no Campeonato Inglês (3,4%) e mínimas no Brasileiro (1,5%).

Os modelos Perfil Categórico e Híbrido apresentaram desempenho praticamente idêntico em todas as ligas (diferenças $< 0{,}06$ pontos), indicando que a classificação categórica (V/E/D) é o componente dominante do modelo Híbrido. A geração de placares específicos contribui marginalmente para a previsão de pontuação final, embora forneça informação adicional sobre placares prováveis. Este resultado confirma que, para prever classificações, acertar o resultado é mais importante que acertar o placar exato.

3.1.2 Inércia de Desempenho

A análise dos valores de h ótimo revela que a maioria das combinações apresenta alta inércia ($h > 0{,}75$), indicando que o histórico acumulado é geralmente mais informativo que a forma recente. Observam-se, contudo, padrões interessantes de convergência e divergência entre modelos.

O Campeonato Alemão apresenta valores muito similares entre os três modelos ($h \in [0{,}760, 0{,}771]$), sugerindo evolução sincronizada de aspectos categóricos e quantitativos. Em contraste, Brasil e França mostram grande divergência: modelos categóricos com $h \approx 0{,}83$ versus Poisson com $h \approx 0{,}75$. Esta diferença indica que resultados categóricos são mais estáveis que taxas de gols nestas ligas, possivelmente devido à maior variabilidade de placares. O Campeonato Inglês apresenta padrão invertido, com Poisson tendo maior inércia ($h = 0{,}814$), sugerindo que taxas de gols são mais estáveis que resultados na Premier League.

3.1.3 Previsibilidade entre Ligas

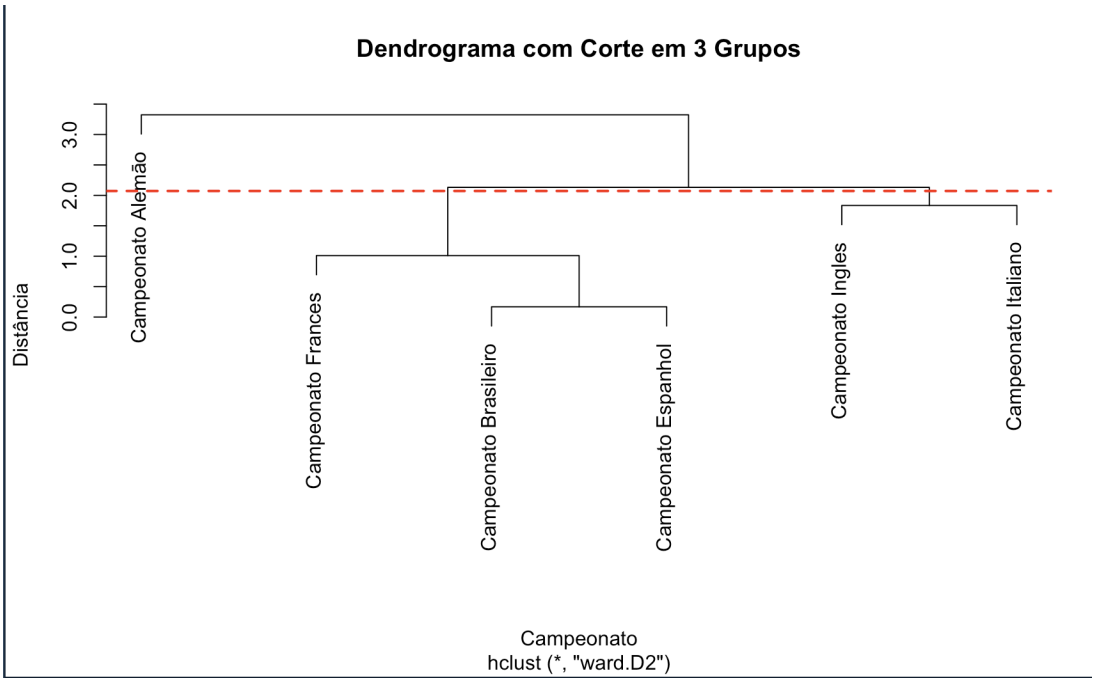
O Campeonato Alemão é o mais previsível em todos os modelos (MAE entre 2,96 e 3,06), beneficiando-se do formato com 18 times. O Campeonato Inglês é o menos previsível (MAE entre 3,71 e 3,84), refletindo o alto equilíbrio competitivo da Premier League. Os demais ocupam posições intermediárias. Um MAE de 3 a 4 pontos representa erro de 4 a 6% da pontuação típica de meio de tabela, equivalendo a errar 1 a 2 jogos por equipe na temporada completa.

3.1.4 Análise de Agrupamento Hierárquico das Ligas

Para identificar objetivamente padrões de similaridade entre as ligas, foi aplicada análise de agrupamento hierárquico utilizando o método de Ward sobre as características do modelo híbrido (h ótimo médio e MAE médio). A Figura 1 apresenta o dendrograma resultante com corte em 3 grupos.

O dendrograma identifica três grupos distintos de ligas com características preditivas similares. O Grupo 1 é formado exclusivamente pelo Campeonato Alemão, que se diferencia significativamente das demais ligas por combinar menor inércia com maior previsibilidade. O Grupo 2 agrupa Campeonatos Brasileiro e Espanhol, que compartilham alta inércia e previsibili-

Figura 1 – Dendrograma hierárquico das ligas com base em h ótimo e MAE do modelo híbrido (corte em 3 grupos).



Fonte: Elaborado pelo autor.

dade intermediária. O Grupo 3 reúne Campeonatos Inglês e Italiano, caracterizados por inércia intermediária-baixa e menor previsibilidade.

Interessantemente, o Campeonato Francês aparece isolado, formando um ramo intermediário entre os grupos, refletindo suas características únicas: alta inércia (similar a Brasil e Espanha) mas previsibilidade moderada-baixa (similar a Inglaterra e Itália). Esta análise hierárquica confirma que inércia e previsibilidade são dimensões parcialmente independentes, e que diferentes ligas podem compartilhar uma característica (e.g., alta inércia) mas diferir substancialmente em outra (e.g., previsibilidade).

3.2 Acurácia Categórica

Além das métricas de erro contínuo (MAE/RMSE), avaliou-se a capacidade prática dos modelos de identificar corretamente as equipes que terminam em posições críticas da tabela. Para cada uma das 126 temporadas-campeonato, comparou-se as equipes com maior probabilidade prevista (campeão, top 4, rebaixados) com as equipes que efetivamente terminaram nessas posições.

A Tabela 4 apresenta o desempenho agregado dos três modelos, enquanto as Tabelas 5, 6 e 7 detalham os resultados por liga para cada modelo.

Tabela 4 – Acurácia categórica global dos modelos (126 temporadas-campeonato).

Modelo	Campeão (%)	G4 4/4 (%)	G4 ≥2 (%)	Rebx 4/4 (%)	Rebx ≥2 (%)	MAE
Poisson	**82,5**	**55,6**	44,4	**27,0**	73,0	**3,37**
Perfil	79,4	57,1	42,9	23,8	76,2	3,43
Híbrido	79,4	56,3	43,7	23,8	76,2	3,43

Fonte: Elaborado pelo autor.

A Tabela 4 demonstra que o modelo de Poisson Puro apresenta desempenho superior não apenas em termos de MAE, mas também em acurácia categórica. O modelo identificou corretamente o campeão em 82,5% das temporadas (104 de 126), acertou completamente o top 4 em 55,6% dos casos (70 temporadas), e identificou os 4 rebaixados corretamente em 27,0% das temporadas (34 casos). Os modelos Perfil e Híbrido apresentaram desempenho muito similar entre si, ligeiramente inferior ao Poisson em todas as métricas categóricas.

3.2.1 Acurácia por Liga

As Tabelas 5, 6 e 7 detalham o desempenho de cada modelo por campeonato.

Tabela 5 – Acurácia categórica do modelo Híbrido por campeonato (2003–2023).

Campeonato	Temp.	Campeão (%)	G4 Compl. (4/4, %)	G4 Parc. (≥2, %)	Rebx Compl. (4/4, %)	Rebx Parc. (≥2, %)	MAE
Alemão	21	90,5	61,9	38,1	33,3	66,7	3,04
Brasileiro	21	66,7	42,9	57,1	14,3	85,7	3,33
Espanhol	21	76,2	76,2	23,8	9,5	90,5	3,38
Francês	21	81,0	38,1	61,9	23,8	76,2	3,48
Inglês	21	71,4	61,9	38,1	23,8	76,2	3,84
Italiano	21	90,5	57,1	42,9	38,1	61,9	3,51

Fonte: Elaborado pelo autor.

Tabela 6 – Acurácia categórica do modelo Poisson Puro por campeonato (2003–2023).

Campeonato	Temp.	Campeão (%)	G4 Compl. (4/4, %)	G4 Parc. (≥2, %)	Rebx Compl. (4/4, %)	Rebx Parc. (≥2, %)	MAE
Alemão	21	95,2	66,7	33,3	38,1	61,9	2,96
Brasileiro	21	66,7	38,1	61,9	19,0	81,0	3,28
Espanhol	21	85,7	76,2	23,8	9,5	90,5	3,33
Francês	21	81,0	33,3	66,7	28,6	71,4	3,54
Inglês	21	76,2	66,7	33,3	19,0	81,0	3,71
Italiano	21	90,5	52,4	47,6	47,6	52,4	3,43

Fonte: Elaborado pelo autor.

Tabela 7 – Acurácia categórica do modelo Perfil Categórico por campeonato (2003–2023).

Campeonato	Temp.	Campeão (%)	G4 Compl. (4/4, %)	G4 Parc. ($\geq$2, %)	Rebx Compl. (4/4, %)	Rebx Parc. ($\geq$2, %)	MAE
Alemão	21	90,5	66,7	33,3	33,3	66,7	3,06
Brasileiro	21	66,7	42,9	57,1	14,3	85,7	3,33
Espanhol	21	76,2	76,2	23,8	9,5	90,5	3,36
Francês	21	81,0	38,1	61,9	23,8	76,2	3,50
Inglês	21	71,4	61,9	38,1	19,0	81,0	3,82
Italiano	21	90,5	57,1	42,9	42,9	57,1	3,49

Fonte: Elaborado pelo autor.

A análise por liga revela variações significativas na capacidade de identificar posições críticas. Para campeões, o modelo de Poisson apresenta as maiores taxas de acerto, destacando-se no Campeonato Alemão (95,2%) e Italiano (90,5%). O Campeonato Brasileiro apresenta a menor taxa de acerto de campeão em todos os modelos (66,7%), refletindo maior imprevisibilidade na disputa pelo título, possivelmente devido ao formato de pontos corridos favorecendo maior equilíbrio.

Para o top 4, o Campeonato Espanhol apresenta as maiores taxas de acerto completo (76,2% em todos os modelos), indicando hierarquia mais definida. O Campeonato Francês apresenta as menores taxas de acerto completo (33 a 38%), mas altas taxas de acerto parcial (62 a 67%), sugerindo que a disputa pelas vagas europeias é mais equilibrada.

No rebaixamento, observa-se maior dificuldade geral de previsão, com taxas de acerto completo variando de 9,5% (Espanha) a 47,6% (Itália no Poisson). O Campeonato Espanhol apresenta consistentemente as menores taxas de acerto de rebaixados mas as maiores de acerto parcial (90,5%), indicando que, embora o modelo identifique corretamente a zona de perigo, a definição exata dos 4 rebaixados é mais difícil. O Campeonato Italiano apresenta as maiores taxas de acerto completo de rebaixados (38 a 48%), sugerindo maior previsibilidade na zona inferior da tabela.

Estes resultados demonstram que os modelos possuem não apenas precisão numérica de pontuação, mas também utilidade prática para identificar zonas críticas da tabela, com desempenho variando significativamente entre ligas e refletindo diferentes níveis de equilíbrio competitivo.

3.3 Aplicação

3.3.1 Previsão dos Campeonatos Brasileiros 2025

Como aplicação prática, o modelo híbrido foi utilizado para prever os Campeonatos Brasileiros de 2025, com h calibrado via validação cruzada: $h = 0{,}75$ (Série A, até rodada 30) e $h = 0{,}80$ (Série B, até rodada 35). A Figura 2 apresenta as probabilidades.

Na Série A, Flamengo (46,09%) e Palmeiras (42,91%) disputam o título de forma equilibrada, ambos com G4 praticamente garantido ($>$ 99,5%). Cruzeiro (91,64%) e Mirassol (81,42%) são favoritos às outras vagas de Libertadores. Sport está praticamente rebaixado (99,34%), seguido por Juventude (82,35%) e Fortaleza (62,37%). Vitória mantém 77,29% de chance de permanência.

Na Série B, Coritiba é amplo favorito (72,23% título; 99,16% acesso). Chapecoense (87,28%), Remo (75,86%) e Goiás (56,48%) disputam as vagas restantes. Paysandu está praticamente rebaixado à Série C (99,87%), seguido por Amazonas (80,90%) e Volta Redonda (65,38%).

Figura 2 – Previsões Campeonatos Brasileiros 2025.

Campeonato Brasileiro Série A - 2025

Probabilidades de Campeão, G4, Pré-Libertadores, Sul Americana e Rebaixamento

■ G4 ■ Pré-Libertadores ■ Sul Americana ■ Misto ■ Rebaixamento

Pos	Time	Pts	Campeão %	G4 %	Reb. %
1	Flamengo	75	46,09	99,76	0,00
2	Palmeiras	75	42,91	99,55	0,00
3	Cruzeiro	69	6,08	91,64	0,00
4	Mirassol	67	4,87	81,42	0,00
5	Bahia	61	0,00	11,76	0,00
6	Fluminense	59	0,04	7,93	0,00
7	Botafogo	58	0,01	5,96	0,00
8	Vasco da Gama	55	0,00	1,59	0,00
9	Corinthians	52	0,00	0,11	0,16
10	São Paulo	52	0,00	0,20	0,12
11	Grêmio	50	0,00	0,05	0,21
12	Atlético MG	49	0,00	0,03	1,30
13	Internacional	47	0,00	0,00	1,94
14	Ceará	45	0,00	0,00	6,34
15	Bragantino	45	0,00	0,00	6,86
16	Santos	43	0,00	0,00	16,30
17	Vitória	42	0,00	0,00	22,71
18	Fortaleza	38	0,00	0,00	62,37
19	Juventude	35	0,00	0,00	82,35
20	Sport	28	0,00	0,00	99,34

Autor: Danilo Carvalho

(a) Série A

Campeonato Brasileiro Série B - 2025

Probabilidades de Campeão, Acesso e Rebaixamento

■ Acesso ■ Misto ■ Rebaixamento

Pos	Time	Pts	Campeão %	Acesso %	Reb. %
1	Coritiba	67	72,23	99,16	0,00
2	Chapecoense	64	13,08	87,28	0,00
3	Remo	63	8,57	75,86	0,00
4	Goiás	61	3,81	56,48	0,00
5	Novorizontino	60	1,65	41,02	0,00
6	Criciúma	59	0,42	19,51	0,00
7	Ath Paranaense	59	0,24	16,85	0,00
8	CRB	57	0,00	2,72	0,00
9	Cuiabá	55	0,00	0,66	0,00
10	Atl Goianiense	54	0,00	0,27	0,00
11	Avaí	54	0,00	0,19	0,00
12	Vila Nova	49	0,00	0,00	0,00
13	Operário	47	0,00	0,00	0,00
14	América (MG)	46	0,00	0,00	0,06
15	Ferroviária	45	0,00	0,00	1,12
16	Athletic	41	0,00	0,00	19,41
17	Botafogo (SP)	40	0,00	0,00	33,26
18	Volta Redonda	39	0,00	0,00	65,38
19	Amazonas	37	0,00	0,00	80,90
20	Paysandu	31	0,00	0,00	99,87

Autor: Danilo Carvalho

(b) Série B

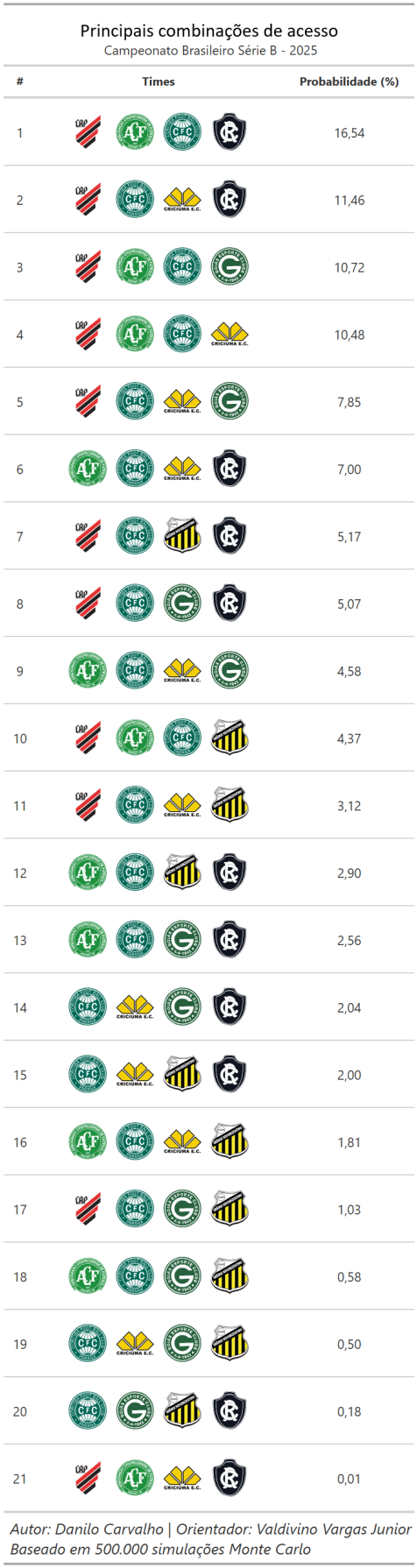
Fonte: Elaborado pelo autor.

3.3.2 Análise das combinações de acesso

Conforme descrito na metodologia (Seção 2.7), foi realizado um estudo de caso detalhado para o Campeonato Brasileiro Série B de 2025. Após a rodada 36 (com duas rodadas restantes), identificou-se que 7 equipes ainda disputavam 4 vagas de acesso, resultando em 35 combinações matemáticas possíveis ($C(7,4) = 35$).

As 500.000 simulações Monte Carlo foram processadas para identificar a probabilidade de ocorrência de cada uma dessas 35 combinações. A Figura 3 apresenta as 21 combinações mais prováveis que foram observadas nas simulações, detalhando os cenários de acesso.

A tabela demonstra a concentração de probabilidade nos cenários mais prováveis. As cinco principais combinações, por exemplo, somam mais de 50% de chance de ocorrência, fornecendo uma visão clara de quais grupos de times tinham maior probabilidade de configurar o G4 final.



Principais combinações de acesso
Campeonato Brasileiro Série B - 2025

#	Times	Probabilidade (%)
1		16,54
2		11,46
3		10,72
4		10,48
5		7,85
6		7,00
7		5,17
8		5,07
9		4,58
10		4,37
11		3,12
12		2,90
13		2,56
14		2,04
15		2,00
16		1,81
17		1,03
18		0,58
19		0,50
20		0,18
21		0,01

Autor: Danilo Carvalho | Orientador: Valdivino Vargas Junior
Baseado em 500.000 simulações Monte Carlo

Figura 3 – Principais combinações de acesso (G4) para a Série B 2025, após a rodada 36. As probabilidades são baseadas em 500.000 simulações Monte Carlo, mostrando os 21 cenários mais prováveis.

Fonte: Elaborado pelo autor.

Conclusão

Este trabalho comparou sistematicamente três abordagens de modelagem preditiva para campeonatos de futebol (Poisson Puro, Perfil Categórico e Híbrido), aplicando-as a seis das principais ligas mundiais ao longo de 21 temporadas (2003 a 2023), totalizando 126 temporadas-campeonato e 46.503 partidas analisadas.

Dentro do escopo desta pesquisa e considerando as métricas avaliadas, os resultados indicaram um melhor desempenho do modelo de Poisson Puro, que obteve as melhores marcas no conjunto de dados das seis ligas (MAE global: 3,37 versus 3,43 dos modelos categóricos) e maior acurácia categórica (82,5% de acerto de campeões versus 79,4%). Este resultado valida empiricamente, para a amostra considerada, a adequação da distribuição de Poisson para modelar gols no futebol. Os modelos Perfil Categórico e Híbrido apresentaram desempenho praticamente idêntico, indicando que a classificação categórica (V/E/D) é o componente dominante e que acertar o resultado é mais importante que acertar o placar exato para prever pontuações finais.

Todas as combinações de liga e modelo apresentaram alta inércia ($h > 0{,}75$), indicando que o histórico acumulado é mais informativo que a forma recente. Observaram-se, contudo, diferenças sistemáticas entre ligas: Brasil e França com maior inércia ($h \approx 0{,}83$), Itália e Alemanha mais reativas ($h \approx 0{,}76$). Em termos de previsibilidade, o Campeonato Alemão foi o mais previsível (MAE = 2,96) e o Inglês o menos previsível (MAE = 3,71). A análise hierárquica identificou três grupos distintos de ligas, revelando que inércia e previsibilidade são dimensões parcialmente independentes.

As principais limitações incluem a não incorporação de fatores externos (lesões, transferências, motivação) e a suposição de independência entre jogos. Trabalhos futuros podem explorar modelos bayesianos hierárquicos, técnicas de aprendizado de máquina, incorporação de covariáveis adicionais e implementação da correção de Dixon-Coles para placares baixos.

A aplicação prática do modelo híbrido para previsão dos Campeonatos Brasileiros 2025 demonstrou sua viabilidade operacional, indicando disputa equilibrada pelo título da Série A entre Flamengo (46%) e Palmeiras (43%), e favorecendo Coritiba na Série B (72% de probabilidade de título). Este trabalho demonstra que, apesar da alta componente aleatória do futebol, abordagens estatísticas rigorosas podem capturar padrões subjacentes e fornecer *insights* valiosos sobre dinâmicas competitivas.

Referências

BERGMEIR, C.; BENÍTEZ, J. M. On the use of cross-validation for time series predictor evaluation. **Information Sciences**, Elsevier, v. 191, p. 192–213, 2012. Citado na página 32.

DIXON, M. J.; COLES, S. G. Modelling association football scores and inefficiencies in the football betting market. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 46, n. 2, p. 265–280, 1997. Citado 5 vezes nas páginas 14, 21, 24, 35 e 37.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2nd. ed. New York: Springer, 2009. Citado na página 19.

JR, J. H. W. Hierarchical grouping to optimize an objective function. **Journal of the American Statistical Association**, Taylor & Francis, v. 58, n. 301, p. 236–244, 1963. Citado na página 20.

KOLMOGOROV, A. N. **Foundations of the Theory of Probability**. New York: Chelsea Publishing Company, 1956. Tradução da obra original alemã de 1933, "Grundbegriffe der Wahrscheinlichkeitsrechnung", que estabeleceu a base axiomática da probabilidade. Citado na página 16.

KUHNERT, F. V.; POSSATO, P. N. Modelo para predição dos resultados de partidas de futebol. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação). 2023. Citado 2 vezes nas páginas 14 e 24.

LIMA, B. N. B. d. *et al.* Probabilidades no futebol. **Revista Matemática Universitária**, n. 48/49, p. artigo 02, 2010. Citado 3 vezes nas páginas 23, 24 e 27.

MAHER, M. J. Modelling association football scores. **Statistica Neerlandica**, Wiley Online Library, v. 36, n. 3, p. 109–118, 1982. Citado 5 vezes nas páginas 14, 21, 27, 29 e 37.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2018. Disponível em: <http://www.R-project.org/>. Citado na página 25.

RAMOS, L. F. P.; LEMOS, H. C. F.; BATISTA, B. D. d. O. **Modelagem Matemática para Previsão de Jogos de Futebol**. Dissertação (Mestrado) — Mestrado Profissional em Matemática PROFMAT, Universidade Federal de São Jõa o Del-Rei - UFSJ, 2019. Dissertação de Mestrado. Citado 3 vezes nas páginas 14, 23 e 24.

ROSS, S. **A First Course in Probability**. 9th. ed. Boston: Pearson, 2014. Citado na página 16.

TASHMAN, L. J. Out-of-sample tests of forecasting accuracy: an analysis and review. **International Journal of Forecasting**, Elsevier, v. 16, n. 4, p. 437–450, 2000. Citado na página 32.

APÊNDICE A – Algoritmo

```
//====================================================================
// ALGORITMO 1: Framework Principal de Análise
// Autor: Danilo Silva Carvalho de Oliveira (TCC 2025)
// Objetivo: Otimizar e avaliar 3 modelos em 6 ligas (21 temporadas)
//====================================================================

// --- 1. Inicialização ---
Carregar_Dados_Completos(arquivo = "partidas_2003_2023.csv")
Lista_Ligas = ["Alemão", "Brasileiro", "Espanhol", "Francês", "Inglês", "Italiano"]
Lista_Temporadas = [2003, ..., 2023]
Lista_Modelos = ["Poisson Puro", "Perfil Categórico", "Híbrido"]
Lista_h = [0.00, 0.05, ..., 1.00] // 21 valores

// Lista para armazenar os resultados da Tabela 3
Resultados_Agregados = []

// --- 2. Loop Principal ---
PARA CADA liga EM Lista_Ligas
    PARA CADA temporada EM Lista_Temporadas

        // Filtra os dados da temporada-campeonato específica
        Dados_Temporada = Filtrar_Dados(Dados_Completos, liga, temporada)
        // Garante a ordem cronológica correta (pela data)
        Ordenar_Por_Data(Dados_Temporada)

        PARA CADA modelo EM Lista_Modelos

            // --- 3. Otimização (Validação Cruzada Temporal) ---
            // Processo descrito na Seção 2.4.2
            h_otimo = Encontrar_h_Otimo(Dados_Temporada, modelo, Lista_h)

            // --- 4. Avaliação Final do Modelo ---
            // Processo descrito na Seção 2.4.3
            Ponto_Corte_Final = 75% da temporada // (Aprox. rodada 28 de 38)
            Dados_Treino_Final = Dados_Temporada[1 ... Ponto_Corte_Final]
            Dados_Teste_Final = Dados_Temporada[Ponto_Corte_Final+1 ... Fim]

            // Roda as 10.000 simulações
            Resultados_Simulacao_Final = Simular_Temporada(
                modelo = modelo,
                h = h_otimo,
                dados_historicos = Dados_Treino_Final,
```

```
                jogos_para_simular = Dados_Teste_Final.jogos,
                k_simulacoes = 10000
            )

            // --- 5. Cálculo de Métricas Finais ---
            Pont_Media_Prevista = Calc_Media_Pontuacoes(Resultados_Simulacao_Final)
            Pont_Real_Final = Calc_Pontos_Reais(Dados_Temporada)

            MAE_final = Calcular_MAE(Pont_Media_Prevista, Pont_Real_Final)
            RMSE_final = Calcular_RMSE(Pont_Media_Prevista, Pont_Real_Final)

            // Cálculo para as Tabelas 4, 5, 6, 7
            Acu_Cat = Calc_Acuracia_Categ(Resul_Simulacao_Final, Pont_Real_Final)

            // --- 6. Armazenamento ---
            Salvar_Resultado(Resultados_Agregados, liga, temporada, modelo,
                             h_otimo, MAE_final, RMSE_final, Acuracia_Cat)
        FIM_PARA
    FIM_PARA
FIM_PARA

// --- 7. Análise Final ---
Gerar_Tabela_Resultados_Agregados(Resultados_Agregados) // Tabela 3
Gerar_Tabelas_Acuracia(Resultados_Agregados)           // Tabelas 4, 5, 6, 7
Gerar_Dendrograma(Resultados_Agregados)                // Figura 1
```

```
//====================================================================
// FUNÇÃO 1: Encontrar_h_Otimo (Seção 2.4.2)
//====================================================================
FUNÇÃO Encontrar_h_Otimo(Dados_Temporada, modelo, Lista_h)
    Definir_Folds(Dados_Temporada, k_cv = 5) // (Fold_1, ..., Fold_5)
    Resultados_CV = []

    PARA CADA h EM Lista_h
        Erros_Folds = []
        // O loop de teste vai de j=1 até k_cv-1 = 4
        PARA j EM [1, ..., 4]
            Dados_Treino_CV = Combinar_Folds(1 ... j)
            Dados_Teste_CV = Fold_[j+1]

            // Roda as 2.000 simulações para este fold
            Resultados_Simulacao_CV = Simular_Temporada(
                modelo = modelo,
                h = h,
                dados_historicos = Dados_Treino_CV,
                jogos_para_simular = Dados_Teste_CV.jogos,
                k_simulacoes = 2000
            )

            // Compara a média das simulações com o real
            Pontuacao_Media_CV = Calcular_Media_Pontuacoes(Resultados_Simulacao_CV)
            // Pontuação real no final do fold de teste
            Pontuacao_Real_CV = Calcular_Pontos_Reais(Combinar_Folds(1 ... j+1))

            MAE_fold = Calcular_MAE(Pontuacao_Media_CV, Pontuacao_Real_CV)
            Adicionar(Erros_Folds, MAE_fold)
        FIM_PARA

        // Calcula o MAE médio para o valor 'h' em todos os 4 folds
        MAE_medio_h = Calcular_Media(Erros_Folds)
        Salvar_Resultado(Resultados_CV, h, MAE_medio_h)
    FIM_PARA

    // Encontra o 'h' que minimizou o MAE médio
    h_otimo = Encontrar_h_minimo(Resultados_CV) // Eq. 2.13
    RETORNAR h_otimo
FIM_FUNÇÃO
```

```
//====================================================================
// FUNÇÃO 2: Simular_Temporada (O "Motor" - Seção 2.4.1)
//====================================================================
FUNÇÃO Simular_Temporada(modelo, h, dados_historicos,
jogos_para_simular, k_simulacoes)
    Resultados_k_Simulacoes = [] // Armazena as pontuações finais de cada simulação

    // --- Etapa 1: Treinamento Base ---
    // Inicializa os parâmetros (categóricos = [1/3, 1/3, 1/3], Poisson = 1° jogo)
    Parametros_Base = Inicializar_Parametros()
    // "Treina" os parâmetros com todos os dados históricos reais
    PARA CADA jogo_real EM dados_historicos
        Parametros_Base = Atualizar_Parametros(Parametros_Base, jogo_real, h)
    FIM_PARA

    Pontuacao_Base = Calcular_Pontos_Reais(dados_historicos)

    // --- Etapa 2: Simulação Monte Carlo (k vezes) ---
    PARA i EM [1, ..., k_simulacoes]
        // Reseta os parâmetros e a pontuação para o estado "treinado"
        Parametros_Sim = Copiar(Parametros_Base)
        Pontuacao_Sim = Copiar(Pontuacao_Base)

        // Simula o restante da temporada
        PARA CADA jogo_sim EM jogos_para_simular
            // Gera o resultado (V/E/D) e o placar
            Resultado_Simulado = Gerar_Resultado_Jogo(Parametros_Sim, modelo,
                                                      jogo_sim.timeA, jogo_sim.timeB)

            // Adiciona os pontos do jogo simulado
            Pontuacao_Sim = Adicionar_Pontos(Pontuacao_Sim, Resultado_Simulado)

            // Atualiza os parâmetros com o resultado SIMULADO
            Parametros_Sim = Atualizar_Parametros(Parametros_Sim, Resultado_Simulado,
            h)
        FIM_PARA

        // Salva a pontuação final desta simulação
        Adicionar(Resultados_k_Simulacoes, Pontuacao_Sim)
    FIM_PARA

    RETORNAR Resultados_k_Simulacoes
FIM_FUNÇÃO
```

```
//====================================================================
// FUNÇÃO 3: Gerar_Resultado_Jogo (Seção 2.2.2)
//====================================================================
FUNÇÃO Gerar_Resultado_Jogo(Parametros, modelo, timeA, timeB)
    SE modelo == "Poisson Puro"
        Taxas_Pois = Calcular_Taxas_Poisson(Parametros,
                                            timeA,
                                            timeB)
        Gols_A = Sortear_Poisson(Taxas_Pois.lambda_A)
        Gols_B = Sortear_Poisson(Taxas_Pois.lambda_B)
        RETORNAR (Resultado=(Gols_A, Gols_B), Placar=(Gols_A, Gols_B))

    SE modelo == "Perfil Categórico"
        Probs_Cat = Calcular_Probs_Categoricas(Parametros,
                                               timeA,
                                               timeB)
        Resultado = Sortear_Multinomial(Probs_Cat)
        Placar_Gen = Resultado_Para_Placar_Generico(Resultado)
        RETORNAR (Resultado=Resultado, Placar=Placar_Gen)

    SE modelo == "Híbrido"
        // Etapa 1: Classificação
        Probs_Cat = Calcular_Probs_Categoricas(Parametros, timeA, timeB)
        Resultado_Cat = Sortear_Multinomial(Probs_Cat)

        // Etapa 2: Geração de Placar
        Taxas_Pois = Calcular_Taxas_Poisson(Parametros, timeA, timeB)
        Placar_Cond = Sortear_Poisson_Condicional(Taxas_Pois, Resultado_Cat)
        RETORNAR (Resultado=Resultado_Cat, Placar=Placar_Cond)
    FIM_SE
FIM_FUNÇÃO

//====================================================================
// FUNÇÃO 4: Atualizar_Parametros (Eq. 2.1)
//====================================================================
FUNÇÃO Atualizar_Parametros(Parametros_Antigos, Jogo_Observado, h)
    // Para parâmetros de Poisson (ex: ataque mandante time A)
    Param_Novos.lambda_A = h * Param_Antigos.lambda_A + (1-h) * Jogo_Obs.Gols_A

    // Para perfis categóricos (ex: perfil mandante time A)
    SE Jogo_Observado.Resultado == "Vitoria_A" ENTÃO Obs_Vetor = [1, 0, 0]
    SE Jogo_Observado.Resultado == "Empate" ENTÃO Obs_Vetor = [0, 1, 0]
    SE Jogo_Observado.Resultado == "Derrota_A" ENTÃO Obs_Vetor = [0, 0, 1]

    Parametros_Novos.PM_A = h * Parametros_Antigos.PM_A + (1-h) * Obs_Vetor

    // ... (repetir para todos os 6 parâmetros de cada time) ...
```

```
    RETORNAR Parametros_Novos
FIM_FUNÇÃO
```

APÊNDICE B – Gráficos de otimização parametro h

As imagens da otimização do parâmetro h estão disponíveis no seguinte repositório GitHub: <https://github.com/DanOakStats/TCC/tree/main/Imagens>