Previsão de Resultados na Premier League: Comparação Sistemática de Modelos de Aprendizado de Máquina

Murilo Pedral Mota[1], **Wilia Francisco da Silva**[1], **Victor Flávio de Andrade Araujo**[2]

[1]Grupo de Pesquisa Interdisciplinar em Tecnologia, Computação e Sociedade (GPITCS) – Universidade Tiradentes (UNIT)
CEP 49032-490 – Aracaju – SE – Brasil

[2]Pontifícia Universidade Católica do Rio Grande do Sul (PUC-RS)

{murilo.pedral, wilia.francisco}@souunit.com.br,
victor.araujo@edu.pucrs.br

Resumo. *Este trabalho compara nove modelos de aprendizado de máquina para a previsão de resultados da Premier League inglesa (16 temporadas, 2010/11–2025/26;* $N = 6.080$ *partidas), sob validação temporal walk-forward, teste independente e bootstrap pareado em dois esquemas de reamostragem, com Log Loss como métrica principal. O protocolo detecta vazamento temporal nas features e na calibração. O Poisson superou com significância estatística os oito modelos discriminativos calibrados e duas das três especificações por força de equipe; dele permanecem indistinguíveis apenas a Regressão Logística ajustada sem calibração e a variante Dixon-Coles com decaimento temporal. No teste final obteve Log Loss de 1,0099 e acurácia de 50,92%, sendo escolhido por parcimônia e interpretabilidade.*

Palavras-chave: *previsão de resultados de futebol; Premier League; aprendizado de máquina; modelo de Poisson; validação temporal; calibração de probabilidades; Log Loss.*

Abstract. *This study compares nine machine learning models for predicting English Premier League outcomes (16 seasons, 2010/11–2025/26;* $N = 6{,}080$ *matches), under walk-forward validation, an independent test set, and paired bootstrap in two resampling schemes, with Log Loss as the primary metric. The pipeline detects temporal leakage in the features and in the calibration step. The Poisson model significantly outperformed all eight calibrated discriminative models and two of the three team-strength specifications; only a tuned, uncalibrated Logistic Regression and the Dixon-Coles variant with time decay remained indistinguishable from it. On the test set it achieved a Log Loss of 1.0099 and 50.92% accuracy, and was selected for its parsimony and interpretability.*

Keywords: *football match prediction; English Premier League; machine learning; Poisson model; temporal validation; probability calibration; Log Loss.*

1. Introdução

A previsão de resultados de futebol é um problema de classificação multiclasse, vitória do mandante (H), empate (D) ou vitória do visitante (A) de alta incerteza, em que mesmo modelos sofisticados raramente superam 55% de acurácia em ligas competitivas [Hubáček et al. 2022, Bunker and Susnjak 2022]. A Premier League é um dos campeonatos mais estudados [Fischer and Heuer 2024]; modelos de Poisson têm demonstrado com-

petitividade surpreendente frente a algoritmos mais complexos [Fischer and Heuer 2024, Penn and Donnelly 2022], motivando investigar o real ganho de cada nível de complexidade. Um problema recorrente é a ausência de protocolos rigorosos de validação temporal: *k-fold* aleatório em dados temporais viola a causalidade, produzindo estimativas otimistas [Hyndman and Athanasopoulos 2021, Kapoor and Narayanan 2023], e estimativas pontuais sem testes de significância levam a conclusões não reproduzíveis [Demšar 2006].

Este trabalho apresenta uma comparação sistemática de nove modelos em quatro níveis de complexidade crescente, sob protocolo de validação temporal rigoroso. As contribuições centrais são: (i) detecção e remoção de um vazamento temporal severo nas *features* originais, que produzia acurácias espúrias de até 99,7%; (ii) avaliação das hipóteses de melhoria via *bootstrap* pareado com 5.000 réplicas em dois esquemas de reamostragem; e (iii) a demonstração de que o Poisson, a solução mais simples do conjunto, supera com significância estatística os oito discriminativos após calibração e duas das três especificações por força de equipe, permanecendo indistinguível de uma Regressão Logística ajustada e sem calibração e da variante Dixon-Coles com decaimento temporal resultado que reforça o princípio da parcimônia em contextos de alta incerteza.

2. Referencial Teórico

2.1. Modelos Probabilísticos, Calibração e Métricas

O Poisson duplo independente assume que gols do mandante e do visitante seguem distribuições de Poisson distintas, derivando P(H), P(D), P(A) pela integração da conjunta; Nguyen [Nguyen 2021] confirma essa distribuição na Premier League, e Dixon e Coles [Dixon and Coles 1997] estendem o modelo com ρ para corrigir a subestimação de empates e vitórias por margem mínima. Penn e Donnelly [Penn and Donnelly 2022] aplicam o Poisson duplo à Euro 2020 superando todos os participantes do concurso da *Royal Statistical Society*, achado convergente com Fischer e Heuer [Fischer and Heuer 2024] e Hubáček [Hubáček et al. 2022]. Bunker [Bunker and Susnjak 2022] aponta a inconsistência de protocolos de validação como obstáculo à comparação entre estudos, e Galekwa [Galekwa et al. 2024] mostra que *odds* superam a maioria dos modelos de *ML*.

XGBoost [Chen and Guestrin 2016] introduz regularização L_1/L_2; LightGBM [Ke et al. 2017] cresce árvores por folha; CatBoost [Prokhorenkova et al. 2018] usa *ordered boosting* contra vazamento no treino, com diferenças pequenas entre os três em $N < 10.000$ [Bentéjac et al. 2021]. SVM e *gradient boosting* tendem a probabilidades mal calibradas; Silva [Silva Filho et al. 2023] revisa a calibração, sendo a regressão isotônica via `CalibratedClassifierCV` a referência para N moderado, relevante pois o *Log Loss* penaliza classificação e calibração simultaneamente, o que motiva Wheatcroft [Wheatcroft 2021] a preferir métricas de calibração completa ao *Ranked Probability Score*.

2.2. Validação Temporal e Inferência Estatística

Kapoor [Kapoor and Narayanan 2023] documenta que vazamento entre treino e teste é o modo de falha mais comum em *ML* aplicado. Hyndman [Hyndman and Athanasopoulos 2021] sustenta a validação *walk-forward* para dados com dependência temporal, e Demšar [Demšar 2006] recomenda IC por rea-

mostragem como critério de superioridade, o critério de IC95% do *delta* não cruzando zero, adotado aqui, aplica essa recomendação [Efron and Tibshirani 1994]. Julian *et al.* [Julian et al. 2021] associam congestão de jogos a queda de desempenho físico, o que motivou as seis *features* de calendário da Linha V3 (Seção 4), sem ganho observado.

3. Metodologia

3.1. Dados e Construção de Features

Os dados foram obtidos do Football-Data.co.uk, cobrindo 16 temporadas da Premier League (2010/11–2025/26, $N = 6.080$ partidas).[1] O conjunto final conta com 44 *features*, em quatro blocos (Tabela 1), após a auditoria de vazamento temporal descrita adiante nesta seção. Todas as variáveis são calculadas sobre o histórico estritamente anterior à partida-alvo m, com `shift(1)` aplicado antes de qualquer agregação móvel.

Tabela 1. As 44 *features* do conjunto final, por bloco.

Bloco	Variáveis	Qtd.
1 – Elo	`home_elo, away_elo, diff_elo, home_elo_prob, away_elo_prob`	5
2 – Forma (últ. 5 jogos)	`{home,away}_form_{gols_marcados, gols_sofridos, pontos, taxa_vitoria}_last5`, geral e restrita ao mando (`*_home_last5`/`*_away_last5`)	16
3 – Temporada + diferenciais	`{home,away}_season_{matches, position, points_pg, goals_scored_pg, goals_conceded_pg, goal_diff_pg}` (12) + 5 diferenciais `diff_season_*`	17
4 – Calendário/agenda (só mandante)	`days_rest_home, rest_difference, fixture_congestion_index, opp_avg_elo_last5, strength_of_schedule_last5, elo_change_last5`	6
Total		**44**

Bloco 1 (Elo). Sistema iterativo em ordem cronológica, *rating* inicial 1.500, $K = 32$, vantagem de mando de 100 pontos, sem reinício entre temporadas: $E_h = [1 + 10^{(R_a - R_h - 100)/400}]^{-1}$, $E_a = 1 - E_h$, atualizado por $R_i' = R_i + K(S_i - E_i)$, $S_i \in \{1;\ 0{,}5;\ 0\}$. **Bloco 2 (forma).** Média móvel de 5 jogos por equipe e estatística, no total e restrita ao mando de campo (4 estatísticas $\times$ 2 escopos $\times$ 2 equipes). **Bloco 3 (temporada).** Acumuladas desde o início da temporada corrente, para mandante e visitante, mais cinco diferenciais $\Delta v = v_h - v_a$. **Bloco 4 (calendário).** Extraído apenas da perspectiva do mandante, sem contrapartida para o visitante, assimetria que ajuda a explicar o efeito adverso destas variáveis sobre a Regressão Logística, quantificado por ablação ao final da Seção 4.

Valores ausentes nas primeiras rodadas permanecem nulos nas agregações (`min_periods=1`) e são tratados por zero de forma uniforme nos nove modelos, antes de qualquer *pipeline*; incidência de 0,3%–2,7% das linhas.

Auditoria de vazamento temporal. Durante a inspeção, modelos simples chegaram a 99,7% de acurácia, resultado trivialmente impossível em três classes aproximadamente equiprováveis. Causa raiz: 19 colunas derivadas do placar final (`FTHG`, `FTAG` e 17 *features* como `home_goals_last5`) eram calculadas incluindo a partida-alvo antes de

[1] Código, *scripts* de preparação de dados e resultados completos: `https://github.com/wiliafsilva/artigopremierligue`. Python 3.11; `scikit-learn` 1.9.0, `XGBoost` 3.2.0, `LightGBM` 4.7.0, `CatBoost` 1.2.10, `pandas` 3.0.3, `NumPy` 2.4.6, `SciPy` 1.17.1. Semente fixa `SEED=42` em todos os experimentos; *bootstrap* com 5.000 réplicas. Dados brutos de domínio público em Football-Data.co.uk.

qualquer divisão temporal [Kapoor and Narayanan 2023]. Removidas, o conjunto limpo foi comparado ao *baseline* por *bootstrap* pareado (IC95%: $[-0{,}053, +0{,}052]$, cruzando zero), confirmando que a melhoria anterior era espúria; as 44 *features* remanescentes constituem a base de todos os experimentos.

3.2. Estratégia de Validação Temporal

Foram adotados dois esquemas complementares. O ***Design* A** (*split* fixo, $N = 380$) treina em 2010/11–2022/23 e avalia em 2023/24; eficiente, mas insuficiente para detectar diferenças menores que $\sim 2\%$ em *Log Loss*. O ***Design* B** (*walk-forward*, $N = 2.280$) executa 6 *folds* com treino crescente e validação sucessiva em 2018/19–2023/24, acumulando poder estatístico para diferenças da ordem de 1% [Hyndman and Athanasopoulos 2021]. O teste final ($N = 760$, 2024/25–2025/26) foi usado uma única vez, após autorização explícita.

A avaliação no teste foi restrita, deliberadamente, aos dois modelos que o *walk-forward* não conseguiu separar: Poisson e Regressão Logística ajustada. Submeter os nove ao conjunto retido o converteria em um segundo conjunto de validação, incorporando o ruído amostral favorável àquelas duas temporadas; por isso a comparação entre os nove é reportada exclusivamente no *walk-forward* ($N = 2.280$), cabendo ao retido apenas verificar se o modelo selecionado se sustenta fora da amostra.

3.3. Modelos Avaliados

Foram comparados 9 modelos em quatro níveis de complexidade crescente: **L1 — Simples** (Regressão Logística, Naive Bayes); **L2 — *Ensemble*** (*Random Forest*, *Gradient Boosting*); **L3 — *Gradient Boosting* Avançado** (XGBoost [Chen and Guestrin 2016], LightGBM [Ke et al. 2017], CatBoost [Prokhorenkova et al. 2018]); **L4 — *Kernel*/Probabilístico** (SVM, Poisson). O Poisson ajusta dois `PoissonRegressor` independentes (alvos `FTHG` e `FTAG`), cada um em *pipeline* `SimpleImputer` $\rightarrow$ `StandardScaler` $\rightarrow$ `PoissonRegressor`, com ligação logarítmica $\lambda = \exp(x'\beta)$ e regularização L2 no valor padrão ($\alpha = 1{,}0$), não ajustada; ambos recebem as 44 *features*. A distribuição conjunta é o produto externo das *pmf* truncadas em 8 gols (matriz 9×9), de onde $P(H)$ é a soma da triangular inferior, $P(D)$ o traço e $P(A)$ a triangular superior, as três renormalizadas pela soma para compensar a massa truncada acima do corte [Penn and Donnelly 2022]; os demais oito recebem calibração isotônica via `CalibratedClassifierCV` com partição temporal (`TimeSeriesSplit`), detalhada na Seção 4 [Silva Filho et al. 2023].

Ajuste de hiperparâmetros. Em *split* interno temporal (treino 2010/11–2020/21, validação 2021/22–2022/23), sem acesso ao teste: Regressão Logística ($C = 0{,}01$, L2, $100\times$ mais forte que o padrão); CatBoost (`depth=6`, `l2_leaf_reg=3`, coincidente com o padrão); LightGBM (`lr=0,03`, `num_leaves=15`, `min_child_samples=20`, capacidade reduzida frente ao padrão, que sobreajustava).

3.4. Protocolo de Comparação Estatística

Toda comparação foi conduzida por *bootstrap* pareado com 5.000 réplicas sobre as contribuições individuais de *Log Loss* por partida, seguindo Efron e Tibshirani [Efron and Tibshirani 1994] e o protocolo de Demšar [Demšar 2006]. O pareamento elimina a variabilidade da composição da amostra, já que todos os modelos são avaliados sobre as mesmas partidas.

Dois esquemas complementares foram empregados. No modo agregado (*pooled*), reamostram-se partidas individuais, pressupondo independência, hipótese violada, já que partidas de uma mesma temporada compartilham elenco, calendário e contexto, o que estreita artificialmente os intervalos. No modo em blocos (*cluster* por temporada), reamostram-se temporadas inteiras, preservando a dependência interna; é a variante de referência, por reconhecer que o número efetivo de observações independentes é da ordem do número de temporadas.

O critério de decisão baseia-se no IC95% de Δ = *Log Loss*(candidato) $-$ *Log Loss*(Poisson): positivo indica candidato significativamente pior; negativo, significativamente melhor; contendo zero, ausência de diferença detectável, não evidência de equivalência, que demandaria teste formal com margem pré-especificada, não adotado aqui. Os dois esquemas concordaram nas dezesseis comparações da Tabela 3 (*walk-forward*, $N = 2.280$): nenhum veredito mudou entre os modos, ainda que os intervalos por temporada sejam, como esperado, ligeiramente mais largos. A métrica principal é o *Log Loss* [Wheatcroft 2021]; *Brier Score*, *Accuracy*, F1-Macro, *Balanced Accuracy*, *Precision* e *Recall* Macro serviram só para diagnóstico.

4. Resultados e Discussão

Os experimentos seguiram quatro linhas independentes: **V2 – *Leakage*** (*features* de gols históricos — *rejeitada*); **V3 – *MVPs*** (6 *features* de agenda/*momentum* — *rejeitada*); **Elo+*Odds*** (*odds* de *bookmaker* — *inconclusiva*); e Comparação (modelos mais complexos superam o Poisson — *Poisson como modelo final*).

Comparação Sistemática dos 9 Modelos. Na avaliação *walk-forward* sem ajuste de hiperparâmetros ($N = 2.280$), o Poisson liderou com *Log Loss* de 0,9672. L2 e L3 apresentaram miscalibração, evidenciada pela inversão F1-Macro/*Log Loss*: LightGBM e *Random Forest* lideraram em F1-Macro mas ficaram nas últimas posições em *Log Loss*, o que motivou calibração isotônica via `CalibratedClassifierCV` [Silva Filho et al. 2023].

A calibração foi originalmente conduzida com validação cruzada aleatória de cinco partições (`cv=5`). Identificou-se, em revisão, que isso permitia que partidas cronologicamente posteriores participassem da calibração de partidas anteriores, vazamento temporal; a piora observada em alguns modelos vinha desse vazamento, não de sobreajuste do calibrador. Os resultados a seguir empregam partição temporal (`TimeSeriesSplit`, 5 divisões), aplicada uniformemente aos oito modelos discriminativos — incluindo Regressão Logística e Naive Bayes, que antes não recebiam calibração, e LightGBM, com API ajustada para compatibilidade. O Poisson permanece sem calibração por ser generativo. A Tabela 2 apresenta o *Log Loss* antes e depois da correção.

Tabela 2. Efeito da calibração temporal no *Log Loss* e no ECE — *walk-forward* ($N = 2.280$).

Modelo	LL pré-cal.	LL pós-cal.	Δ LL	ECE pré	ECE pós
Poisson (referência)	*0,9672*	—	—	—	—
Regressão Logística	0,9807	**0,9746**	$-0,0061$	0,0208	0,0252
CatBoost	0,9746	0,9789	$+0,0043$	0,0104	0,0143
Naive Bayes	3,8970	0,9797	$-2,9172$	0,3940	0,0245
Gradient Boosting	0,9853	0,9905	$+0,0053$	0,0155	0,0317
XGBoost	1,0047	0,9916	$-0,0130$	0,0434	0,0394
Random Forest	1,0011	0,9917	$-0,0094$	0,0246	0,0264
LightGBM	1,0132	0,9928	$-0,0204$	0,0535	0,0399
SVM	1,0007	1,0046	$+0,0040$	0,0504	0,0316

Δ LL = pós-cal. − pré-cal.; negativo = calibração melhora. ECE com 10 *bins*.

Com a partição temporal, o efeito da calibração deixa de ser uniformemente favorável: o *Log Loss* melhora em cinco dos oito modelos e o ECE em quatro. As duas métricas divergem em dois casos, a Regressão Logística melhora em *Log Loss* e piora em ECE, enquanto o SVM faz o inverso, sinal de que o ganho de calibração é compensado por perda de resolução e no *Random Forest* o *Log Loss* melhora sem efeito apreciável sobre o ECE. O Naive Bayes, antes tido por miscalibração catastrófica, torna-se competitivo (3,897 → 0,9797; ECE 0,394 → 0,025) com o mesmo tratamento dos demais, o problema nunca foi o modelo, e sim a ausência de calibração, da qual ele depende mais que os outros. A Regressão Logística também se beneficia em *Log Loss* ($-0,0061$), contrariando a premissa de que dispensaria calibração; com esse ganho, passa ao segundo lugar geral (0,9746), atrás só do Poisson. A Figura 1 traz os diagramas de confiabilidade de quatro modelos representativos, o caso extremo (Naive Bayes), os dois em que *Log Loss* e ECE divergem (Regressão Logística e SVM) e o caso sem efeito sobre o ECE (*Random Forest*); os oito estão na Tabela 2.

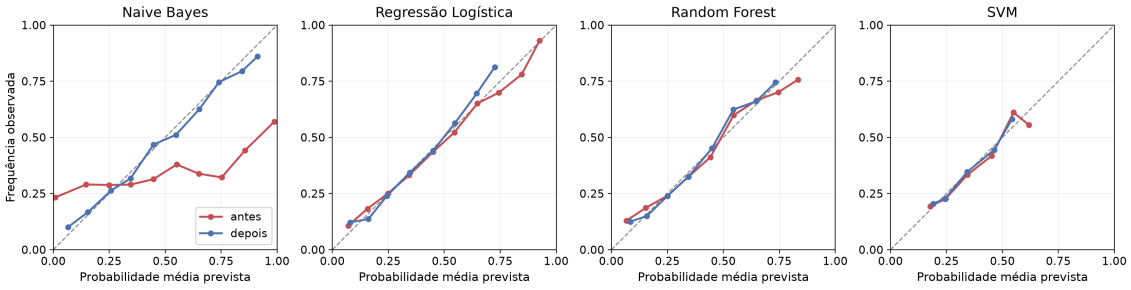


Figura 1. Diagramas de confiabilidade *one-vs-rest* de quatro modelos representativos — antes vs. depois da calibração temporal (*walk-forward*, $N = 2.280$). O ECE da Tabela 2 usa confiança máxima (*top-label*).

Comparação por Bootstrap Pareado. A escolha do Poisson é sustentada pelas dezesseis comparações da Tabela 3 contra os oito discriminativos calibrados, os candidatos tunados e as três especificações generativas, por *bootstrap* pareado (5.000 réplicas) nos dois esquemas: agregado e em blocos.

O Poisson supera, com significância estatística em ambos os esquemas, todos os oito modelos calibrados, inclusive a Regressão Logística calibrada o que inverte o veredito de "sem diferença detectável" reportado antes da correção do vazamento na calibração. Entre os tunados, só a Regressão Logística tunada ($C = 0,01$, L2), sem calibração, tem IC95% contendo zero em ambos os esquemas: sem diferença detectável em relação ao

Tabela 3. Poisson vs. modelos calibrados, candidatos tunados e especificações generativas *walk-forward* ($N = 2.280$).

Grupo	Candidato	Δ	IC95% agregado	IC95% em blocos
Calibrados	Regressão Logística (cal.)	$+0,0075$	$[+0,0014,\ +0,0134]$	$[+0,0011,\ +0,0146]$
	CatBoost (cal.)	$+0,0118$	$[+0,0053,\ +0,0186]$	$[+0,0027,\ +0,0206]$
	Naive Bayes (cal.)	$+0,0126$	$[+0,0060,\ +0,0189]$	$[+0,0066,\ +0,0179]$
	Gradient Boosting (cal.)	$+0,0234$	$[+0,0158,\ +0,0307]$	$[+0,0077,\ +0,0392]$
	XGBoost (cal.)	$+0,0245$	$[+0,0164,\ +0,0324]$	$[+0,0115,\ +0,0400]$
	Random Forest (cal.)	$+0,0246$	$[+0,0165,\ +0,0326]$	$[+0,0115,\ +0,0374]$
	LightGBM (cal.)	$+0,0256$	$[+0,0169,\ +0,0343]$	$[+0,0102,\ +0,0411]$
	SVM (cal.)	$+0,0375$	$[+0,0281,\ +0,0474]$	$[+0,0215,\ +0,0534]$
Tunados	CatBoost (padrão = tunado)	$+0,0067$	$[+0,0014,\ +0,0121]$	$[+0,0013,\ +0,0115]$
	LightGBM (*default*)	$+0,0887$	$[+0,0697,\ +0,1084]$	$[+0,0622,\ +0,1180]$
	LightGBM (tunado)	$+0,0182$	$[+0,0101,\ +0,0261]$	$[+0,0083,\ +0,0287]$
	Regressão Logística (*default*)	$+0,0135$	$[+0,0004,\ +0,0318]$	$[+0,0020,\ +0,0283]$
	Regressão Logística (tunada)	$+0,0058$	$[-0,0036,\ +0,0186]$	$[-0,0036,\ +0,0178]$
Especificações generativas	Força de equipe, $\rho = 0$	$+0,0177$	$[+0,0070,\ +0,0290]$	$[+0,0065,\ +0,0307]$
	Dixon-Coles (ρ livre)	$+0,0191$	$[+0,0082,\ +0,0305]$	$[+0,0076,\ +0,0321]$
	Dixon-Coles + decaimento	$+0,0083$	$[-0,0016,\ +0,0183]$	$[-0,0061,\ +0,0286]$

Δ = *Log Loss*(candidato) − *Log Loss*(Poisson); positivo = Poisson melhor. Nas linhas *default*, os hiperparâmetros são os padrões da biblioteca (LightGBM `lr` = 0,1); a Tabela 2 (pré-cal.) usa `lr` = 0,05.

Poisson. As dezesseis comparações da Tabela 3 concordam integralmente entre agregado e em blocos.

Teste Final Independente. No teste final ($N = 760$, 2024/25–2025/26, avaliado uma única vez), o Poisson obteve *Log Loss* de 1,0099 e acurácia de 50,92% (Tabela 4).

Tabela 4. Métricas no teste final independente ($N = 760$).

Métrica	Poisson	*Logistic* (C=0,01)	Melhor
Log Loss	**1,0099**	1,0142	Poisson
Brier Score	**0,2017**	0,2029	Poisson
Accuracy	**50,92%**	50,92%	Empate
Balanced Acc.†	**44,43%**	44,40%	Poisson
F1-Macro	0,3784	**0,3788**	Logistic ($\Delta = 0{,}0004$)
Precision Macro	0,3375	**0,3392**	Logistic ($\Delta = 0{,}0017$)

† Idêntica ao *Recall* Macro por definição neste desenho.

As vantagens da Logística em F1-Macro ($\Delta = 0{,}0004$) e *Precision* Macro ($\Delta = 0{,}0017$) são marginais frente ao ruído do *bootstrap* (IC95%: agregado $[-0{,}0045, +0{,}0132]$; em blocos $[-0{,}0034, +0{,}0122]$). Nenhuma combinação de critérios mostra a Logística claramente superior, coerente com a literatura [Hubáček et al. 2022, Fischer and Heuer 2024].

Desempenho por temporada. Desagregando o teste por temporada ($N = 380$ cada): em 2024/25, *Log Loss* de 0,9855 (Poisson) vs. 0,9821 (Logística); em 2025/26, 1,0342 vs. 1,0464 a distância inverte de sinal entre as temporadas, então o agregado não é dominado por uma única temporada favorável.

Nenhum modelo prevê empates. A Tabela 5 apresenta a matriz de confusão do Poisson no teste final. Nas 760 partidas, o modelo prevê zero empates; a Regressão Logística ajustada prevê apenas dois. A consequência direta é F1 = 0,000 para a classe D (empate) em ambos os modelos, com *Recall* e *Precision* também nulos. Nas demais classes, o Poisson atinge F1 = 0,636 em H (*Precision* 0,515; *Recall* 0,833) e F1 = 0,499 em A (0,498; 0,500); o F1-Macro de 0,3784 é, portanto, a média de dois valores moderados e de um zero.

Tabela 5. Matriz de confusão do Poisson — teste final ($N = 760$).

Real \ Previsto	H	D	A
H (317)	264	**0**	53
D (197)	126	**0**	71
A (246)	123	**0**	123

Empates representam 22,1% do *walk-forward* (503/2.280) e 25,9% do teste (197/760), mas nunca são a classe modal sob argmax em nenhum dos nove modelos; a calibração isotônica, embora melhore o *Log Loss*, reduz ainda mais essa já escassa capacidade de prever empates, efeito colateral não documentado antes. Isso justifica o *Log Loss* como métrica principal: sob *Accuracy* ou F1, "nunca prever empate" seria indistinguível de um modelo que atribui $P(D) \approx 0{,}26$ de forma consistentemente correta o que o Poisson faz.

Comparação com Dixon-Coles. Para verificar se a omissão de uma parametrização clássica escondia um modelo superior, comparou-se o Poisson por regressão com três especificações por força de equipe ($N = 2.280$): sem correção ($\rho = 0$), Dixon-Coles com ρ livre [Dixon and Coles 1997] e com decaimento temporal ($\xi = 0{,}0018$/dia), bloco "Especificações generativas" da Tabela 3.

O ρ estimado é consistentemente negativo ($-0,037$ a $-0,054$ nos seis *folds*), na direção prevista por Dixon e Coles [Dixon and Coles 1997]: marginais independentes subestimam placares baixos, sobretudo empates. Mas o mecanismo não se traduz em ganho, ρ livre (0,9863) é pior que a mesma parametrização sem correção (0,9849), sugerindo que essa dependência já é capturada pelas *features* de forma e Elo, enquanto o decaimento temporal isolado (0,9754) indica maior peso da recência. Em ambos os modos, o Poisson supera duas das três alternativas; da variante com decaimento não se distingue, com IC95% contendo zero.

Incorporação de *odds* de *bookmakers*. As probabilidades implícitas do mercado (Bet365, colunas `B365H/D/A` do Football-Data.co.uk, cobertura $\geq 95\%$) foram extraídas das *odds* decimais pela remoção proporcional da margem da casa (*overround*): $p_i = (1/\text{odd}_i)/\sum_{j\in\{H,D,A\}} 1/\text{odd}_j$. Como *feature* adicional ao Poisson (linha de investigação Elo+*Odds*), essas probabilidades não produziram ganho estatisticamente significativo em nenhum dos três desenhos de *bootstrap* avaliados, *split* único, *walk-forward* agregado e em blocos, todos com IC95% contendo zero.

Aplicabilidade a competições regionais e cenários de baixa disponibilidade de dados. Os resultados têm implicação direta para competições regionais (estaduais, divisões de acesso, ligas menores), cuja disponibilidade de dados é ordens de grandeza inferior à das 4.940 partidas de treino aqui empregadas, regime em que se espera que a vantagem da parcimônia se amplie. As 44 *features* exigem janelas móveis e Elo contínuo, que um histórico curto dificilmente sustenta, enquanto o Dixon-Coles opera desde a primeira rodada. São hipóteses derivadas dos resultados, não achados empíricos.

Origem e retenção das *features* da Linha V3. As seis variáveis do Bloco 4 originaram-se da própria Linha V3 e foram classificadas como não promissoras ao seu encerramento, tendo sido descontinuadas no construtor de *features*. Elas permaneceram, contudo, no conjunto processado que alimenta os experimentos, porque os *scripts* de avaliação selecionam colunas por exclusão de metadados e alvos, e não pela lista

do construtor: a retenção é uma limitação do *pipeline*, não uma escolha metodológica. Um teste de ablação ($N = 2.280$) quantifica o efeito. A exclusão das seis não altera o Poisson ($\Delta = 0{,}0000$; IC95% $[-0{,}0010, +0{,}0011]$ agregado e $[-0{,}0015, +0{,}0012]$ em blocos) e melhora significativamente a Regressão Logística tunada ($\Delta = -0{,}0078$; $[-0{,}0198, -0{,}0009]$ e $[-0{,}0201, -0{,}0009]$), o que inverte o ponto-estimado do confronto direto entre os dois. O veredito não muda: com 44 ou com 38 *features*, o IC95% do confronto contém zero nos dois esquemas, e a ausência de diferença detectável entre Poisson e Regressão Logística tunada se sustenta em ambas as variantes. A assimetria é coerente com o Bloco 4 ser extraído apenas da perspectiva do mandante (Seção 3.1) e constitui limitação do conjunto de *features*, não do protocolo.

Limitações, os sete modelos descartados no *walk-forward* não foram avaliados no teste, preservando a estimativa não enviesada do modelo selecionado mas impedindo afirmações sobre seu desempenho absoluto em 2024/25–2025/26. A reamostragem por temporada no teste dispõe de apenas duas unidades: ali o modo em blocos é verificação de robustez, não teste primário; a inferência principal apoia-se no *walk-forward*, com seis temporadas.

5. Conclusão

Nove modelos foram comparados sob protocolo de validação temporal rigoroso, com todas as comparações sustentadas por *bootstrap* pareado (5.000 réplicas) em dois esquemas concordantes. O Poisson foi selecionado por parcimônia e interpretabilidade: superou com significância estatística os oito discriminativos calibrados, incluindo a Regressão Logística calibrada, e duas das três especificações por força de equipe. As exceções são a Regressão Logística tunada ($C = 0{,}01$, L2) sem calibração e a variante Dixon-Coles com decaimento temporal, das quais permanece indistinguível.

A contribuição metodológica mais relevante é mostrar que vazamento temporal pode inflar acurácias de 51% para 99,7%, e que testes inadequados levam a conclusões não reprodutíveis, mesmo quando o vazamento é sutil, como na calibração, problemas que *walk-forward* [Hyndman and Athanasopoulos 2021] e *bootstrap* pareado em dois modos [Efron and Tibshirani 1994] eliminaram.

Como trabalhos futuros: *odds* como *prior* bayesiano; dados contextuais ausentes (lesões, suspensões, viagem); outras ligas; e, sobretudo, replicação em competições regionais, para testar se a vantagem de parcimônia cresce com menos dados e quando o Dixon-Coles supera a especificação por variáveis.

Referências

Bentéjac, C., Csörgő, A., and Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3):1937–1967.

Bunker, R. and Susnjak, T. (2022). The application of machine learning techniques for predicting match results in team sport: a review. *Journal of Artificial Intelligence Research*, 73:1285–1322.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proc.*

22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD '16), pages 785–794. ACM.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

Dixon, M. J. and Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265–280.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1st edition.

Fischer, M. and Heuer, A. (2024). Match predictions in soccer: machine learning vs. Poisson approaches. arXiv preprint.

Galekwa, R. M. et al. (2024). A systematic review of machine learning in sports betting: techniques, challenges, and future directions. arXiv preprint.

Hubáček, O., Šourek, G., and Železný, F. (2022). Forty years of score-based soccer match outcome prediction: an experimental review. *IMA Journal of Management Mathematics*, 33(1):1–18.

Hyndman, R. J. and Athanasopoulos, G. (2021). *Forecasting: principles and practice*. OTexts, Melbourne, 3rd edition.

Julian, R., Page, R. M., and Harper, L. D. (2021). The effect of fixture congestion on performance during professional male soccer match-play: A systematic critical review with meta-analysis. *Sports Medicine*, 51(2):255–273.

Kapoor, S. and Narayanan, A. (2023). Leakage and the reproducibility crisis in machine learning-based science. *Patterns*, 4(9):100804.

Ke, G. et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Inf. Proc. Systems (NeurIPS)*, volume 30, pages 3146–3154.

Nguyen, Q. (2021). Poisson modeling and predicting English Premier League goal scoring. *The New England Journal of Statistics in Data Science*, 1(1):95–101.

Penn, M. J. and Donnelly, C. A. (2022). Analysis of a double Poisson model for predicting football results in Euro 2020. *PLOS ONE*, 17(5):e0268511.

Prokhorenkova, L. et al. (2018). CatBoost: unbiased boosting with categorical features. In *Advances in Neural Inf. Proc. Systems (NeurIPS)*, volume 31, pages 6638–6648.

Silva Filho, T. et al. (2023). Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112:3211–3260.

Wheatcroft, E. (2021). Evaluating probabilistic forecasts of football matches: the case against the ranked probability score. *Journal of Quantitative Analysis in Sports*, 17(4):273–287.