

Lucas Teixeira Duarte

Previsão de jogos do Campeonato Brasileiro 2026: uma abordagem comparativa
de modelos estatísticos e de Machine Learning

PROJETO DE GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO
APRESENTADO AO DEPARTAMENTO DE ENGENHARIA INDUSTRIAL
DA PUC-RIO, COMO PARTE DOS REQUISITOS PARA OBTENÇÃO
DO TÍTULO DE BACHAREL EM ENGENHARIA DE PRODUÇÃO

Orientador: Igor Tona Peres

Departamento de Engenharia Industrial
Rio de Janeiro, 01 de junho de 2026.

RESUMO

A tentativa de prever resultados de jogos de futebol vem ganhando destaque no mundo, por conta de alguns fatores que ajudam no planejamento dos diretores dos departamentos de futebol dos clubes, mas principalmente pelo grande crescimento do número de casas de aposta e de seus usuários, que movimentam um intenso valor financeiro. Este estudo teve como objetivo aplicar três tipos de modelagem estatística em uma base de dados contendo os resultados de todos os jogos do Campeonato Brasileiro desde 2003 até a décima oitava rodada do campeonato de 2026, além do valor anual medido no início do ano dos elencos dos times participantes do Campeonato Brasileiro 2026, e comparar seus resultados. A base de dados foi reestruturada, sendo necessária a retirada de um time participante do campeonato que foi previsto por falta de histórico de dados, além da transformação dos dados dos resultados em outras variáveis que aumentam a assertividade da previsão. As modelagens escolhidas para as previsões foram utilizando probabilidades de Distribuição de Poisson, Regressão Logística e Machine Learning, com o algoritmo XGBoost. Os resultados indicam uma superioridade na previsão feita pelo modelo de Regressão Logística em relação ao modelo de XGBoost, observada através das métricas de avaliação dos modelos utilizadas. Foi observado também que os modelos tendem a favorecer a vitória do mandante como resultado da previsão. O presente trabalho contribui para o entendimento de fatores que interferem em um jogo de futebol e na classificação geral de um campeonato e abre a possibilidade de estudos futuros que podem utilizar outras variáveis e testar a assertividade de cada variável separadamente.

Palavras-chave: futebol, Poisson, Regressão Logística, Machine Learning, XGBoost, previsão, resultados, Campeonato Brasileiro 2026.

ABSTRACT

Attempts to predict football match results has been gaining attention around the world, due to a number of factors that assist club football department managers in their planning, but primarily because of the significant growth of betting houses and their users, who generate substantial amounts of money. The aim of this study was to apply three types of statistical modelling to a database containing the results of all Campeonato Brasileiro matches from 2003 up to the 18th round of 2026, as well as the annual value measured at the start of the year for the squads of the teams participating in the Campeonato Brasileiro 2026, and to compare their results. The database was restructured, requiring the removal of one team participating in the championship that was excluded due to a lack of historical data, as well as the transformation of the results data into other variables that increase the accuracy of the prediction. The models chosen for the predictions utilised Poisson distribution probabilities, logistic regression and machine learning, using the XGBoost algorithm. The results indicate that the Logistic Regression model outperforms the XGBoost model, as evidenced by the model evaluation metrics used. It was also observed that the models tend to predict a home win. This study contributes to the understanding of factors that influence a football match and the overall league standings, and provides the basis for future research that may utilise other variables and also test the accuracy of each variable separately.

Keywords: soccer, Poisson, Logistic Regression, Machine Learning, XGBoost, prediction, score, Campeonato Brasileiro 2026.

LISTA DE FIGURAS

Figura 3.1 – parte inicial da base de dados bruta	8
Figura 3.2 – parte inicial da base de dados reestruturada	9
Figura 3.5 - Exemplo de tabela das probabilidades de Poisson, confronto entre Athletico-PR e Atlético-MG	13

LISTA DE TABELAS

Tabela 3.1 – Caracterização de variáveis	11
Tabela 3.2 - Início da matriz de confrontos dos Valores Esperados dos mandantes	13
Tabela 3.3 - Início da matriz de confrontos dos Valores Esperados dos visitantes	13
Tabela 3.4 - Exemplo de tabela das probabilidades de Poisson, confronto entre Athletico-PR e Atlético-MG	13
Tabela 3.5 - Hiperparâmetros do modelo de XGBoost	15
Tabela 4.1 – Análise Descritiva (Variável de valor de elenco em milhões de euros)	16
Tabela 4.2 - Métricas de avaliação do modelo de Poisson	17
Tabela 4.3 – Matriz de confusão do modelo de Poisson	18
Tabela 4.4 – Tabela de classificação prevista pelo método de Distribuição de Poisson do Campeonato Brasileiro 2026	18
Tabela 4.5 – Métricas de avaliação do modelo de Regressão Logística	19
Tabela 4.6 – Matriz de confusão do modelo de Regressão Logística	19
Tabela 4.7 – Tabela de classificação prevista pelo método de Regressão Logística do Campeonato Brasileiro 2026	20
Tabela 4.8 - Métricas de avaliação do modelo de XGBoost	21
Tabela 4.9 – Matriz de confusão do modelo de XGBoost	21
Tabela 4.10 – Tabela de classificação prevista pelo método de Regressão Logística do Campeonato Brasileiro 2026	21

SUMÁRIO

1. INTRODUÇÃO	1
2. REFERENCIAL TEÓRICO	3
2.1. Formato de campeonato de futebol de pontos corridos	3
2.2. Modelos preditivos	3
2.2.1. Distribuição de Poisson	3
2.2.2. Regressão logística	4
2.2.3. Machine Learning	5
2.2.3.1. XGBoost	5
3. METODOLOGIA	7
3.1. Desenho do estudo	7
3.2. Framework metodológico	7
3.3. Ferramentas	7
3.4. Base de dados	8
3.5. Pré-processamento e Feature Engineering	10
3.6. Caracterização de variáveis	10
3.7. Métodos	12
3.7.1. Distribuição de Poisson	12
3.7.2. Regressão Logística	14
3.7.3. XGBoost	15
4. RESULTADOS	16
4.1. Análise Descritiva	16
4.2. Distribuição de Poisson	17
4.3. Regressão Logística Multinomial	18
4.4. XGBoost	20
5. DISCUSSÃO	23
6. CONCLUSÃO	24
7. REFERÊNCIAS BIBLIOGRÁFICAS	25

1. INTRODUÇÃO

O futebol faz parte da cultura de muitas pessoas mundo afora, e a relação que seus adeptos têm com o esporte é muito intensa, sendo milhões de torcedores fanáticos. Esse fanatismo gera um nível altíssimo de engajamento, que logo pode ser visto como uma oportunidade de negócio. Esses torcedores compram produtos, consomem conteúdo nas redes sociais e na televisão, participam de eventos das marcas relacionadas ao futebol, das federações, dos clubes, transformando esse grande engajamento em retorno financeiro (NIELSEN, 2025).

Em 2025, todos os 30 maiores faturamentos de clubes de futebol no mundo foram superiores a 188 milhões de euros, sendo os dois maiores perto de 1 bilhão de euros (GLOBO ESPORTE, 2026). No Brasil, as 3 maiores quantias de premiações recebidas por clubes são superiores a 300 milhões de reais (GLOBO ESPORTE, 2025). Uma pesquisa realizada pela Serasa indica que 61% da população brasileira afirma gastar com futebol. Portanto, conclui-se que o esporte futebol movimenta cifras muito altas. Entre outros motivos, o interesse no esporte vem crescendo, atraindo cada vez mais pessoas de diferentes áreas do conhecimento, sendo a Engenharia uma das mais aplicadas na atualidade (GRAND VIEW RESEARCH, 2026).

O interesse na previsibilidade nos resultados dos jogos de futebol tem crescido junto com o aumento das casas de apostas, ao redor do mundo, e principalmente no Brasil. Globalmente, o setor de apostas online cresceu 734% de 2021 até 2024. No Brasil havia 26 empresas que têm como atividade principal apostas online em 2021, e no ano seguinte, em 2022, esse número cresceu 203%, totalizando 79 empresas (CNN BRASIL, 2024). Dos 20 clubes da série A do Campeonato Brasileiro 2025, todos são patrocinados por casas de aposta, sendo em 90% dos clubes o patrocinador master (GLOBO ESPORTE, 2025). Apostar em resultados de jogos de futebol com base em dados e estatística pode aumentar significativamente a chance de ter resultados positivos nas apostas, e consequentemente, ganho financeiro.

Com as informações que temos disponíveis hoje sobre o que envolve uma partida de futebol, é possível perceber que existem muitos fatores sem ser apenas os dados históricos dos resultados que influenciam. Por exemplo, o clima na hora e no dia da partida, o aspecto psicológico dos envolvidos nos jogos, mudança/transferência de jogadores e treinadores nos clubes. Portanto, basear previsões somente em modelagem pode ser desafiador.

Este trabalho terá como principal objetivo a aplicação de métodos de análise preditiva utilizando Poisson, regressão logística e o algoritmo XGBoost de Machine Learning em

resultados de partidas do Campeonato Brasileiro de futebol, que é a competição mais antiga disputada por clubes no Brasil e quem o ganha é chamado de Campeão Brasileiro. É o torneio de maior prestígio no território nacional. Os objetivos específicos do trabalho são: comparar a eficiência dos modelos aplicados por base de métricas de avaliação; discutir o impacto de variáveis não analisadas nos resultados dos modelos.

Este trabalho está estruturado em 6 capítulos. O capítulo 2 apresenta o Referencial Teórico, detalhando os conceitos de formato de campeonato de pontos corridos e dos modelos aplicados. O capítulo 3 apresenta a Metodologia, incluindo framework metodológico, base de dados, ferramentas utilizadas, pré-processamento de dados e feature engineering. O capítulo 4 apresenta os resultados das aplicações, os quais são interpretados e discutidos no capítulo 5 (Discussão). Por fim, o capítulo 6 apresenta as conclusões do estudo e sugestões para trabalhos futuros.

2. REFERENCIAL TEÓRICO

O capítulo tem como objetivo apresentar os fundamentos teóricos importantes para a compreensão do trabalho.

2.1. Formato de campeonato de futebol de pontos corridos

Um campeonato de pontos corridos, como é a forma de disputa do Campeonato Brasileiro desde 2003 até a edição atual, é uma competição onde todas as equipes se enfrentam duas vezes, cada vez em um turno, que são chamados de ‘Primeiro Turno’ e ‘Segundo Turno’. Para exemplificar, em um confronto fictício da Equipe 1 contra a Equipe 2, no primeiro turno a partida seria realizada no estádio que pertence à Equipe 1, que estaria jogando em casa, como mandante, e a Equipe 2 estaria jogando fora de casa, como visitante. No segundo turno, a partida seria realizada no estádio que pertence à Equipe 2, e nesse caso, a Equipe 2 estaria jogando como mandante e a Equipe 1 como visitante (CBF, 2026).

Em relação à pontuação do campeonato, existem 3 possibilidades em cada jogo para uma equipe: uma vitória faz a equipe somar 3 pontos na classificação geral; uma derrota faz a equipe não somar pontos; um empate distribui 1 ponto para as duas equipes do confronto. A cada rodada os pontos conquistados pelas equipes são somados, e ao final de todas as rodadas temos a classificação do campeonato (CBF, 2026).

No Brasil e na maioria das ligas nacionais dos países pelo mundo, os 4 primeiros da classificação geral se qualificam diretamente para o maior torneio internacional do continente, que no caso da América do Sul, esse torneio chama-se Libertadores da América. Os 4 últimos colocados na classificação geral são rebaixados para a divisão inferior, não disputando a série A (primeira divisão) no ano seguinte (CBF, 2026).

2.2. Modelos preditivos

Neste tópico são detalhados os modelos que foram aplicados no presente estudo, Distribuição de Poisson, Regressão Logística e o algoritmo XGBoost de Machine Learning.

2.2.1. Distribuição de Poisson

A Distribuição de Poisson é uma distribuição de probabilidade discreta que descreve a probabilidade de um evento ocorrer um certo número de vezes (k) dentro de um intervalo específico de tempo. Esta distribuição é caracterizada por apenas um parâmetro λ (lambda) que representa o número médio de eventos. Ela pode ser utilizada para encontrar a probabilidade de ocorrer 0,1,2,... eventos em um intervalo de tempo a partir do conhecimento do número médio de eventos independentes que ocorrem nesse intervalo (FONSECA JÚNIOR, 2025).

A variável aleatória em uma Poisson, que assume um valor dentre os possíveis, normalmente chamada de X , é sempre o número de ocorrências em um intervalo de tempo. Por exemplo, o número de chamadas telefônicas em 1 minuto ou o número de gols em uma partida. Suas premissas se encaixam no contexto de uma partida de futebol, pois envolvem independência entre os eventos, taxa constante de ocorrência, baixa probabilidade de múltiplos eventos em um intervalo muito pequeno de tempo e proporcionalidade da taxa ao tamanho do intervalo (FONSECA JÚNIOR, 2025). A equação (1) representa a fórmula da Distribuição de Poisson, que calcula a probabilidade do valor assumido por X acontecer em um certo intervalo de tempo é apresentada abaixo:

$$P(X = k) = (\lambda^k * e^{-\lambda}) / k! \quad (1)$$

2.2.2. Regressão Logística

Regressão logística é uma técnica de mineração de dados, que são organizadas com o intuito de descobrir padrões úteis nos dados. Entre as tarefas de mineração de dados, está a tarefa de previsão, que consiste em prever o valor de um atributo com base no valor de outros atributos. Em regressão logística, a variável dependente é categórica e frequentemente binária, isto é, pode assumir somente dois valores, normalmente é "sucesso" ou "fracasso", "1" ou "0" (GONZALEZ, 2018).

A regressão logística também é uma técnica estatística que tem como objetivo produzir um modelo que permita a predição de valores tomados por uma variável dependente em função de uma ou mais variáveis independentes (LOHMANN, 2011). Segundo Hosmer e Lemeshow, métodos de regressão têm se tornado um componente integral para qualquer análise de dados interessada em descrever a relação entre uma variável resposta (dependente) e uma ou mais variáveis.

A técnica de regressão logística é capaz de obter a influência que cada variável independente tem no evento estudado. Para estimativa dos parâmetros de um modelo de regressão logística é utilizado o método de máxima verossimilhança. Este método, assim como o método de mínimos quadrados, permite a estimação de parâmetros de modelos e a realização de testes de hipóteses relativos a restrições lineares e não lineares. As estimações de máxima verossimilhança destes parâmetros são obtidas por métodos iterativos (LOHMANN, 2011).

2.2.3. Machine Learning

São métodos computacionais que tratam e processam uma grande quantidade de dados. Os algoritmos de Machine Learning são divididos em três grupos: Aprendizado Supervisionado (AS), Aprendizado Não Supervisionado (ANS) e o Aprendizado por Reforço (AR). Dentro da primeira opção, AS, os algoritmos têm como objetivo prever resultados futuros a partir de uma entrada de dados e um treinamento. O ANS visa uma análise descritiva buscando agrupamento dos dados e o AR toma decisões ótimas em tempo real. Serão utilizados somente os algoritmos de AS, pois são os modelos mais indicados para eventos esportivos. Eles são utilizados para treinar um modelo e achar padrões em uma base de dados com variáveis bem definidas, e posteriormente, usa o modelo treinado para prever os valores da variável dependente no tempo de interesse (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

2.2.3.1. XGBoost

Por questões de simplificação, vamos testar um dos métodos mais utilizados de Machine Learning, o XGBoost. Esse método é comumente reportado como um método que obtém resultados superiores na literatura de Machine Learning (PERES et al, 2022).

O XGBoost ou Extreme Gradient Boosting, traduzido em Impulsioneamento Extremo de Gradiente, é uma biblioteca de aprendizado de máquina (Machine Learning) que emprega o algoritmo de boosting com árvores de decisão (CHEN; GUESTRIN, 2016).

Este algoritmo envolve criar árvores de decisão de maneira sequencial. Todas as variáveis independentes são ponderadas de acordo com sua influência no resultado real. Se a árvore faz uma decisão errada em relação ao que aconteceu na realidade, o peso das variáveis é ajustado para a próxima árvore (BHATT et al, 2023).

O termo "gradient boosting" vem da ideia de "boosting" um modelo fraco, ou seja, utilizar vários modelos fracos e combinar a informação adquirida em cada um para gerar um modelo forte. O termo "gradient" complementa o "boosting", pois o processo de trabalhar com vários modelos fracos é conhecido como um algoritmo de gradiente descendente, o qual é aplicado a uma função objetivo e funciona achando sua melhor resposta (NVIDIA, s.d.). A função objetivo contém uma função de perda que representa a diferença entre os valores corretos e os previstos e um termo de regularização (EKANAYAKE; MEDDAGE; RATHNAYAKE, 2022).

Assim como outras metodologias de boosting, o gradient boosting começa com um aprendiz fraco e a partir das previsões desse aprendiz fraco, cria outras árvores de decisão com base nos erros do aprendiz inicial. O algoritmo então, calcula o erro dos valores da previsão de cada árvore em relação aos valores reais. Esses erros são utilizados para pontuar o modelo e agregados a função de perda da função objetivo. O gradiente descendente é um algoritmo de otimização utilizado para minimizar a função de perda em problemas de Machine Learning (CHEN e GUESTRIN, 2016).

3. METODOLOGIA

Neste capítulo, é detalhada a metodologia do trabalho, passando por fonte de dados, pré-processamento dos dados, feature engineering, caracterização de variáveis, métodos e ferramentas utilizadas.

3.1. Desenho do estudo

O presente estudo consiste na aplicação de 3 métodos probabilísticos preditivos, Distribuição de Poisson, Regressão Logística e XGBoost em uma base de dados API de jogos de futebol, extraída do site GitHub, com o objetivo de ter a maior porcentagem de assertividade possível na previsão de jogos do Campeonato Brasileiro 2026, e comparar os métodos utilizados. O trabalho caracteriza-se como uma pesquisa exploratória comparativa, de caráter experimental e com abordagem quantitativa.

3.2. Framework metodológico

O framework metodológico do trabalho seguiu os seguintes passos: (i) foi realizada a extração dos dados necessários e a criação de variáveis no mesmo padrão em que veio a base de dados inicial, com todas as informações sendo separadas por mandante e visitante; (ii) análise descritiva inicial, comparando os grupos dos mandantes vencedores, mandantes perdedores, visitantes vencedores e visitantes perdedores; (iii) aplicação dos métodos probabilísticos preditivos escolhidos, Distribuição de Poisson, Regressão Logística e XGBoost.

3.3. Ferramentas

Para a organização e pré-processamento dos dados e aplicação do método preditivo utilizando Distribuição de Poisson foi utilizada a ferramenta Excel, e para a aplicação dos outros métodos probabilísticos preditivos, Regressão Logística e do algoritmo de Machine Learning, XGBoost, foi utilizada a ferramenta Python.

3.4. Base de dados

A base de dados foi primária, extraída em Excel, contendo as informações dos resultados de todos os jogos do Campeonato Brasileiro desde 2003 até 2025. A planilha bruta vem com todas as informações de uma linha em apenas uma célula, separadas por aspas e vírgula, conforme mostrado na Figura 3.1.

ID,"rodada","data","hora","mandante","visitante","formacao_mandante","formacao_visitante","tecnico_mandante","tecnico_visitante","vencedor","arena",
1,"1","29/03/2003","16:00","Guarani","Vasco",,,,,,"Guarani","À Brinco de Ouro","4","2","SP","RJ"
2,"1","29/03/2003","16:00","Athletico-PR","Gremio",,,,,,"Athletico-PR","À Arena da Baixada","2","0","PR","RS"
3,"1","30/03/2003","16:00","Flamengo","Coritiba",,,,,,"À Maracanã","1","1","RJ","PR"
4,"1","30/03/2003","16:00","Goiás","Paysandu",,,,,,"À Serra Dourada","2","2","GO","PA"
5,"1","30/03/2003","16:00","Internacional","Ponte Preta",,,,,,"À Beira Rio","1","1","RS","SP"
6,"1","30/03/2003","16:00","Criciúma","Fluminense",,,,,,"Criciúma","À Heriberto Hulse","2","0","SC","RJ"
7,"1","30/03/2003","16:00","Juventude","Sao Paulo",,,,,,"À Alfredo Jaconi","2","2","RS","SP"
8,"1","30/03/2003","16:00","Fortaleza","Bahia",,,,,,"À Castelão","0","0","CE","BA"
9,"1","30/03/2003","16:00","Cruzeiro","Sao Caetano",,,,,,"À Mineirão","2","2","MG","SP"
10,"1","30/03/2003","16:00","Vitoria","Figueirense",,,,,,"À Barradão","1","1","BA","SC"
11,"1","30/03/2003","18:00","Santos","Parana",,,,,,"À Vila Belmiro","2","2","SP","PR"
12,"1","30/03/2003","18:00","Corinthians","Atletico-MG",,,,,,"Atletico-MG","À Pacaembu","0","3","SP","MG"
13,"2","05/04/2003","16:00","Fluminense","Fortaleza",,,,,,"À Maracanã","1","1","RJ","CE"
14,"2","05/04/2003","16:00","Atletico-MG","Santos",,,,,,"À Mineirão","0","0","MG","SP"
15,"2","05/04/2003","16:00","Coritiba","Internacional",,,,,,"Internacional","À Couto Pereira","0","1","PR","RS"
16,"2","05/04/2003","18:00","Gremio","Guarani",,,,,,"Gremio","À Olímpico","3","1","RS","SP"
17,"2","06/04/2003","16:00","Bahia","Flamengo",,,,,,"Flamengo","À Fonte Nova","1","2","BA","RJ"
18,"2","06/04/2003","16:00","Figueirense","Corinthians",,,,,,"À Orlando Scarpelli","3","3","SC","SP"
19,"2","06/04/2003","16:00","Paysandu","Vitoria",,,,,,"Vitoria","À Mangueirão","1","2","PA","BA"
20,"2","06/04/2003","16:00","Ponte Preta","Juventude",,,,,,"Ponte Preta","À Moisés Lucarelli","1","0","SP","RS"
21,"2","06/04/2003","16:00","Parana","Athletico-PR",,,,,,"Parana","À Pinheirão","3","0","PR","PR"
22,"2","06/04/2003","16:00","Sao Caetano","Criciúma",,,,,,"Sao Caetano","À Anacleto Campanella","3","2","SP","SC"
23,"2","06/04/2003","18:00","Sao Paulo","Cruzeiro",,,,,,"Cruzeiro","À Morumbi","2","4","SP","MG"
24,"2","06/04/2003","18:00","Vasco","Goiás",,,,,,"Vasco","À São Januário","6","4","RJ","GO"

Figura 3.1 – Parte inicial da base de dados bruta

Por isso, foi necessário utilizar a ferramenta “Texto para colunas” e separar todas as informações em colunas, usando aspas e vírgula como delimitador de separação. Após essa ação, cada jogo é representado por uma linha e contém informações sobre o nome do time mandante (que joga dentro de casa), o nome do time visitante (que joga fora de casa), a quantidade de gols marcados pelo time mandante, a quantidade de gols marcados pelo time visitante, a rodada do campeonato em que o jogo foi disputado e sua data, as quais estavam distribuídas nas colunas. Existiam outras colunas com outras informações que não foram utilizadas no estudo e, portanto, excluídas. A base de dados reestruturada é mostrada na figura 3.2 abaixo.

ID	rodata	data	mandante	visitante	mandante_Placar	visitante_Placar
1	1	29/03/2003	Guarani	Vasco	4	2
2	1	29/03/2003	Athletico-PR	Gremio	2	0
3	1	30/03/2003	Flamengo	Coritiba	1	1
4	1	30/03/2003	Goias	Paysandu	2	2
5	1	30/03/2003	Internacional	Ponte Preta	1	1
6	1	30/03/2003	Criciuma	Fluminense	2	0
7	1	30/03/2003	Juventude	Sao Paulo	2	2
8	1	30/03/2003	Fortaleza	Bahia	0	0
9	1	30/03/2003	Cruzeiro	Sao Caetano	2	2
10	1	30/03/2003	Vitoria	Figueirense	1	1
11	1	30/03/2003	Santos	Parana	2	2
12	1	30/03/2003	Corinthians	Athletico-MG	0	3
13	2	05/04/2003	Fluminense	Fortaleza	1	1
14	2	05/04/2003	Athletico-MG	Santos	0	0
15	2	05/04/2003	Coritiba	Internacional	0	1
16	2	05/04/2003	Gremio	Guarani	3	1
17	2	06/04/2003	Bahia *	Flamengo	1	2
18	2	06/04/2003	Figueirense	Corinthians	3	3
19	2	06/04/2003	Paysandu	Vitoria	1	2
20	2	06/04/2003	Ponte Preta	Juventude	1	0
21	2	06/04/2003	Parana	Athletico-PR	3	0
22	2	06/04/2003	Sao Caetano	Criciuma	3	2
23	2	06/04/2003	Sao Paulo	Cruzeiro	2	4
24	2	06/04/2003	Vasco	Goias	6	4
25	4	09/04/2003	Paysandu	Santos	2	1
37	3	12/04/2003	Internacional	Bahia	2	0
38	3	12/04/2003	Santos	Figueirense	2	0
39	3	12/04/2003	Athletico-PR	Athletico-MG	1	2
40	3	12/04/2003	Corinthians	Paysandu	6	1
41	3	13/04/2003	Vitoria	Vasco	2	1
42	3	13/04/2003	Criciuma	Sao Paulo	1	1
43	3	13/04/2003	Fortaleza	Sao Caetano	1	1
44	3	13/04/2003	Juventude	Coritiba	2	1
45	3	13/04/2003	Guarani	Parana	2	0

Figura 3.2 – Parte inicial da base de dados reestruturada

Também fez parte da base de dados pronta o valor do elenco de cada time no início do ano, de cada ano desde 2006 até 2025, retirada do site Transfermarkt, criado na Alemanha em 2000 e considerado uma referência global como banco de dados de futebol. É de conhecimento geral que jornalistas esportivos dos maiores canais de comunicação do mundo consultam o site.

Os Campeonatos Brasileiros dos anos 2003 e 2004 foram disputados com 24 equipes, totalizando 46 rodadas, e cada rodada contendo 12 jogos. Já o Campeonato Brasileiro de 2005 foi disputado com 22 equipes, totalizando 42 rodadas, cada uma contendo 11 jogos. Do Campeonato Brasileiro 2006 em diante, todos foram disputados com 20 equipes, totalizando 38 rodadas, com 10 jogos em cada uma.

3.5. Pré-processamento e Feature Engineering

A base de dados bruta foi reestruturada e a ordenação dos confrontos foi por critério cronológico, do mais antigo para o mais recente. A partir dessa reestruturação, foram criadas variáveis para o estudo se aproximar dos reais fatores que influenciam um jogo de futebol e a colocação de um time em um campeonato, e então formar a base de dados que foi utilizada na aplicação de Regressão Logística e do algoritmo de Machine Learning, XGBoost. Por exemplo, o número de participações de um time na série A do Campeonato Brasileiro, é uma característica de um time de futebol que deve ser levada em consideração na previsão.

Não tem disponível no site Transfermarkt todos os dados dos valores do elenco de cada time no início do ano. Nos anos de 2003, 2004 e 2005, a informação não está disponível para nenhuma equipe. Nos anos seguintes, a grande maioria dos clubes tem essa informação disponível. A partir do ano de 2006, os valores faltantes relacionados ao valor do elenco de cada equipe no início do ano foram preenchidos com o menor valor de elenco da equipe dentre os anos de 2006 e 2025. O preenchimento foi feito dessa forma pois, nesses casos, os valores não estavam preenchidos no site Transfermarkt em razão dos jogadores do elenco da equipe não terem valor de mercado conhecido, ou seja, valiam muito pouco. Os preenchimentos de valores faltantes com o menor valor disponível foram analisados e nenhum valor foi discrepante dos demais, não interferindo na análise preditiva.

3.6. Caracterização de variáveis

Na tabela 3.1 de Caracterização de Variáveis, a seguir, estão listadas todas as variáveis que foram levantadas e tiveram sua participação analisada para este estudo, suas respectivas descrições, unidades de medida e se foram utilizadas ou não.

Nome	Descrição	Unidade	Utilizada?
Gols casa marcados	Média de gols de uma equipe feitos em casa	Numérica	Sim
Gols casa sofridos	Média de gols de uma equipe sofridos em casa	Numérica	Sim
Gols fora marcados	Média de gols de uma equipe feitos fora de casa	Numérica	Sim

Gols fora sofridos	Média de gols de uma equipe sofridos fora de casa	Numérica	Sim
Anos série A	Quantidade de participações na série A do Campeonato Brasileiro	Numérica	Sim
Pontuação acumulada	Pontuação acumulada nos últimos três campeonatos, sem contar com o corrente	Numérica	Sim
Valor elenco	Média do valor do elenco medido no início do ano	Numérica	Sim
Finalizações feitas	Número de finalizações tentadas na meta do adversário	Numérica	Não
Finalizações feitas no gol	Número de finalizações que acertaram o gol do adversário	Numérica	Não
Finalizações sofridas	Número de finalizações sofridas	Numérica	Não
Finalizações sofridas no gol	Número de finalizações sofridas que acertaram o gol	Numérica	Não
Escanteios batidos	Número de escanteios a favor de uma equipe	Numérica	Não
Escanteios fornecidos	Número de escanteios contra uma equipe	Numérica	Não
Cartões amarelos	Número de cartões amarelos de uma equipe	Numérica	Não
Cartões vermelhos	Número de cartões vermelhos de uma equipe	Numérica	Não
Faltas cometidas	Número de faltas cometidas por uma equipe	Numérica	Não
Faltas sofridas	Número de faltas sofridas por uma equipe	Numérica	Não
Performance individual	Média das notas dos jogadores de uma equipe nos 5 jogos anteriores	Numérica	Não
Performance recente	Pontuação conquistada por uma equipe nos 5 jogos anteriores	Numérica	Não
Clássico regional	Se a partida representa um clássico regional. Acontece quando os adversários são da mesma região	Binária	Não
Posse de bola	Tempo em que uma equipe ficou com a posse da bola	Percentual	Não
Confrontos por posição	Pontuação de uma equipe no campeonato	Numérica	Não

Tabela 3.1 – Caracterização de variáveis

O método utilizando Distribuição de Poisson utilizou somente parte delas, que foram as médias de gols feitos nas situações de mandante e visitante as médias de gols sofridos nas situações de mandante e visitante, sendo atualizadas a cada jogo.

As outras variáveis que compõem a base de dados do estudo foram o número de participações de cada time na série A do Campeonato Brasileiro; a pontuação acumulada nos últimos 3 anos na série A do Campeonato Brasileiro; e o valor do elenco de cada time no início do ano. Essas variáveis foram construídas a partir da base de dados bruta do GitHub.

3.7. Métodos

3.7.1. Distribuição de Poisson

A Distribuição de Poisson foi aplicada somente nos dados de média de gols dos anos de 2020 até a décima oitava rodada do campeonato de 2026, sendo separados por gols feitos por mandantes e visitantes e gols sofridos por mandantes e visitantes.

Foi utilizado o Valor Esperado como lambda de Poisson. Ele foi calculado, para o mandante, como sendo a média entre as médias do número de gols (número de gols dividido pelo número de partidas) feitos pela equipe mandante e do número de gols sofridos pela equipe visitante. No caso do Valor Esperado da equipe visitante, foi calculado sendo a média entre as médias do número de gols sofridos pela equipe mandante e do número de gols feitos pela equipe visitante. Este cálculo foi realizado seguindo a proposta de FONSECA JÚNIOR (2025).

Dessa forma, um mesmo time terá um lambda diferente em cada confronto, pois é levado em consideração o desempenho do seu adversário dentro e fora de casa, tanto no ataque, com o número de gols feitos, quanto na defesa, com o número de gols sofridos, ambos entrando no cálculo em todos os 380 confrontos.

Os cálculos foram realizados utilizando a ferramenta Microsoft Excel, onde todos os times participantes do campeonato foram postos em uma matriz de confrontos, como mostrado abaixo. Foram construídas duas matrizes, uma para o Valor Esperado correspondente aos mandantes, e outra correspondente aos visitantes. Na matriz, quando um time se cruza com ele próprio, a célula correspondente ao Valor Esperado desse confronto fica vazia, como mostrado nas figuras 3.3 e 3.4.

Esperança mandante	Athletico-PR	Athletico-MG	Bahia	Botafogo-RJ
Athletico-PR		1,3092	1,5658	1,2171
Athletico-MG	1,4842		1,6355	1,2868
Bahia	1,4342	1,3289		1,2368
Botafogo-RJ	1,5000	1,3947	1,6513	
Chapecoense	1,0263	0,9211	1,1776	0,8289
Corinthians	1,4211	1,3158	1,5724	1,2237
Coritiba	1,2368	1,1316	1,3882	1,0395
Cruzeiro	1,3684	1,2632	1,5197	1,1711

Tabela 3.2 - Início da matriz de confrontos dos valores esperados dos mandantes

Esperança visitante	Athletico-PR	Athletico-MG	Bahia	Botafogo-RJ
Athletico-PR		1,0500	1,0197	1,1184
Athletico-MG	0,9329		0,9855	1,0842
Bahia	0,9276	1,0105		1,0789
Botafogo-RJ	0,9342	1,0171	0,9868	
Chapecoense	1,2961	1,3789	1,3487	1,4474
Corinthians	0,9171	1,0000	0,9697	1,0684
Coritiba	1,0329	1,1158	1,0855	1,1842
Cruzeiro	0,8662	0,9491	0,9189	1,0175

Tabela 3.3 - Início da matriz de confrontos dos valores esperados dos visitantes

Após o cálculo das Esperanças de todos os confrontos, tomando os times como mandantes e visitantes, foi calculado o valor da probabilidade de um time fazer certa quantidade de gols utilizando a Distribuição de Poisson. As possibilidades de número de gols marcados consideradas no estudo foram de 0 a 6, pois as probabilidades de um time fazer mais de 6 gols foram consideradas desprezíveis.

Foram montadas tabelas para todos os 380 confrontos, todas contendo os valores das probabilidades do mandante e do visitante marcarem certa quantidade de gols, como o exemplo abaixo. A quantidades de gols marcados com a maior probabilidade dentre as possíveis do mandante e do visitante são as assumidas no modelo.

	Mandante	Visitante
Gols	Athletico-PR	Athletico-MG
0	27,00%	34,99%
1	35,35%	36,74%
2	23,14%	19,29%
3	10,10%	6,75%
4	3,31%	1,77%
5	0,87%	0,37%
6	0,19%	0,07%

Tabela 3.4 - Exemplo de tabela das probabilidades de Poisson, confronto entre Athletico-PR e Atlético-MG

No exemplo do jogo da figura 3.5, o resultado da modelagem utilizando Distribuição de Poisson para o confronto entre Athletico-PR e Atlético-MG foi um empate de 1 x 1.

3.7.2. Regressão Logística

Com a base de dados reestruturada e realizado o pré-processamento, além do estudo e seleção de variáveis, foi realizada uma análise descritiva inicial com o objetivo de explorar o comportamento das variáveis, separados em 3 grupos de interesse, mandantes vitoriosos, visitante vitoriosos, empates.

Foi realizada uma regressão logística multinomial com o objetivo de prever, ponderando as variáveis escolhidas de acordo com histórico dos confrontos, os resultados dos jogos do Campeonato Brasileiro 2026. Como antes mencionado, as informações sobre data, rodada, placar do mandante e placar do visitante não foram incluídas no modelo de regressão logística, foram utilizadas somente para o pré-processamento dos dados e criação de variáveis no processo de feature engineering.

Foi escolhida a regressão logística multinomial para ser possível estimar a probabilidade de mais de duas classes, diferente da regressão logística binária, que só permite trabalhar com duas classes, visto que no futebol temos três possíveis desfechos. As três classes consideradas no modelo para a variável resposta foram: classe 0 = vitória do visitante; classe 1 = empate; classe 2 = vitória do mandante.

As variáveis de participações na série A do Campeonato Brasileiro, pontuação acumulada nos últimos três anos e valor do elenco precisaram ser desmembradas em mandantes e visitantes, para então todas as variáveis fiquem com o mesmo padrão na base de dados. Por exemplo, a variável valor do elenco foi desmembrada em valor do elenco do mandante e valor do elenco do visitante. O valor do elenco de um time específico se repete em todos os 38 jogos que esse time disputou em um campeonato, 19 vezes na coluna de mandante e 19 vezes na coluna de visitante. O mesmo acontece para as outras duas variáveis citadas neste parágrafo. As quatro variáveis que envolvem média de gols, feitos e sofridos como mandante e visitante, são atualizadas a cada rodada e independem do campeonato.

Antes do treinamento, todas as variáveis que fizeram parte do modelo foram padronizadas por meio da técnica StandardScaler, que transforma os dados de acordo com uma distribuição de média igual a 0 e desvio padrão igual a 1, para reduzir problemas de escalas entre as variáveis. Os dados de 2003 até 2024 foram utilizados como dados de treinamento, e os dados de 2025 foram utilizados como dados de teste.

Após o treinamento, ocorreu a etapa de calibração das probabilidades, utilizando o método Sigmoid (Platt Scaling) nos dados de teste. Essa calibração é feita para aproximar as probabilidades estimadas com as frequências dos dados históricos.

O desempenho do modelo foi avaliado por meio das métricas de acurácia, precisão, recall, F1-Score e Matriz de Confusão. A acurácia mede a proporção total de previsões corretas. As métricas de precisão, recall e F1-Score avaliam a capacidade do modelo de identificar cada uma das três classes e a matriz de confusão analisa os padrões de erro do modelo.

Para cada partida, foram obtidas as probabilidades de vitória do mandante, vitória do visitante ou empate. O resultado da previsão foi definido pela classe que apresentou maior probabilidade em cada confronto.

3.7.3. XGBoost

A metodologia aplicada no algoritmo de XGBoost é praticamente a mesma da aplicada na Regressão Logística multinomial. A etapa de definição da variável resposta, validação temporal, avaliação do desempenho e definição dos resultados são idênticas. O que difere os dois modelos é a parametrização e a forma como as probabilidades são calculadas, especificadas no capítulo de Referencial Teórico deste trabalho.

A parametrização utilizada no treinamento do modelo de XGBoost foi estimada por validação cruzada e está especificada na tabela 3.6 abaixo:

Parâmetro	Representação	Valor
Número de árvores	n_estimators	400
Profundidade máxima das árvores	max_depth	5
Taxa de aprendizado	learning_rate	0,05
Subamostragem das observações	subsample	0,8
Subamostragem das variáveis	colsample bytree	0,8
Função objetivo	-	classificação multiclasse (multi)

Tabela 3.5 – Hiperparâmetros do modelo de XGBoost

4. RESULTADOS

Neste capítulo são apresentados os resultados da análise descritiva e dos três métodos aplicados na base de dados.

4.1. Análise Descritiva

A análise descritiva das variáveis utilizadas nos modelos de Regressão Logística e XGBoost é apresentada abaixo na Tabela 4.2. Foram selecionados somente média e desvio padrão (entre parênteses) como métricas de análise. Assim, podemos já começar a identificar alguns padrões do que é necessário para vencer dentro e fora de casa, assim como empatar.

Variável	Em caso de Mandantes Vencedores	Em caso de Visitantes Vencedores	Em caso de Empates
Média de gols marcados pelos mandantes	1,69 (0,28)	1,58 (0,29)	1,63 (0,30)
Média de gols marcados pelos visitantes	1,05 (0,23)	1,14 (0,24)	1,09 (0,22)
Média de gols sofridos pelos mandantes	1,05 (0,18)	1,11 (0,22)	1,072(0,23)
Média de gols sofridos pelos visitantes	1,66 (0,32)	1,55 (0,24)	1,58 (0,27)
Pontuação acumulada (3 camp.) - Mandante	106,63 (78,03)	99,65 (75,88)	106,93 (77,13)
Pontuação acumulada (3 camp.) - Visitante	100,35 (77,49)	112,41 (76,95)	107,13 (76,71)
Valor do elenco - Mandante	€ 45,41 MM (€ 38,81 MM)	€ 42,13 MM (€ 36,15 MM)	€ 44,71 MM (€ 38,06 MM)
Valor do elenco - Visitante	41,93 (36,23)	48,55 (40,41)	45,32 (38,59)
Número de participações - Mandante	8,80 (6,24)	8,10 (5,91)	8,44 (6,12)
Número de participações - Visitante	8,03 (5,98)	9,41 (6,31)	8,70 (6,18)

Tabela 4.1 – Análise Descritiva (Variável de valor de elenco em milhões de euros)

Através da análise descritiva, é possível aferir que os mandantes apresentam maior média de gols marcados e menor média de gols sofridos, o que mostra um desempenho ofensivo e defensivo das equipes quando jogam em seus domínios.

Os visitantes vencedores apresentaram maior média de gols marcados em relação aos mandantes vencedores, o que indica uma necessidade maior de marcar gols para vencer fora de casa.

Uma observação interessante é que quando o mandante vence, ele possui pontuação acumulada nos últimos três campeonatos maior que o visitante, e quando o visitante vence, ele possui pontuação acumulada nos últimos três campeonatos maior que o mandante, além dessa pontuação também ser maior que a do mandante quando o mandante vence, indicando que essa variável tem um comportamento coerente com os resultados das partidas.

É possível observar o padrão de que, em média, quando o mandante vence, ele tem valor de elenco superior, e quando, em média, o visitante vence, ele tem valor de elenco superior. Portanto podemos assumir que essa variável funciona como um medidor de qualidade técnica do elenco.

4.2. Distribuição de Poisson

Os resultados da aplicação dos modelos foram avaliados através das métricas de acurácia, precisão, recall, F1-Score e Matriz de Confusão. O conjunto de avaliação foram as 18 primeiras rodadas do Campeonato Brasileiro 2026, totalizando 180 jogos. As métricas relacionadas à aplicação do modelo de Poisson estão expostas nas tabelas 4.2 e 4.3.

Classe	Precisão	Recall	F1-Score	Suporte
Mandante	0,4222	0,437	0,4294	87
Empate	0,2444	0,44	0,3143	50
Visitante	0	0	0	43

Tabela 4.2 - Métricas de avaliação do modelo de Poisson

A matriz de confusão apresentada na tabela 4.3 abaixo mostra quantos resultados de partidas foram corretamente previstos no teste. As classes nas linhas representam os resultados reais, enquanto as classes nas colunas representam os resultados previstos. Por exemplo, o cruzamento da classe vitória do mandante nas linhas com vitória do visitante nas colunas nos mostra o número de jogos que na realidade foram vitória do mandante, mas na previsão foram vitórias do visitante.

Real \ Previsto	Mandante	Empate	Visitante
Mandante	38	49	0
Empate	28	22	0
Visitante	24	19	0

Tabela 4.3 – Matriz de confusão do modelo de Poisson

O modelo não foi capaz de prever nenhuma vitória de visitante. Todas as partidas reais que o resultado foi uma vitória do visitante foram classificados de forma errada, como vitória do mandante ou empate. De 87 partidas que terminaram com uma vitória do mandante, o modelo acertou 38, enquanto 49 foram classificados incorretamente como empate.

Após o cálculo das probabilidades do número de gols de cada equipe em cada um dos 380 confrontos e a consequente determinação dos resultados pela previsão, somando a pontuação de todos os times chegamos à classificação final prevista do Campeonato Brasileiro 2026, apresentada na tabela 4.4 abaixo.

Tabela - Campeonato Brasileiro 2026				
COLOCAÇÃO	TIMES	MANDANTE	VISITANTE	TOTAL
1º	Flamengo	52	20	72
2º	Palmeiras	50	20	70
3º	Mirassol	46	15	61
3º	Botafogo-RJ	42	19	61
5º	Cruzeiro	48	8	56
6º	Atlético-MG	42	13	55
7º	Fluminense	48	5	53
8º	São Paulo	46	5	51
9º	Grêmio	34	14	48
9º	Bahia	40	8	48
11º	Corinthians	42	5	47
12º	Bragantino	34	12	46
13º	Internacional	40	5	45
14º	Vitória	38	2	40
15º	Athletico-PR	36	3	39
16º	Santos	32	2	34
17º	Coritiba	22	6	28
18º	Vasco	18	5	23
18º	Chapecoense	8	1	9

Tabela 4.4 – Tabela de classificação prevista pelo método de Distribuição de Poisson do Campeonato Brasileiro 2026

Como não modelamos os critérios de desempate, times com a mesma pontuação foram colocados ocupando a mesma posição. Portanto, Flamengo seria o campeão, se classificando para a Libertadores junto de Palmeiras, Mirassol e Botafogo. Santos, Coritiba, Vasco e Chapecoense seriam rebaixados. Times com mesma pontuação foram considerados ocupando a mesma colocação, pois não foram adotados critérios de desempate, como saldo de gols ou número de vitórias.

4.3. Regressão Logística Multinomial

Os resultados da aplicação da Regressão Logística Multinomial estão expostos nas tabelas 4.5 e 4.6.

Classe	Precisão	Recall	F1-Score	Suporte
Visitante	0,31	0,63	0,42	35
Empate	0,31	0,12	0,17	43
Mandante	0,58	0,52	0,55	75

Tabela 4.5 – Métricas de avaliação do modelo de Regressão Logística

A acurácia global do modelo foi de 43%, mostrando que menos da metade das partidas tiveram seus resultados corretamente previstos.

A matriz de confusão referente à aplicação do método de Regressão Logística é apresentada na tabela 4.6 abaixo.

Real \ Previsto	Visitante	Empate	Mandante
Visitante	22	3	10
Empate	20	5	18
Mandante	28	8	39

Tabela 4.6 – Matriz de confusão do modelo de Regressão Logística

O modelo apresentou um desempenho parecido na previsão de vitórias do mandante e do visitante, acertando 39 resultados de 75 e 22 de 35, respectivamente. Porém, o desempenho na classe de empate foi bastante limitado, acertando apenas 5 resultados que terminaram realmente como empate em um total de 43.

A tabela de classificação final prevista para o Campeonato Brasileiro 2026 pelo modelo de Regressão Logística é mostrada abaixo.

Tabela - Campeonato Brasileiro 2026		
COLOCAÇÃO	TIMES	PONTOS
1º	Palmeiras	91
2º	Flamengo	86
3º	Sao Paulo	74
4º	Cruzeiro	72
5º	Fluminense	66
6º	Santos	58
6º	Gremio	58
6º	Athletico-PR	58
9º	Bragantino	53
10º	Athletico-MG	52
11º	Internacional	48
12º	Corinthians	45
13º	Mirassol	44
14º	Bahia	41
15º	Coritiba	32
16º	Vasco	29
17º	Vitoria	26
18º	Botafogo-RJ	24
19º	Remo	21
20º	Chapecoense	11

Tabela 4.7 – Tabela de classificação do Campeonato Brasileiro 2026 prevista pelo método de Regressão Logística

Segundo a tabela 4.7, Palmeiras seria o campeão, se classificando para a Libertadores junto de Flamengo, São Paulo e Cruzeiro. Vitória, Botafogo, Remo e Chapecoense seriam rebaixados.

4.4. XGBoost

Os resultados da aplicação do algoritmo XGBoost de Machine learning foram avaliados da mesma maneira do modelo de Regressão Logística. A acurácia global do modelo foi de 56%. Nas tabelas 4.8 e 4.9 abaixo são apresentadas as outras métricas de avaliação.

Classe	Precisão	Recall	F1-Score	Suporte
Visitante	0,46	0,34	0,39	35
Empate	0,45	0,12	0,19	43

Mandante	0,59	0,91	0,71	75
-----------------	------	------	------	----

Tabela 4.8 - Métricas de avaliação do modelo de XGBoost

A matriz de confusão do modelo de XGBoost apresentada abaixo identificou quase todas as vitórias dos mandantes, 68 de 75 ocorrências. Porém, acertou apenas 5 empates de um total de 43, continuando com uma baixa taxa de assertividade nesta classe.

Real \ Previsto	Visitante	Empate	Mandante
Visitante	12	4	19
Empate	9	5	29
Mandante	5	2	68

Tabela 4.9 – Matriz de confusão do modelo de XGBoost

A tabela de classificação prevista para o Campeonato Brasileiro 2026 pelo modelo de XGBoost é mostrada abaixo.

Tabela - Campeonato Brasileiro 2026		
COLOCAÇÃO	TIMES	PONTOS
1º	Palmeiras	89
2º	Flamengo	88
3º	Bragantino	63
4º	Atletico-MG	60
4º	Sao Paulo	60
6º	Cruzeiro	59
6º	Fluminense	59
8º	Athletico-PR	58
9º	Vitoria	54
10º	Mirassol	51
11º	Gremio	50
12º	Santos	46
13º	Internacional	45
14º	Coritiba	44
15º	Bahia	42
16º	Vasco	41
17º	Corinthians	31
18º	Botafogo-RJ	20
19º	Remo	19
20º	Chapecoense	14

Tabela 4.10 – Tabela de classificação do Campeonato Brasileiro 2026 prevista pelo método XGBoost

Segundo a tabela 4.10, Palmeiras seria o campeão, se classificando para a Libertadores junto de Flamengo, Bragantino e Atlético-MG. Corinthians, Botafogo, Remo e Chapecoense seriam rebaixados.

5. DISCUSSÃO

Analisando os resultados obtidos, podemos identificar que o modelo de XGBoost apresentou melhores resultados, apresentando uma melhora na acurácia em aproximadamente 12 pontos percentuais em relação a Regressão Logística e de 22 pontos percentuais em relação a Poisson. Esses resultados eram esperados, já que o modelo de XGBoost é considerado mais robusto e mais eficiente para problemas de classificação.

Uma hipótese que pode ser levantada é que os resultados no contexto do futebol não são lineares desfavorecendo o uso de Regressão Logística. Porém, a base de dados é relativamente curta, e o XGBoost tende a apresentar melhor desempenho quando é aplicado em uma grande quantidade de dados.

Apesar de 56% de acurácia parecer uma probabilidade baixa, dentro do contexto de futebol é uma boa porcentagem pois é um esporte com alto grau de imprevisibilidade decorrente de fatores aleatórios como expulsões, lesões, decisões arbitrais, entre diversos outros.

A dificuldade na previsão de empates também chamou atenção. Ela pode ser explicada no modelo de Regressão também por ser um modelo linear e o empate representar uma classe intermediária, o que dificulta a separação por meio de fronteiras lineares. Porém a dificuldade permaneceu no modelo de XGBoost, sugerindo que o problema não está apenas no algoritmo, mas também nas variáveis disponíveis. As variáveis utilizadas representam principalmente força relativa das equipes, essas características tendem a explicar quem é mais forte, mas não necessariamente quais jogos terminarão empatados.

É possível identificar que os modelos tenderam a favorecer a classe “Vitória do mandante”, o que era esperado, considerando que, historicamente, no futebol, os mandantes costumam sair vitoriosos. O pior desempenho nos modelos foi a previsão de empates, com uma porcentagem de acerto muito baixa, como mostrado nas Matrizes de Confusão. Esses resultados mostram que os modelos conseguiram identificar com mais facilidade times superiores e inferiores, porém não conseguiram identificar quando os confrontos eram mais equilibrados.

6. CONCLUSÃO

Este trabalho teve como objetivo aplicar e comparar métodos de análise preditiva para prever resultados de jogos de campeonatos de futebol, utilizando dados do Campeonato Brasileiro de 2003 até a décima oitava rodada do campeonato de 2026. Os métodos utilizados foram Distribuição de Poisson, Regressão Logística Multinomial e o algoritmo XGBoost de Machine Learning.

Foi realizada a obtenção, tratamento e estruturação dos dados, além de criação de variáveis. A análise descritiva realizada mostrou que as variáveis escolhidas têm influência nos resultados das partidas, justificando o seu uso no estudo.

Os resultados mostraram que o modelo de XGBoost teve um desempenho superior a Regressão, ele já é considerado na literatura por ter alta capacidade de previsão em problemas de classificação, além do fato do futebol não ter comportamentos lineares e a Regressão ser um modelo linear.

Pode-se deduzir que as relações entre as variáveis escolhidas e os resultados são predominantemente lineares. A análise das matrizes de confusão evidenciou um grande acerto quando a previsão era de um jogo onde o mandante saiu vitorioso e mostrou uma enorme dificuldade na previsão de empates.

De forma geral, o trabalho mostrou que técnicas estatísticas e de aprendizado de máquina podem ser utilizados na previsão de resultados de jogos de futebol, embora existam limitações provenientes da aleatoriedade do esporte e especificamente deste trabalho, de tempo hábil para o aperfeiçoamento dos modelos e no tratamento das variáveis.

Por fim, é sugerido que o trabalho tenha uma continuidade em futuras pesquisas, que podem incluir novas variáveis, como indicadores de forma recente das equipes, pontuação no campeonato corrente, mudanças de elenco e treinador, métricas individuais dos atletas, histórico direto entre equipes, entre outras.

Além da aplicação acadêmica, o trabalho mostra que a Engenharia e a Ciência de Dados podem ser úteis dentro do futebol em apoio à tomada de decisão, estudos de desempenho de equipes e avaliação de probabilidades em competições.

7. REFERÊNCIAS BIBLIOGRÁFICAS

BHATT, C. M.; PATEL, P.; GHETIA, T.; MAZZEO, P. L. Effective heart disease prediction using machine learning techniques. *Algorithms*, v. 16, n. 2, p. 88, 2023. DOI: 10.3390/a16020088.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, 2016. p. 785–794.

CNN BRASIL. Da tática ao VAR: como a IA promete inovar na Copa do Mundo de 2026. Disponível em: <https://www.cnnbrasil.com.br/esportes/futebol/copa-do-mundo/da-tatica-ao-var-como-a-ia-promete-inovar-na-copa-do-mundo-de-2026/>. Acesso em: 12 jun. 2026.

CNN BRASIL. Setor de apostas online cresceu 734% desde 2021, aponta pesquisa. Disponível em: https://www.cnnbrasil.com.br/economia/negocios/setor-de-apostas-online-cresceu-734-desde-2021-aponta-pesquisa/#goog_rewarded. Acesso em: 12 jun. 2026.

CONFEDERAÇÃO BRASILEIRA DE FUTEBOL. Manual de Competições 2026. Disponível em: <https://www.cbf.com.br/a-cbf/noticias/informes-cbf/a/cbf-divulga-manual-de-competicoes-documento-que-substituirá-o-regulamento-geral-das-competicoes>. Acesso em: 12 jun. 2026.

EKANAYAKE, I. U.; MEDDAGE, D. P. P.; RATHNAYAKE, U. A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). *Case Studies in Construction Materials*, v. 16, e01059, 2022. DOI: 10.1016/j.cscm.2022.e01059.

FONSECA JUNIOR, J. M. da. Modelagem de resultados de futebol: aplicação da distribuição de Poisson. 2025. Dissertação (Mestrado Profissional em Matemática em Rede Nacional – PROFMAT) – Universidade Federal de São Paulo, São José dos Campos, 2025.

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

GLOBO ESPORTE. Flamengo sobe uma posição na lista de clubes mais ricos do mundo; veja ranking. Disponível em: <https://ge.globo.com/futebol/futebol-internacional/noticia/2026/01/22/flamengo-sobe-uma-posicao-na-lista-de-clubes-mais-ricos-do-mundo-veja-ranking.ghml>. Acesso em: 12 jun. 2026.

GLOBO ESPORTE. Premiações 2025: Flamengo ultrapassa o Fluminense e encerra ano na liderança. Disponível em: <https://ge.globo.com/gato-mestre/noticia/2025/12/22/premiacoes-2025-flamengo-ultrapassa-o-fluminense-e-encerra-ano-na-lideranca.ghml>. Acesso em: 12 jun. 2026.

GLOBO ESPORTE. Todos os clubes do Brasileirão 2025 são patrocinados por bets. Disponível em: <https://ge.globo.com/futebol/futebol-internacional/noticia/2025/03/11/todos-os-clubes-do-brasileirao-2025-sao-patrocinados-por-bets.ghml>. Acesso em: 12 jun. 2026.

GRAND VIEW RESEARCH. Sports Event Market Size, Share & Trends Analysis Report By Event Scale, By Event Format, By Target Audience, By Revenue Stream, By Type of Event, By Region, And Segment Forecasts, 2026–2033. San Francisco: Grand View Research, 2026.

GONZALEZ, L. de A. Regressão logística e suas aplicações. 2018. Monografia (Graduação em Ciência da Computação) – Universidade Federal do Maranhão, São Luís, 2018.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The elements of statistical learning. 2. ed. New York: Springer, 2009.

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. Applied logistic regression. 3. ed. Hoboken: John Wiley & Sons, 2013.

LOHMANN, M. Regressão logística e redes neurais aplicadas à previsão probabilística de alagamentos no município de Curitiba, PR. 2011. Tese (Doutorado em Geografia) – Universidade Federal do Paraná, Curitiba, 2011.

NIELSEN. Relatório Global de Esportes 2025: três fatores-chave de crescimento a serem observados nos esportes globais. Nova York: Nielsen, 2025. Disponível em: <https://www.nielsen.com/pt/insights/2025/global-sports-report-2025/>. Acesso em: 12 jun. 2026.

NVIDIA CORPORATION. XGBoost. NVIDIA Glossary, [s.d.]. Disponível em: <https://www.nvidia.com/en-us/glossary/xgboost/>. Acesso em: 12 jun. 2026.

PERES, I. T.; HAMACHER, S.; OLIVEIRA, F. L. C.; BOZZA, F. A.; SALLUH, J. I. F. Data-driven methodology to predict the ICU length of stay: a multicentre study of 99,492 admissions in 109 Brazilian units. *Anaesthesia Critical Care & Pain Medicine*, 2022.

SERASA. Gastos com o futebol afeta o bolso de 61% dos brasileiros, revela Serasa. Disponível em: <https://www.serasa.com.br/imprensa/gastos-com-o-futebol-afeta-o-bolso-dos-brasileiros-revela-serasa/>