UNIVERSIDADE FEDERAL DE ALFENAS - UNIFAL-MG

NIKOLAS NEVES DE FIGUEIREDO

COMPARAÇÃO DE MODELOS DE PREDIÇÃO: O CAMPEONATO BRASILEIRO DE FUTEBOL ENTRE 2017 A 2022

ALFENAS, MG
2023

UNIVERSIDADE FEDERAL DE ALFENAS - UNIFAL-MG

NIKOLAS NEVES DE FIGUEIREDO

COMPARAÇÃO DE MODELOS DE PREDIÇÃO: O CAMPEONATO BRASILEIRO DE FUTEBOL ENTRE 2017 A 2022

Trabalho de Conclusão de Curso para obtenção do título de licenciado em Matemática pela Universidade Federal de Alfenas. Área de concentração: Estatística Aplicada.
Orientador: Prof. Dr. Eric Batista Ferreira.
Coorientador: Prof. Me. Giovani Festa Paludo.

ALFENAS, MG
2023

Dedico este trabalho à minha mãe, que sempre me incentivou e ensinou a persistir naquilo que acredito.

Agradecimentos

Ao meus pais, Acioni e Dirceu, que me proporcionaram todo suporte financeiro, mental, emocional e afetivo que me fez ser capaz de chegar aqui. Toda dedicação, expectativa e incentivo foram essenciais para a realização deste trabalho.

À minha namorada Clara, por todo amor, carinho e compreensão. Todas as nossas conversas, as inúmeras vezes que recebi seu colo e o sonhos que compartilhamos estão profundamente gravadas em quem eu sou hoje. Ainda, não poderia deixar de mencionar sua família, Daniel, Cleide, Gustavo, Henrique e Laura. O acolhimento que vocês me proporcionaram foi o que manteve minha cabeça no lugar durante esse tempo.

Aos meus queridos amigos Rafael, Ronaldo e Edmara, ambos foram suporte e incentivo durante essa caminhada. Fico muito feliz por ter encontrado pessoas tão especiais que pude compartilhar momentos alegres, preocupações e experiências.

Ao meu coorientador e amigo Giovani, que junto a mim, se dedicou muitas horas buscando erros ou decifrando os códigos dos modelos. Ao meu orientador Eric, pela paciência, dedicação, amizade e ensinamentos. Ambos me permitiram amadurecer muito durante essa pesquisa.

Meus profundos agradecimentos ao professor Gilcione que disponibilizou os dados de um dos modelos utilizados nesse trabalho.

Por fim, aos meus queridos amigos, André, Catarina, Glória, João Paulo, José, Laura, Malu, Matheus e Vinícius que estiveram comigo sempre que precisei.

Resumo

Nas últimas décadas a indústria do futebol se fortaleceu muito graças ao desenvolvimento das mídias de comunicação. Mais recentemente, o mercado de apostas também passou por um expansão no meio digital, que revolucionou esse segmento, atraindo milhões de pessoas e movimentando quantias bilionárias. Nesse sentido, surgiram diversos modelos estatísticos com o interesse de predizer resultados de partidas de futebol. Este trabalho surge com o objetivo de buscar entre os modelos existentes, aquele que tenha a melhor capacidade de predizer os resultados de partidas de futebol. Para isso, foram escolhidos o modelo *SD 0* e *Chance I* propostos por Arruda e o modelo do departamento de matemática da Universidade Federal de Minas Geais. Para a aplicação, foram tomadas as partidas dos Campeonatos Brasileiros de Futebol Série A, dos anos 2017 à 2022. Os modelos utilizam estimação pelo método de momentos, método da máxima verossimilhança e um algoritmo de pesos, respectivamente. As 4 métricas escolhidas para comparar os modelos foram: taxa de acerto, medida de DeFinetti, erro preditivo ponderado e medida de Definetti detalhada, sendo as duas últimas propostas deste trabalho. Os três modelos se alternaram entre as melhores e piores métricas a depender do ano, porém *SD 0* se destacou em relação aos demais, apesar de cada um apresentar propriedades interessantes em cada aplicação.

Palavras-chave: Modelos Estatísticos, Predição Esportiva, Comparação de Previsões.

Abstract

In recent decades, the football industry has greatly strengthened, thanks to the development of communication media. More recently, the betting market has also experienced a digital expansion that revolutionized this sector, attracting millions of people and involving billions of financial transactions. In this context, various statistical models have emerged with the aim of predicting football match outcomes. This study aims to identify among existing models the one with the best ability to predict football match results. For this purpose, the models *SD 0* and *Chance I*, proposed by Arruda, and the model from the mathematics department of the Federal University of Minas Gerais were chosen. For the application, matches from the Brazilian Football Championships Serie A, from 2017 to 2022, were considered. The models use estimation through the method of moments, the maximum likelihood method, and a weight algorithm, respectively. The four metrics chosen to compare the models were: accuracy rate, DeFinetti measure, weighted predictive error, and detailed DeFinetti measure, the last two being proposed in this study. The three models alternated between the best and worst metrics depending on the year; however, *SD 0* stood out in comparison to the others, despite each presenting interesting properties in various applications.

Keywords: Statistical Models, Sports Prediction, Forecast Comparison .

Lista de Figuras

Lista de Tabelas

Sumário

1 INTRODUÇÃO

Nas últimas décadas, o esporte se consolidou com uma industria muito influente economicamente e socialmente. Com a ascensão da tecnologia e o desenvolvimento do mercado de entretenimento, os eventos esportivos conquistaram um papel fundamental no engajamento de uma ampla variedade de meios de comunicação. Esse fenômeno não apenas alimentou a paixão dos fãs, mas também gerou oportunidades econômicas significativas, desde merchandising e publicidade associada a eventos esportivos de grande magnitude.

Nesse contexto, a indústria de apostas esportivas passou por uma transformação digital que possibilitou uma expansão substancial e rápida. Atualmente, esse mercado movimenta milhões de reais mensalmente, tornando-se um ator poderoso que, em alguns casos, pode exercer influência negativa sobre o próprio esporte. A integração de plataformas online e aplicativos móveis tornou as apostas acessíveis a um público ainda maior, com um impacto considerável na paisagem esportiva e social.

Com base na situação descrita acima, justifica-se o desejo de predizer os resultados de competições esportivas. Essas predições não são apenas de interesse para apostadores, mas também para clubes esportivos, treinadores, patrocinadores e até mesmo torcedores que desejam entender melhor as chances de suas equipes favoritas em um determinado jogo.

O futebol se destaca como uma modalidade esportiva favorável para a previsão de resultados, sendo considerado o esporte mais popular do mundo, atraindo bilhões de entusiastas e movimentando milhões em cada campeonato. Dada a diversidade de variáveis envolvidas, que vão desde a qualidade das equipes até as condições climáticas, o futebol oferece um terreno fértil para análises estatísticas. A possibilidade de prever resultados torna-se evidente ao analisar e quantificar essas variáveis minunciosamente. Embora essa abordagem não seja nova, diversos autores têm dedicado seus esforços à proposição de modelos para predizer os resultados das partidas de futebol.

Esses modelos apresentam abordagens que variam desde métodos estatísticas tradicionais, até algoritmos de aprendizado de máquina. Alguns se concentram exclusivamente em estatísticas de jogos passados, enquanto outros incorporam fatores mais complexos, como as condições física dos jogadores, as táticas da equipe, a motivação e até mesmo o impacto da torcida. A partir disso pergunta-se qual desses modelos seria o melhor.

Este estudo tem como problema de pesquisa: Buscar entre os modelos existentes na literatura, aquele que demonstra a capacidade de predição mais acurada e precisa ao predizer resultados de partidas de futebol. Para abordar essa questão, foram selecionados os últimos 6 anos do Campeonato Brasileiro de Futebol Série A como base de aplicação para três modelos distintos mencionados na literatura. A avaliação da acurácia desses modelos é realizada por meio de quatro métricas, estabelecendo condições iguais para todas as predições.

2 OBJETIVOS

O objetivo norteador desse trabalho é determinar entre os modelos escolhidos, aquele que apresenta melhores resultados das métricas aplicadas em relação as predições do Campeonato Brasileiro de Futebol Série A dos anos de 2017 a 2022.

Junto à isso, existem objetivos específicos que se deseja alcançar nesse processo, estes estão listados abaixo:

- Analisar descritivamente o banco de dados.
- Implementar os modelos *SD 0* e *Chance I*.
- Obter as predições do Modelo *UFMG*.
- Propor métricas que avaliem a precisão dos modelos escolhidos.
- Calcular métricas de predição para todos os modelos.
- Comparar dos modelos em relação a cada uma das métricas.

3 REFERENCIAL TEÓRICO

Essa seção busca trazer a fundamentação teórica que o presente trabalho se baseia, detalhando os conceitos que posteriormente serão citados. Ainda, serão destacados trabalhos presentes na literatura que se assemelham em partes ao estudo que foi feito nessa monografia. Os temas abordados contemplam tanto assuntos específicos, que se aproximam do problema de pesquisa, quanto os mais gerais que estruturam a própria estatística.

3.1 AUDIÊNCIA ESPORTIVA E MERCADO DE APOSTAS

Ao que tudo indica, as primeiras demonstrações esportivas surgiram na Grécia antiga, com o exibicionismo de heróis e guerreiros com o intuito de adorar, entreter e agradecer os deuses. Além de uma função religiosa o esporte também adquiriu uma função política, tanto como forma de exibir poder militar de imperadores, quanto de manipular e influenciar massas. Ao se observar criticamente, ainda é possível observar cenários semelhantes nos dias atuais.

Ainda, segundo Barbanti (2006), o esporte se caracteriza como uma atividade competitiva que exige o uso de habilidades e/ou esforço físico. Isso delimita o esporte como uma atividade humana e cultural, que nesse contexto destaca-se pela competitividade que aflora em todos os envolvidos, desde os atletas, torcedores e até as pessoas mais desinteressadas. Esse sentimento também auxilia no fomento do esporte como atividade de entretenimento, desenvolvendo-se fortemente dentro dessa indústria. Pode-se observar isso na Figura 1, que apresenta os resultados de um estudo feito pela empresa Horizm em 2021, que analisou a audiência digital global dentro do Facebook, Instragram, Twitter e Youtube (*Big 4*) em relação aos esportes e instituições. Nesse estudo, foram destacadas as 11 ligas esportivas que ostentam uma audiência superior a 100 milhões de pessoas.

Tendo em mente esses números, evidencia-se também a enorme influência econômica desse segmento. Estes valores circulam, tanto para o pagamento de atletas, funcionários e infraestrutura, quanto por parte de patrocínio e apostas esportivas que recentemente passou por uma expansão digital.

Quanto ao mercado de apostas, em 2022 na Europa, sua receita atingiu 108,5 bilhões de libras que representa um crescente aumento em relação a 2021 e 2019 (H2 Gambling Capital, 2022). Também em 2022, constatou-se que o Brasil gerou o maior volume de visitas nos sites de apostas conhecidos mundialmente, alcançando 3,19 bilhões de visitas e em 2023 esse número foi de 3,78 bilhões, segundo Almeida (2023).

Figura 1: Ligas com as maiores audiência digitais no mundo, de diferentes esportes, expressas em milhões (m) de espectadores.

Fonte: Adaptado de Horizm (2023).

Em particular, o futebol se destaca em cada um dos segmentos citados acima. Por exemplo, sua audiência digital global atingiu aproximadamente 1,9 bilhão de pessoas em 2022 (ANDRADA, 2022). A receita dos cinco principais clubes brasileiros somou R$ 3,9 bilhões neste mesmo ano, segundo estudo da SportsValue (SOMOGGI, 2023). E a expressiva relevância do Campeonato Brasileiro Série A, como visto na Figura 1.

Haja vista tudo já citado, é possível imaginar o quanto seria impactante se uma pessoa pudesse prever o resultado final de uma competição esportiva. Patrocinadores, torcedores, atletas, apostadores, enfim, todos os atores que participam de alguma forma desse recorte, podem encontrar benefícios e interesse nessa habilidade. Atualmente, existem modelos matemáticos, estatísticos e computacionais que buscam fazer isso. Exemplos incluem Arruda (2000) e Lima *et al.* (2012), que são os modelos discutidos neste trabalho; Araújo *et al.* (2015), Tavares e Suzuki (2015), Degam (2019), Suzuki *et al.* (2010), Junior e Gamerman (2004) e Matos (2017), que apresentam modelos estatísticos ou computacionais para prever os resultados de partidas de futebol. Esses modelos serão discutidos com mais detalhes na seção 5.2.

3.2 PREDIÇÃO ESTATÍSTICA

Segundo Hyndman e Athanasopoulos (2018), a predição estatística trata-se de utilizar todas as informações possíveis, incluindo estatísticas de uma base de dados histórica e

conhecimento de fatores que podem impactar na predição, para projetar eventos futuros da maneira mais acurada possível. Ao relacionar essa definição aos conceitos da inferência estatística, pode-se dizer que estimar parâmetros e ajustar modelos com o intuito de descrever os eventos de uma população não conhecida a partir de uma amostra também se trata de um caso de predição. Este foi o caminho tomado pelos autores para predizer o resultado de determinada partida.

Em particular, no presente trabalho, dois dos três modelos selecionados utilizam modelos de regressão linear e de regressão linear generalizada. Para mais detalhes sobre esses métodos, vide Draper e Smith (1998).

Ao realizar uma predição sobre um evento, deseja-se estabelecer medidas ou métricas que meçam a precisão e acurácia dos resultados. Uma vez que a precisão trata-se de uma medida de proximidade dos valores verdadeiros em relação aos valores previstos. Já a acurácia pauta o grau de exatidão que o modelo atinge, comparando o valor verdadeiro ou aceito como verdadeiro ao valor previsto.

Em suma, o esquema de uma predição estatística passa por três etapas. Primeiro, a escolha e entendimento daquilo que se deseja predizer. Segundo, o desenvolvimento de um algoritmo e lógica para a estimação de parâmetros ou ajuste da regressão. Terceiro, avaliação da predição com métricas que permitam quantificar a acurácia e precisão da predição.

3.3 A PREDIÇÃO NO FUTEBOL

Como já introduzido, conseguir prever o resultado de um confronto esportivo ou a colocação de uma equipe em determinada competição pode ser muito vantajoso. Assim, muitos modelos se concentram em prever resultados esportivos, seja os placares de partidas de futebol, vôlei, basquete, colocações em corridas, resultados de lutas, entre outros, apresentando abordagens estatísticas, computacionais e matemáticas, e uma imensa variedade de metodologias.

Valendo-se dos modelos que visam prever resultados de partidas de futebol, é possível relacionar semelhanças e diferenças entre vários trabalhos. Como Arruda (2000), que propõe 6 modelos, dois deles sendo utilizados neste trabalho. Esses modelos utilizam técnicas que vão desde o método de momentos, método da máxima verossimilhança, com e sem o cálculo de covariantes, até métodos menos hierarquizados que simplesmente enunciam parâmetros da distribuição. Em particular, ele determina que o número de gols do mandante e visitante segue uma distribuição de Poisson bivariada da classe de Holgate.

Outros modelos na literatura concordam com a escolha da distribuição de Poisson para modelar o número de gols de um time. De acordo com Junior e Gamerman (2004), a distribuição de Poisson se ajusta bem ao número de gols dos times no campeonato de 2002. Ainda, trabalhos como Araújo *et al.* (2015), Degam (2019) e Tavares e Suzuki

(2015) aplicam justamente um dos modelos propostos por Arruda (2000) em seus estudos.

3.4 DISTRIBUIÇÕES DE PROBABILIDADE

Segundo Magalhães (2006), uma distribuição de probabilidade descreve como os possíveis resultados de um experimento aleatório ou evento são distribuídos em termos de suas probabilidades. Essa distribuição, em geral, está atrelada a uma função ou a uma tabela que relaciona evento e probabilidade. Vale destacar que essa probabilidade não deve assumir valores negativos, e o somatório das probabilidades de todos os eventos deve ser igual a 1.

As distribuições Normal, Binomial, Poisson, Bernoulli, Uniforme, Geométrica, Hipergeométrica, Gama, F, t e Beta estão entre as mais conhecidas e utilizadas para modelar a probabilidade de eventos e experimentos. No entanto, existe uma infinidade de outras distribuições que podem ser aplicadas em contextos gerais e específicos. A distribuição destacada neste trabalho foi a distribuição de Poisson, que foi usada para modelar a probabilidade de gols feitos por cada time em dois modelos.

3.4.1 Distribuição de Poisson Univariada

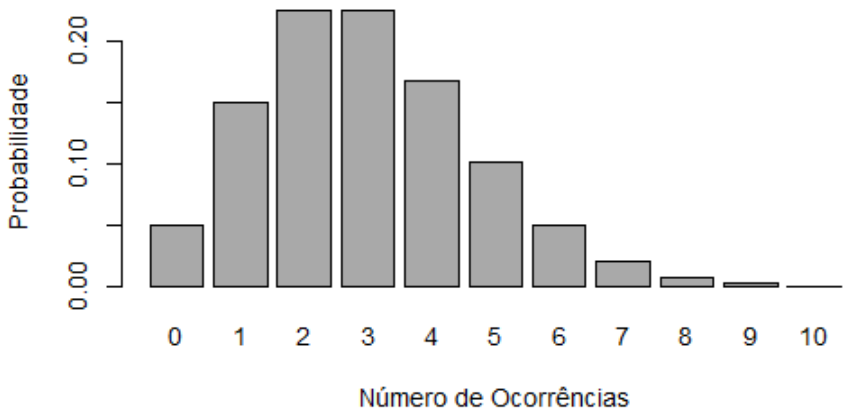
Essa distribuição, que originalmente foi introduzida em 1837 por Siméon Denis Poisson, é utilizada para modelar a probabilidade de eventos ditos raros, uma vez que o número de ocorrências não é muito alto (ROSS, 1976). Além disso, esse número é discreto, ou seja, essa variável aleatória assume apenas valores inteiros.

Essa distribuição é dita univariada já que observa-se o comportamento de apenas um variável, como representado abaixo, a variável X. A função de probabilidade dessa distribuição é dada por:

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \tag{1}$$

O único parâmetro dessa distribuição é o λ, ele é interpretado como a taxa de ocorrência de um determinado evento em relação a uma unidade amostral. Em particular, esse valor também representa a média e a variância dessa distribuição. Observe a seguir a representação gráfica de uma distribuição de Poisson gerada a partir da expressão 1 e com $\lambda = 3$

Figura 2: Exemplo gráfico de uma distribuição Poisson Univariada com $\lambda = 3$.



Fonte: Dos Autores.

Note que a distribuição ilustrada pela Figura 2, apresenta uma assimetria a direita que se localiza próxima ao valor da média dessa distribuição. Para intervalos maiores é possível constatar sua distribuição se assemelha a distribuição normal fixada em torno do meu valor de λ.

3.4.2 Distribuição de Poisson Bivariada

O caso bivariado dessa distribuição, é análogo ao caso mais simples com apenas um variável. Entretanto, observa-se agora o comportamento de X e Y, e cada uma dessas variáveis aleatórias tem seu próprio parâmetro λ_X e λ_Y da distribuição que elas foram amostradas. Assim sua função de probabilidade é dada por:

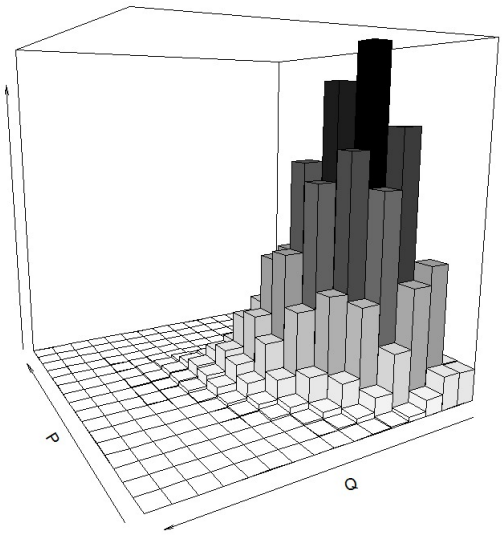
$$P(X = x, Y = y) = \frac{e^{-(\lambda_X + \lambda_Y)}(\lambda_X^x \lambda_Y^y)}{x!y!} \tag{2}$$

Dado a distribuição mostrada na equação 2, é possível denotar uma representação gráfica para ela. Como ela varia em duas dimensões sua representação é feita no espaço como mostra o exemplo da Figura 3.

Nesse estudo, ressaltou-se um caso especial dessa distribuição, que foi utilizada por um dos autores para modelar a probabilidade do número de gols do mandante e do visitante simultaneamente. Essa distribuição pertence a classe de Holgate e recebe o nome de distribuição de Poisson Bivariada "de Holgate" (ARRUDA, 2000). Em particular, ela considera que $X \sim \text{Poisson}(\lambda_X + \lambda_{XY})$ e $Y \sim \text{Poisson}(\lambda_Y + \lambda_{XY})$. Portanto sua função de probabilidade está representada abaixo. Para mais detalhes vide Arruda (2000, p.4).

$$P(X = x, Y = y) = e^{-(\lambda_X + \lambda_Y + \lambda_{XY})} \sum_{i=0}^{min(x,y)} \frac{\lambda_X^{x-i} \lambda_Y^{y-i} \lambda_{XY}^i}{(x-i)!(y-i)!i!}$$

Figura 3: Exemplo gráfico de uma distribuição Poisson Bivariada de P e Q.



Fonte: Adaptado de Astivia (2020).

Nessa distribuição, tanto λ_X, λ_Y e λ_{XY} são o parâmetros de distribuição de Poisson univariadas e independentes. Entretanto, λ_X e λ_Y representam uma taxa de ocorrência de gols dos times de maneira isolada. Já λ_{XY}, diz respeito a essa taxa dado o confronto, ou seja e escolha dos adversários.

3.5 ESTIMAÇÃO PONTUAL

Antes de definir estimação pontual, é importante situar esse assunto dentro da própria estatística, uma vez que ele é tratado no contexto da inferência estatística, ramo que se concentra em utilizar informações amostrais para obter estimativas de uma população de difícil ou até impossível acesso. Vale destacar que a inferência tratada aqui é a inferência paramétrica, ou seja, considera-se que a população à qual deseja-se inferir segue uma distribuição de probabilidade conhecida, mas com seus parâmetros desconhecidos (MOOD; GRAYBILL; BOES, 1974).

Assim, segundo Casella e Berger (2021), estimação pontual é o processo pelo qual se obtém um valor único para o parâmetro desejado. Ao tratar dessa temática, é muito comum que os livros abordem esse tema em dois subtópicos. Primeiro, métodos para encontrar estimadores pontuais, ou seja, estatísticas de uma amostra que assumem o valor do parâmetro desconhecido. Haja vista que existem casos em que não é possível obter um estimador por determinado método. Segundo, formas de avaliar os melhores estimadores com o intuito de escolher aqueles que apresentam melhores condições em relação à avaliação. Nas subseções a seguir, serão tratadas duas formas de se obter esses estimadores, uma vez que esses métodos foram escolhidos pelo autor de um dos modelos buscando estimar parâmetros λ de uma distribuição.

3.5.1 Método dos momentos

Esse método apresenta uma lógica bastante simples e coerente, geralmente é o ponto de partida ao buscar um bom estimador. Na maioria dos casos, esse método sempre retorna um estimador, mesmo que ele possa precisar de aprimoramento.

Em poucas palavras, esse método consiste em igualar os momentos amostrais ao momentos teóricos ou populacionais da distribuição. Assim, segue a definição do k-ésimo momento amostral ordinário (M'_k) de uma amostra de tamanho n, e a definição do k-ésimo momento populacional ordinário (μ'_k) (CASELLA; BERGER, 2021).

$$M'_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

$$\mu'_k = E[X^k]$$

Assim, considerando uma distribuições com parâmetros $\theta_1, \ldots, \theta_k$, é possível estimá-los resolvendo o seguinte sistema:

$$\begin{cases} M'_1 = \mu'_1 \\ \vdots \\ M'_k = \mu'_k \end{cases}$$

Esse sistema faz uma correspondência entre cada um dos momentos amostrais pelos seu respectivo momento populacional.

3.5.2 Método da máxima verossimilhança

Essa técnica é considerada a mais popular e poderosa para a estimação pontual, uma vez que quase sempre gera bons estimadores. Ainda, para utilizar esse método é preciso que a amostra seja independente e identicamente distribuída, para garantir que distribuição conjunta fatore como produto das marginas. Dessa forma, considerando uma população com função de probabilidade $f(x|\theta_1, \ldots, \theta_k)$, o estimador de máxima verossimilhança para os parâmetros $\theta_1, \ldots, \theta_k$; é obtido através da derivada da função de verossimilhança definida por (CASELLA; BERGER, 2021):

$$L(\theta|x) = L(\theta_1, \ldots, \theta_k|x_1, \ldots, x_k) = \prod_{i=1}^{n} f(x_i|\theta_1, \ldots, \theta_k)$$

Agora, para encontrar o valor estimado de cada θ_i, basta encontrar o ponto de máxima dessa função em relação a este parâmetro. Isso pode ser feito igualando a zero, a derivada parcial da função de verossimilhança, como mostrado abaixo:

$$\frac{\partial}{\partial \theta_i} L(\theta|x) = 0$$

Vale ressaltar que ao tomar-se o logaritmo dessa função os pontos de máximo não mudam. Por isso, é comum adotar a função log-verossimilhança para encontrar esse estimador, uma vez que esse cálculo pode ser mais fácil.

4 MATERIAL E MÉTODOS

Em termos gerais, neste estudo foram realizadas comparações entre três modelos de previsão de resultados. Essas comparações foram feitas utilizando quatro métricas, que podem ser vistas com mais detalhes na seção 4.3. As probabilidades dos modelos *SD 0* e *Chance I* foram obtidas utilizando o código disponibilizado pelo autor, Marcelo Arruda, com suas devidas implementações. Já as probabilidades do Modelo *UFMG* foram disponibilizadas pelo departamento de matemática da UFMG. Além disso, a implementação dos modelos e o cálculo das métricas foram feitos na linguagem de programação R (R Core Team, 2023)

4.1 BANCO DE DADOS

Os dados utilizados neste trabalho estão disponíveis na página da Confederação Brasileira de Futebol (CBF). Eles são apresentados no site por meio da Tabela Detalhada que registra todos os jogos que compõem o Campeonato Brasileiro de Futebol Série A de um determinado ano. Apenas as informações relevantes para os modelos foram filtradas, resultando no banco representado pela Tabela 1.

A escolha desse campeonato se deu pela regularidade, importância, estrutura e formato que é popularmente conhecido como sistema de pontos corridos. Mais detalhadamente, o campeonato conta com a participação de 20 times, 38 rodadas e 10 jogos por rodada. Cada time enfrenta os outros 19 duas vezes, uma como mandante, ou seja, o jogo ocorre no campo que ele escolheu, e outra como visitante, que ocorre no campo do adversário. O sistema de pontos distribui 3 pontos para uma vitória, 1 ponto para o empate e 0 para a derrota. Os times ocupam suas colocações no campeonato de acordo com o número de pontos. Além disso, os primeiros colocados garantem uma vaga para disputar a Copa Libertadores, enquanto os 4 piores colocados são rebaixados para o Campeonato Brasileiro Série B, não podendo disputar a Série A no próximo ano.

Tabela 1: Exemplo da estrutura do banco de dados

Data	**Mandante**	gol_M	$gols_V$	**Visitante**	**Neutro**	**Semestre**	**Peso**
13/05/17	Flamengo	1	1	Atlético Mineiro	NA	0	0.29843
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
27/12/20	Grêmio	2	1	Atlético Goianiense	NA	1	0.82411
06/01/21	Flamengo	1	2	Fluminense	*	0	1.02958
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
13/11/22	Ceara	4	1	Juventude	NA	1	1.20000

Fonte: Adaptado de: Confederação Brasileira de Futebol (2023).

Dado o exemplo acima, destacam-se as colunas chamadas “Neutro”e “Peso”que não são tão intuitivas quanto as demais. Detalhando a coluna da variável “Neutro”, que se

refere ao efeito casa, ela apresenta dois tipos de valores, "NA"para denotar que existe mando de campo e "*"para denotar que o jogo aconteceu em um campo neutro. Vale ressaltar que Arruda (2000) considera campo neutro quando dois times se enfrentam no campo que ambos utilizam como casa, popularmente esses jogos são chamados de clássicos. Um exemplo disso é o confronto entre Atlético Mineiro e Cruzeiro no Estádio Governador Magalhães Pinto (Mineirão), essa partida é conhecida como Superclássico Mineiro. Outra situação que caracteriza um campo neutro é quando o jogo acontece em um campo que nenhum dos dois times considera como casa ou está habituado a jogar. Para o modelo *UFMG*, apenas os clássicos são considerados campos neutros.

Quanto ao "Peso", é uma variável que atribui maior relevância para partidas que aconteceram mais próximas da data de realização da predição. Essa métrica parte do pressuposto de que os times se modificam durante o campeonato, tornando o modelo mais sensível a essas mudanças.

Portanto, foram feitas previsões para cada jogo do banco de dados em ordem cronológica, utilizando o próprio banco para calcular os parâmetros de previsão. Isso ocorreu da seguinte forma: para prever um jogo n do banco, considerou-se todos os $n-1$ jogos anteriores como informações para obter os parâmetros necessários para a previsão.

4.2 MODELOS DE PREDIÇÃO

Como o trabalho propõe a comparação dos resultados de modelos já existentes na literatura, pretende-se implementá-los aproximando-se o máximo possível do que foi feito pelos autores. Dessa forma, destaca-se a seguir as considerações e as metodologias que cada um dos modelos segue em sua implementação.

Além disso, a escolha dos modelos se pautou na popularidade, na metodologia e nos resultados apresentados. Uma vez que os modelos de Arruda (2000) foram pioneiros nesse tema e foram reproduzidos e modificados por alguns trabalhos, permitindo compará-lo com outras aplicações. E o modelo de Lima *et al.* (2012), denominado modelo *UFMG*, é um modelo patenteado, influente nas mídias digitais e com resultados promissores.

4.2.1 Modelo *SD 0* de Arruda (2000)

Esse modelo, utiliza uma abordagem estatística para tratar os dados e realizar as predições. Em especial, ele se concentra em prever X e Y que representam respectivamente o número de gols do time mandante ($time_m$) e do time visitante ($time_v$). Segundo Arruda (2000), X e Y seguem uma distribuição Poisson, porém o vetor (X,Y) tem uma distribuição de Poisson Bivariada, em particular da Classe de Holgate.

Assim, o autor considera as seguintes esperanças marginais para as variáveis X e Y: $E[X] = \lambda_x + \lambda_{xy}$ e $E[Y] = \lambda_y + \lambda_{xy}$. Sendo λ_x e λ_y parâmetros de uma distribuição Poisson Bivariada de Holgate que refletem a frequência de gols do $time_m$ e do $time_v$,

respectivamente. Já λ_{xy}, representa a covariância entre X e Y. Porém, o modelo *SD 0* considera essa covariância nula, ou seja, $\lambda_{xy} = 0$. Para maiores explicações vide Arruda (2000).

O autor utiliza o método dos momentos para a estimação pontual e um estimador indireto para calcular λ_x e λ_y. Obtemos os estimadores através da solução do sistema apresentado acima.da seguinte forma:

$$\begin{cases} E[X-Y] = \lambda_x - \lambda_y \\ E[X+Y] = \lambda_x + \lambda_y \end{cases} \qquad \begin{cases} \hat{\lambda_x} = \frac{\hat{E}[X-Y]+\hat{E}[X+Y]}{2} \\ \hat{\lambda_y} = \frac{\hat{E}[X+Y]-\hat{E}[X-Y]}{2} \end{cases}$$

Dessa forma, por meio de modelos lineares, como mostrado abaixo estima se $E[X+Y]$ e $E[X-Y]$.

$$(X+Y)_i = S_i\alpha + \varepsilon_{ai}$$
$$(X-Y)_i = T_i\beta + \varepsilon_{bi}$$

em que S_i e T_i são matrizes linhas com $n+1$ elementos, n referente ao número de times que estão participando no campeonato em questão, organizados em ordem alfabética e o último elemento se refere ao fator campo da partida i. Além disso, essa matriz é composta apenas por 0 e 1 sendo que, para S_i o elemento que ocupa o posição do time mandante na i-ésima partida assume 1 e o que está na posição do time visitante também assume valor 1, o restante é preenchido com 0, salvo a última posição relativa ao campo, que assume valor 0 se o campo for neutro, e 1 caso contrário. Já a matriz T_i difere da matriz S_i somente no valor que assume o time visitante que no caso é -1, o restante permanece igual. O exemplo abaixo ilustra a estrutura dessas matrizes, considerando S a união das matrizes linha S_i e T a união das matrizes linha T_i, ambas organizado pelo decorrer das partidas. Como exemplo observe três partidas de um Campeonato Brasileiro hipotético:

- Cruzeiro x Palmeiras no Mineirão
- Flamengo x Fluminense no Maracanã
- Avaí x Ceará no Estádio da Ressacada

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} Matriz\, S \qquad \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} Matriz\, T$$

Por meio de uma regressão linear simples, usando os placares de partidas anteriores, calcula-se os parâmetros α e β de cada um dos times presente no banco de dados, até uma dada rodada, como mostra o exemplo a seguir:

$$\hat{\alpha} = (S'S)^{-1}S'(X+Y)$$
$$\hat{\beta} = (T'T)^{-1}T'(X-Y)$$

4.2.2 Modelo *Chance I* de Arruda (2000)

Dentre os seis modelos apresentados pelo autor, o *Chance I* foi um dos que obtiveram os melhores resultados em sua aplicação. O funcionamento dele é muito semelhante ao que já foi estabelecido para o modelo *SD 0*, sendo que λ_x e λ_y são consideradas independentes e $X \sim Poisson(\lambda_x)$ e $Y \sim Poisson(\lambda_y)$, para uma dada partida do campeonato. Agora, para estimar λ_x e λ_y utiliza-se um modelo log-linear de Poisson e o método de máxima verossimilhança para estimar os parâmetros desse modelo. Assim, os logaritmos de suas funções de verossimilhança são dadas por:

$$L(\lambda_x, X) = -\lambda_x + xlog\lambda_x - log(x!)$$
$$L(\lambda_y, Y) = -\lambda_y + ylog\lambda_y - log(y!)$$

Segundo Arruda (2000), por meio da variável explicativa $\lambda = e^{U\beta}$, é possível relacionar a distribuição do número de gols marcados com as variáveis indicadoras, como mostrada a seguir.

$$L(\lambda_x, X) = -e^{U_x\beta} + xU_x\beta - log(x!)$$
$$L(\lambda_y, Y) = -e^{U_y\beta} + yU_y\beta - log(y!)$$

Uma vez que, U_x é um vetor que indica os times que estão participando do jogo em questão, de maneira análoga ao que foi definido para S_i do modelo *SD 0*. Contudo, U_x não se restringe apenas ao time mandante. E U_y da mesma forma que T_i indica além dos times, o mando de campo de cada um deles. Ainda, β informa parâmetros de ataque e defesa para os times do campeonato. Aplicando isso a um número k de jogos utilizados para estimar os parâmetros e unindo as informações de mandante e visitante têm-se a seguinte função de verossimilhança:

$$L(\lambda_{x_1}, \lambda_{y_1}, \ldots, \lambda_{x_k}, \lambda_{y_k}, X_1, Y_1, \ldots, X_k, Y_k) = \sum_{i=1}^{k}(-e^{U_{x_i}\beta} - e^{U_{y_i}\beta} + x_iU_{x_i}\beta + y_iU_{y_i}\beta - log(x_i!) - log(y_i!))$$

4.2.3 Modelo *UFMG* de Lima *et al.* (2012)

Pode-se dizer que este modelo tem caráter matemático, devido ao seu algoritmo que utiliza constantes para aumentar ou diminuir as métricas atribuídas à cada time. Apesar

disso, ele utiliza simulações para adicionar aleatoriedade ao processo de predição. Assim, o modelo trabalha com o seguinte esquema de vetores de probabilidades descrito abaixo:

Inicialmente, cada time possui dois vetores que buscam caracterizar o desempenho desse time quando joga como mandante PM e quando joga como visitante PV. Isso vale, para cada time que participa da competição que o autor deseja prever.

$$PM = (pvm, pem, pdm)$$
$$PV = (pdv, pev, pvv)$$

As incógnitas presentes nesses vetores de probabilidade são: probabilidade de vitória com mandante (pvm), probabilidade de empate com mandante (pem), probabilidade de derrota com mandante (pdm), probabilidade de derrota com visitante (pdv), probabilidade de empate com visitante (pev),probabilidade de vitória com visitante (pvv). Já o vetor que carrega as primeiras estimativas dos resultados de determinado confronto, é obtido pela média dos vetores PM do time mandante (tm) e PV do time visitante (tv), como mostrado a seguir

$$\begin{aligned}(vm, em, vv)_{tm \times tv} &= \frac{PM_{tm} + PV_{tv}}{2} \\ &= \left(\frac{pvm_{tm} + pdv_{tv}}{2}, \frac{pem_{tm} + pev_{tv}}{2}, \frac{pdm_{tm} + pvv_{tv}}{2}\right)\end{aligned}$$

Entretanto, ainda existe o problema de determinar PM e PV de cada time. Esse vetores são dinâmicos e atualizados após cada confronto que o time participa e de acordo com o evento resultante observado (vitória mandante, empate ou vitória visitante). Os valores dos vetores PM^0 e PV^0 são escolhidos arbitrariamente pelo autor, que geralmente inicia escolhendo valores próximos de 0.33 para cada uma das componentes. Em seguida, esse vetor é atualizado segundo o algoritmo apresentado abaixo, o qual ilustra o confronto de um time tm mandante contra tv visitante.

Caso ocorra vitória do mandante:

$$PM_{tm}^{k+1} = \frac{p \cdot PM_{tm}^{k} + R_{tv}(1,0,0)}{p + R_{tv}}$$
$$PV_{tv}^{k+1} = \frac{p \cdot PV_{tv}^{k} + (1 - R_{tm})(1,0,0)}{p + (1 - R_{tm})}$$

Caso ocorra vitória do visitante:

$$PM_{tm}^{k+1} = \frac{p \cdot PM_{tm}^{k} + (1 - R_{tv})(0,0,1)}{p + (1 - R_{tv})}$$
$$PV_{tv}^{k+1} = \frac{p \cdot PV_{tv}^{k} + R_{tm}(0,0,1)}{p + R_{tm}}$$

Onde p é um fator peso que regula a sensibilidade do ajuste dos vetores. Uma vez que, quanto maior o valor de p maior será o impacto da ultima atualização. Já R representa o rendimento de um determinado times que pode ser calculado da seguinte forma:

$$R = \frac{\text{nº de vitórias} + 0.5 \cdot \text{nº de empates}}{\text{nº total de jogos}}$$

Agora para o caso de empate, considera-se ainda o rendimento dos times que se enfrentam, como mostrado abaixo:

Caso de empate em que o $R_{tv} \leqslant 0.50$

$$PM_{tm}^{k+1} = \frac{p \cdot PM_{tm}^{k} + (1 - 2 \cdot R_{tv})(0, 0.5, 0.5) + 2 \cdot R_{tv}(0,1,0)}{p + 1}$$

Caso de empate em que o $R_{tv} \geqslant 0.50$

$$PM_{tm}^{k+1} = \frac{p \cdot PM_{tm}^{k} + (2 \cdot R_{tv} - 1)(0.5, 0.5, 0) + 2 \cdot (1 - R_{tv})(0,1,0)}{p + 1}$$

Aplica-se de maneira análoga para o vetor PV_{tv}^{k+1}, o mesmo raciocínio, considerando o rendimento do time mandante. Além disso, para jogos considerados clássicos os autores consideram ambos os times como mandante para as atualizações dos vetores.

O próximo passo trata-se de encorporar aleatoriedade no processo de predição por meio de simulações, que consistem em sortear um número aleatório γ, compreendido no intervalo "real"[0,1]. Demarcado o valor de γ tem-se:

- se $0 \leqslant \gamma \leqslant vm$, considera-se que ocorreu vitória do mandante.
- se $vm \leqslant \gamma < em + vm$, considera-se que ocorreu empate.
- se $em + vm \leqslant \gamma \leqslant 1$, considera-se que ocorreu vitória do visitante.

Para obter as proporções de ocorrência de cada evento, os autores reproduzem esse sorteio 10000 vezes. Obtendo agora, a probabilidade de ocorrência de cada evento.

4.3 MÉTRICAS DE ACURÁCIA E PRECISÃO

Parte fundamental da aplicação de um modelo são: medir sua acurácia, saber se ele acertou ou não e se possível determinar o quanto acertou ou errou. Portanto, para esse estudo utilizou-se a taxa de acerto (TA), o erro preditivo médio ponderado (EPMP), a medida de DeFinetti (DF) e a medida de Definetti detalhada (DFD) para avaliar o modelo. Abaixo estão definidas as métricas que não foram detalhadas na seção anterior.

4.3.1 Taxa de Acerto

Para estabelecer a taxa de acerto, primeiro é necessário definir outra métrica, o erro preditivo médio (EPM). Uma vez que, quando se trata de predições e estimação de parâmetros é bastante comum utilizar o erro preditivo quadrático médio. Porém, como nesse estudo essa métrica não assume valores negativos, considerou-se apenas erro preditivo médio.

Para observar o comportamento do modelo com o passar das rodadas e consequentemente o aumento da amostra, foi calculado o EPM para cada uma das rodadas, da seguinte maneira:

$$EPM_j = \frac{1}{n}\sum_{i=1}^{n} I(\theta_{ij}, \hat{\theta}_{ij})$$

Em que EPM_j é o Erro Preditivo Médio do modelo na rodada j; θ_{ij} é o resultado do i-ésimo jogo da j-ésima rodada (utilizando a escala: 1 para vitória do mandante, 0 para empate, e -1 para derrota do mandante) e $\hat{\theta}_{ij}$ é o resultado predito pelo modelo para o jogo i da rodada j, na mesma escala; e $I(\cdot, \cdot)$ é uma função indicadora de erro, ou seja, que assume 0 quando $\theta_{ij} = \hat{\theta}_{ij}$ (acerto) e 1, caso contrário (erro). O EPM calculado dessa forma, expressa se o modelo acertou ou não e faz a média de quanto o modelo errou naquela rodada.

A taxa de acerto é um métrica bastante simples, uma vez compreendido o erro preditivo médio. Como nesse estudo, almeja-se determinar o modelo que apresenta melhores resultados, faz mais sentido classifica-los por um medida de acerto no lugar de uma medida de erro. Por isso, TA ficou definida como:

$$TA = 1 - EPM$$

Ainda, todas as demais métricas mensuram erro. Assim, para um leitor que busca identificar o "melhor" modelo, acredita-se que uma medida que representa o quanto o modelo acertou, atrairia mais a curiosidade desse leitor.

4.3.2 Erro preditivo médio ponderado

Ao se avaliar a acurácia de modelos como esses, que distribuem as probabilidades de vitória do mandante, empate e vitória do visitante, predizendo que o resultado que ocorrerá é aquele que apresenta maior probabilidade entre os três eventos possíveis, parece razoável ponderar o quanto esse modelo se desviou do resultado verdadeiro. Por exemplo, se o modelo 1 estabelece que ocorrerá empate com 80% de chance em determinado confronto, e o modelo 2 afirma que ocorrerá empate com 55% de chance, no mesmo confronto, caso o resultado verdadeira não seja empate, é evidente que o modelo 1 errou mais que o modelo 2. Esse métrica surge com o proposito de quantificar esse tipo de erro, para além do determinismo do EPM. Então, define-se o erro preditivo médio ponderado (EPMP) como:

$$EPMP_j = \frac{1}{n}\sum_{i=1}^{n} I(\theta_{ij}, \hat{\theta}_{ij}) p^*_{ij}$$

sendo

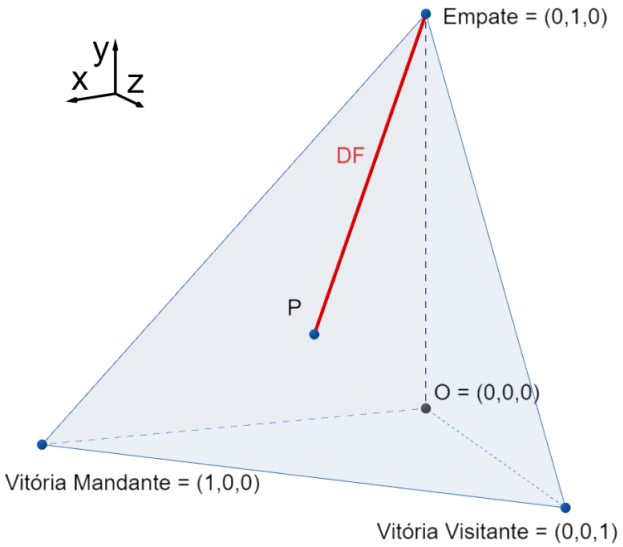
$$p^*_{ij} = \max(p^v_{ij}, p^e_{ij}, p^d_{ij})$$

em que p^v_{ij}, p^e_{ij} e p^d_{ij} representam as probabilidades de vitória, empate e derrota do mandante preditas pelo modelo no jogo i da rodada j; e n é o número de partidas naquela rodada. O restante da formula é análoga ao que já foi definido para o EPM.

4.3.3 Medida de De Finetti

Outro medida de acurácia bastante utilizada para avaliar a qualidade das predições de modelos como os utilizados nesse trabalho é a medida de DeFinetti. Essa medida foi originalmente proposta por DeFinetti (1972), como uma alternativa de métrica para eventos tricotômicos. Assim, esse medida consiste na distância euclidiana quadrática de dois pontos em um espaço $\mathbb{R}^3$. Os 3 eixos desse espaço são ortogonais entre si e limitados de 0 a 1, tendo sua origem no 0. A distância que denota essa métrica é calculada a partir do ponto P com coordenadas (VM, EM, VV) dadas pelas probabilidades do modelos, até o vértice V desse tetraedro que represente o evento observado. Sendo que, VM, EM e VV representam respectivamente, a probabilidade de acontecer vitória do mandante, empate e vitória do visitante. Como ilustrado na Figura 4.

Figura 4: Representação gráfica da medida de DeFinetti (DF)



Fonte: Dos autores.

O cálculo dessa medida acontece de três maneiras a depender do evento observado segundo as seguintes equações:

Se ocorreu vitoria do mandante $V = (1, 0, 0)$, logo:

$$DF = (VM - 1)^2 + (EM - 0)^2 + (VV - 0)^2$$

Se ocorreu empate $V = (0, 1, 0)$, logo:

$$DF = (VM - 0)^2 + (EM - 1)^2 + (VV - 0)^2$$

Se ocorreu vitoria do visitante $V = (0, 0, 1)$, logo:

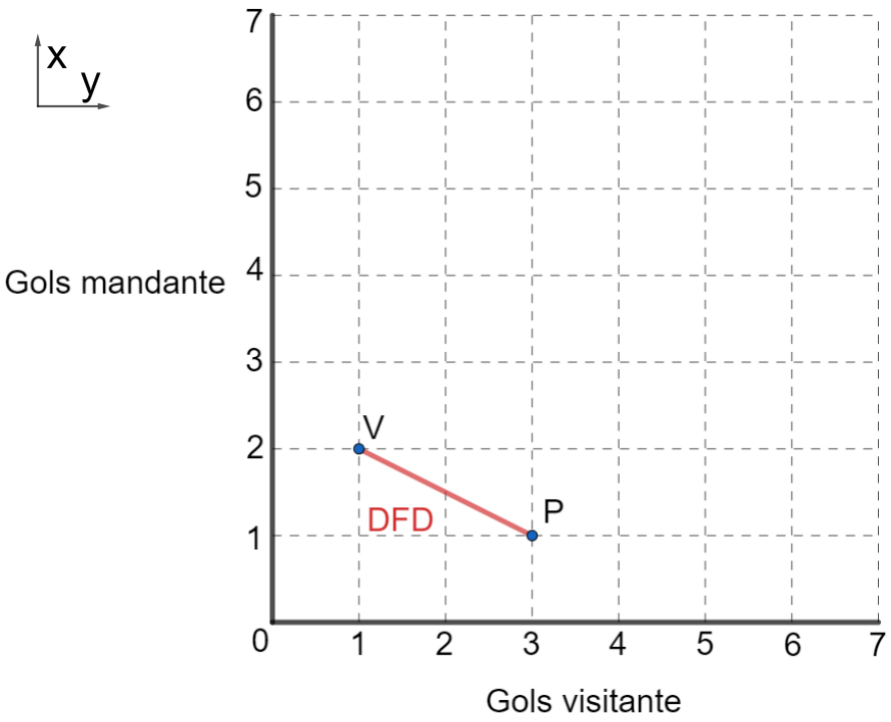
$$DF = (VM - 0)^2 + (EM - 0)^2 + (VV - 1)^2$$

Para ilustrar esse calculo imagina-se a partida hipotética Flamengo $\times$ Atletico Mineiro, que o modelo determinou vitória do visitante com as seguintes probabilidades $(0.34, 0.28, 0.38)$. Entretanto, ocorreu empate, assim o valor de DF será igual à 0.7784.

4.3.4 Medida de DeFinetti Detalhada

É proposta desse trabalho a seguinte métrica que se inspira na medida de DeFinetti e segue os pressupostos do erro preditivo médio ponderado. Nesse sentido, considere a medida de DeFinetti detalhada como a distância euclidiana entre o ponto P e V em um plano cartesiano. Sendo que P tem coordenadas dadas pelo placar que o modelo considera mais provável e V pelo placar observado, como mostra a Figura 5.

Figura 5: Representação gráfica da medida de DeFinetti Detalhada (DFD).



Fonte: Dos autores.

O calculo dessa medida ocorre da seguinte forma:

$$DFD = \sqrt{(gmp - gmo)^2 + (gvp - gvo)^2} \tag{3}$$

Onde *gmp* e *gmo* representam os números de gols do mandante previsto e observado respectivamente. Já *gvp* e *gvo* são os números de gols do visitante previsto e observado respectivamente. Diferente do exemplo anterior esse métrica é sensível ao quanto a predição se aproximou do resultado verdadeiro. Assim, considerando o exemplo do confronto Atlético Mineiro × Internacional, se o modelo determinou um placar de 1×3, mas o placar verdadeiro foi de 2×1, a DFD foi de 2.236068.

5 RESULTADOS E DISCUSSÃO

Nessa seção, serão apresentados os resultados alcançados após a aplicação dos modelos e a análise do banco de dados. Primeiro, serão discutidos aspectos descritivos dos campeonatos observados e dos times participantes. Em seguida, serão discutidos os valores das métricas alcançadas por cada modelo.

5.1 ANÁLISE DESCRITIVA

Uma vez que os modelos se baseiam no número de gols, ou no número de vitórias e rendimento que os times conquistam ao decorrer do campeonato. Vale a pena observar e descrever o banco de dados em relação à esses fatores. Assim, a primeira relação estabelecida pelos modelos que pode ser constatada observando o banco de dados, é o fator casa. Uma vez que, os modelos demonstram uma "preocupação" no que se refere ao local de realização das partidas, logo espera-se que haja uma vantagem em jogar em casa. De fato, como observado na Tabela 2 que representa um resumo da Tabela 5 trazida no Anexo I, é possível observar uma diferença positiva no número de gols feitos como mandante em relação aos feitos como visitante.

Tabela 2: Número de gols feitos em relação ao mando de Campo.

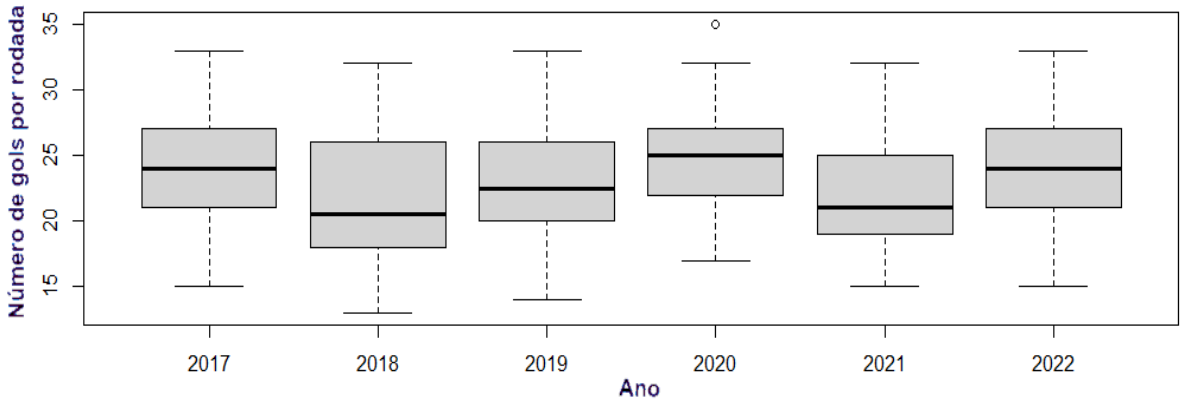
Ano	2017	2018	2019	2020	2021	2022	Total
Mandante	526	524	525	535	483	534	3127
Visitante	397	302	350	409	359	370	2187
Total	923	826	875	944	842	904	5314

Fonte: Dos Autores

Considerando todos os anos analisados, 58.8% dos gols foram feitos por times na posição de mandante. Em 2018 essa diferença foi mais expressiva, cerca de 63.4%. A diferença miníma aconteceu em 2020, 56.6%. Vale recordar que o campeonato referente a este ano aconteceu com as arquibancadas vazias, devido as medidas de proteção contra o coronavírus. Trabalhos futuros podem se debruçar sobre essa especifidade com o intuito de compreender os impactos das arquibancadas no desempenho do time. Além disso, as médias de gols observadas nos campeonatos vão em consonância com o que já foi dito anteriormente. Já que a média geral de gols feitos considerando todos os anos foi de 1.16, mas a média de gols feitos como mandante e vistante foram de 1.37 e 0.95, respectivamente. Mostrando novamente uma predisposição maior para os times mandante marcarem mais gols.

É possível analisar a disposição do número de gols por outra perspectiva. Em particular em relação as rodadas de cada um dos campeonatos como mostra a Figura 6.

Figura 6: Gráfico do número de gols por rodada de cada ano.



Fonte: Dos Autores

O número de gols feitos por rodada dentro dos campeonatos observados flutuou de 13 à 35, sendo 23.3 gols, o seu valor médio. Baseando se nessa média, pode-se pensar que como são realizados 10 jogos por rodada, são feitos 2.33 jogos por confronto. Assim, espera-se uma maior frequência de "*0 x 2*", "*1 x 1*", "*2 x 0*"como placares. Nos anos de 2020 e 2018 observou-se, respectivamente, a maior (24.8) e a menor (21.7) média de gols por rodada. Apesar dos dados apresentarem 22 gols de amplitude entre as rodadas, a amplitude das médias é igual a 3.1 gols.

Outra perspectiva possível para analisar o número de gols, é em relação a cada times que participou de pelo menos um dos campeonatos analisados, como mostra a Tabela 3. Nessa tabela, é possível observar o número de gols que cada time fez como mandante (M), visitante (V) e o total (T) para cada campeonato que ele participou. Ainda, uma importante informação é a média de gols que o time faz por campeonato (Mgols). Esse valor pode ser interpretado como a constância de ataque do time.

Vale lembrar, que como o Campeonato Brasileiro Série A acontece na modalidade de pontos corridos, ou seja, um maior número de pontos garante que o time será campeão e não o maior número de gols. Dessa forma, mesmo um time com uma baixa Mgols pode ser campeão, desde que garante sua vitória que equivale a 3 pontos.

Trinta times estiveram envolvidos nesse intervalo de competições, oito deles, Athletico Paranaense, Atlético Mineiro, Corinthians, Flamengo, Fluminense, Palmeiras, Santos e São Paulo, estiveram presentes em todos. Por outro lado, onze times participaram em 3 ou menos anos de competições, mostrando o caráter rotativo da competição.

Tabela 3: Número de gols e média de gols (Mgols) de cada time.

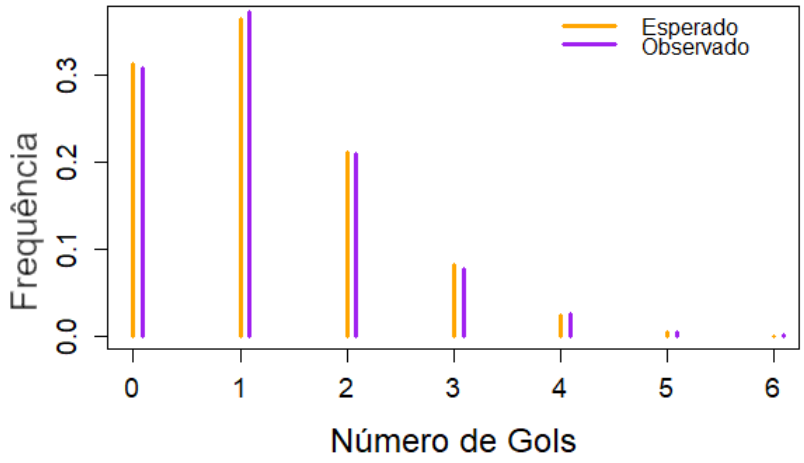
	2017			2018			2019			2020			2021			2022			
times	M	V	T	M	V	T	M	V	T	M	V	T	M	V	T	M	V	T	Mgols
America	-	-	-	22	8	30	-	-	-	-	-	-	22	19	41	22	18	40	37,0
AthleticoPr	29	16	45	44	10	54	26	25	51	20	18	38	23	18	41	31	17	48	46,2
AtleticoGo	18	20	38	-	-	-	-	-	-	24	16	40	17	16	33	22	17	39	37,5
AtleticoMg	27	25	52	31	25	56	28	17	45	41	23	64	42	25	67	25	20	45	54,7
Avaí	15	14	29	-	-	-	10	8	18	-	-	-	-	-	-	24	10	34	27,0
Bahia	34	16	50	27	12	39	24	20	44	26	21	47	27	15	42	-	-	-	44,4
Botafogo	30	15	45	22	16	38	19	12	31	16	17	33	-	-	-	18	23	41	37,6
Bragantino	-	-	-	-	-	-	-	-	-	34	17	51	28	27	55	29	20	49	51,7
Ceará	-	-	-	18	14	32	23	13	36	22	28	50	24	15	39	20	14	34	38,2
Chapecoense	24	23	47	24	10	34	16	15	31	-	-	-	13	14	27	-	-	-	34,7
Corinthians	32	18	50	19	15	34	25	17	42	29	16	45	26	14	40	24	20	44	42,5
Coritiba	21	21	42	-	-	-	-	-	-	12	18	30	-	-	-	25	14	39	37,0
Cruzeiro	26	21	47	25	9	34	13	14	27	-	-	-	-	-	-	-	-	-	36,0
CSA	-	-	-	-	-	-	17	7	24	-	-	-	-	-	-	-	-	-	24,0
Cuiabá	-	-	-	-	-	-	-	-	-	-	-	-	18	16	34	17	14	31	32,5
Flamengo	32	17	49	30	29	59	56	30	86	35	33	68	31	38	69	37	23	60	65,2
Fluminense	27	23	50	17	15	32	18	19	37	34	24	58	24	14	38	39	24	63	46.3
Fortaleza	-	-	-	-	-	-	26	24	50	20	13	33	26	18	44	24	22	46	43,3
Goiás	-	-	-	-	-	-	31	15	46	19	25	44	-	-	-	20	20	40	43,3
Grêmio	26	29	55	35	12	47	38	26	64	33	18	51	29	15	44	-	-	-	52,2
Internacional	-	-	-	32	19	51	28	16	44	34	27	61	29	15	44	40	18	58	51,6
Juventude	-	-	-	-	-	-	-	-	-	-	-	-	19	17	36	18	11	29	32,5
Palmeiras	35	26	61	42	22	64	40	21	61	33	18	51	33	25	58	39	27	66	60,2
Paraná	-	-	-	13	5	18	-	-	-	-	-	-	-	-	-	-	-	-	18,0
Ponte Preta	26	11	37	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	37,0
Santos	25	17	42	28	18	46	44	16	60	35	17	52	21	14	35	28	15	43	46,3
São Paulo	29	19	48	25	21	46	23	16	39	29	30	59	18	13	31	32	23	55	46,3
Sport	27	19	48	19	16	35	-	-	-	19	13	32	13	11	24	-	-	-	34,8
Vasco	21	19	40	29	12	41	20	19	39	20	17	37	-	-	-	-	-	-	39,3
Vitória	22	28	50	22	14	36	-	-	-	-	-	-	-	-	-	-	-	-	43,0

Fonte: Dos Autores

Os três times com as melhores médias de gols analisadas foram o Flamengo, o Palmeiras e o Atlético Mineiro, esses times foram campeões dessa competição nos anos de 2020, 2022 e 2021 respectivamente. Ainda, os times que apresentaram a maior diferença relativa do gols feitos como mandante e visitante foram: Paraná e CSA . Contudo, esses times participaram somente de um dos seis campeonatos observados. Esse fato, levanta hipóteses sobre a métrica usada para dimensionar o efeito casa. Nesse sentido, uma sugestão de abordagem e métrica para determinar esse fator seria a métrica baseada em pontos proposta em Paludo, Figueiredo e Ferreira (2023), que segundo os autores essa métrica é menos inflacionada por dados não uniformes.

Outra pergunta que pode ser respondida utilizando a análise descritiva dos dados, é sobre a escolha da distribuição de Poisson para modelar o número de gols de um time nos anos do campeonato analisados. Nesse sentido, utilizando a frequência do número de gols feitos por cada time do banco de dados em comparação com a frequência teórica produzida a partir da média do número de gols, foi possível obter a Figura 7.

Figura 7: Comparação entre a distribuição do número de gols observada e esperada.



Fonte: Dos Autores

Como visto na Figura 7, os observações empíricas e teóricas são visualmente muito semelhantes. Uma vez que houve uma sobreposição da distribuição dos dados e de uma distribuição de Poisson com $\lambda = 1.16$. Além dessa constatação visual, o teste de Kolmogorov-Smirnov que é usado para avaliar se duas amostras provêm da mesma distribuição, constatou com 99% de confiança, que não existem evidências estatística significativas para rejeitar a hipótese de que essas duas distribuições amostrais derivam de uma mesma distribuição populacional. Portanto, a distribuição de Poisson é amplamente usada para modelas a probabilidade desse evento.

5.2 MODELOS E MÉTRICAS

Ao final dos processos computacionais e matemáticos, para os modelos *SD 0* e *Chance I* foram obtidas 2250 predições, apesar do banco conter 2280 partidas. Foram necessárias 30 partidas, ou seja, as 3 primeiras rodadas do campeonato 2017, para o modelo de regressão linear generalizado do modelo *Chance I* convergir. Com tais predições somadas as probabilidades do modelo *UFMG*, foi possível calcular as medidas de acurácia para os modelos. A Tabela 4, ilustra a saída de cada um dos modelos para a 4ª rodada do campeonato de 2019. Os demais dados frutos da implementação dos modelos podem ser encontrados em: https://github.com/Arcocotg/Previs-o-esportiva.git .

Tabela 4: Probabilidades dos resultados dos jogos da 4ª rodada do Campeonato de 2019 de cada um dos modelos analisado.

Modelo	Time Man	Time Vis	Prob. VM	Prob. E	Prob. VV	Gols_m	Gols_v
SD 0	athleticopr	bahia	0,5527	0,2472	0,1996	1	0
	atleticomg	palmeiras	0,3816	0,2478	0,3702	1	1
	avai	csa	0,6907	0,2457	0,0633	1	0
	corinthians	gremio	0,4214	0,3076	0,2708	1	0
	flamengo	chapecoense	0,6783	0,2130	0,1079	1	0
	fluminense	botafogo	0,4569	0,2690	0,2738	1	1
	fortaleza	saopaulo	0,2431	0,2530	0,5036	1	1
	goias	ceara	0,4690	0,2943	0,2365	1	0
	internacional	cruzeiro	0,5947	0,2826	0,1226	1	0
	santos	vasco	0,6008	0,2524	0,1464	1	0
Chance I	athleticopr	bahia	0,4265	0,2679	0,3054	1	1
	atleticomg	palmeiras	0,2540	0,2388	0,5066	1	1
	avai	csa	0,4885	0,3552	0,1561	0	0
	corinthians	gremio	0,3009	0,3138	0,3851	0	1
	flamengo	chapecoense	0,5576	0,2471	0,1948	1	0
	fluminense	botafogo	0,3346	0,2743	0,3909	1	1
	fortaleza	saopaulo	0,1292	0,2153	0,6547	0	1
	goias	ceara	0,3048	0,3065	0,3885	0	1
	internacional	cruzeiro	0,4302	0,3205	0,2491	1	0
	santos	vasco	0,4503	0,2986	0,2509	1	0
UFMG	athleticopr	bahia	0,4666	0,2667	0,2667	-	-
	atleticomg	palmeiras	0,3900	0,3550	0,2550	-	-
	avai	csa	0,3966	0,3517	0,2517	-	-
	corinthians	gremio	0,3900	0,3550	0,2550	-	-
	flamengo	chapecoense	0,4666	0,2667	0,2667	-	-
	fluminense	botafogo	0,2971	0,2532	0,4497	-	-
	fortaleza	saopaulo	0,4100	0,2700	0,3200	-	-
	goias	ceara	0,4122	0,2634	0,3244	-	-
	internacional	cruzeiro	0,4666	0,2667	0,2667	-	-
	santos	vasco	0,4666	0,2667	0,2667	-	-

Prob. VM = Probabilidade de vitória do mandante; Prob.EM = Probabilidade de empate; Prob. VV = Probabilidade de vitória do visitante; Gols_m = Número de gols mais provável para o mandante; Gols_v = Número de gols mais provável para o visitante.

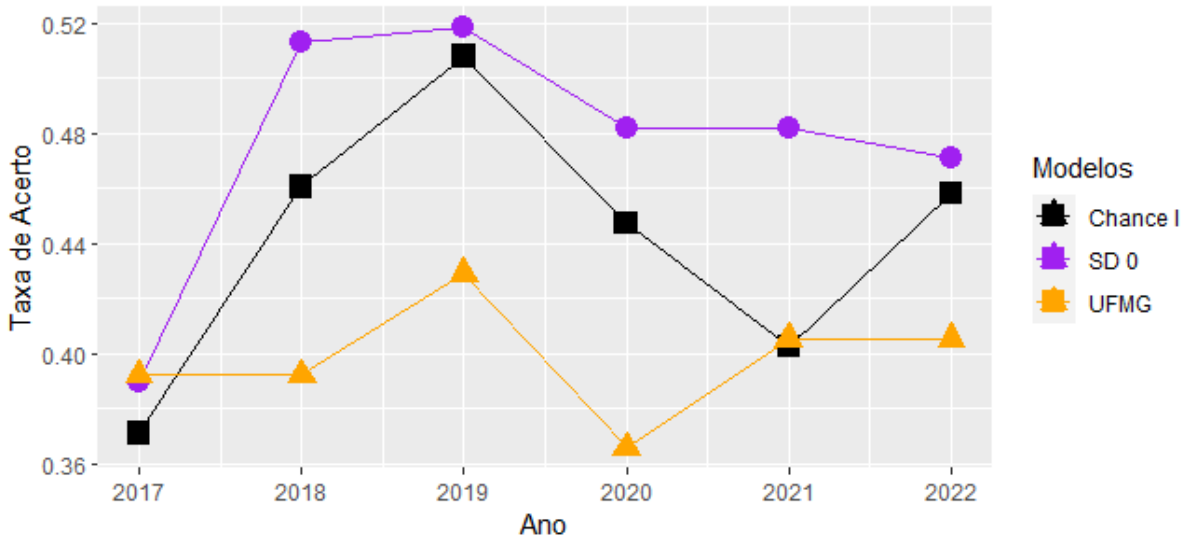
Fonte: Dos autores.

Observando a Tabela 4, nota-se que o modelo *SD 0* apresenta, em sua maioria, valores de probabilidades menos equilibradas, isto é, existem probabilidades acima de 50% para

um dos três eventos possíveis. Ao contrario do modelo *UFMG*, que não apresenta probabilidades maiores que 47% para nenhum resultado. Ainda, as colunas $Gols_m$ e $Gols_v$ para este modelo não apresentam nenhum valor, pois ele não foi implementado pelos autores do presente estudo e foi obtidos do Departamento de Matemática de UFMG, apenas as probabilidades de ocorrência de cada evento.

Os gráficos a seguir buscam estratificar as métricas, permitindo analisar e visualizar o desempenho dos modelos estudados em relação à cada uma dos critérios avaliados. Como cada métrica carrega seu algoritmo próprio e representa atributos distintos, torna-se necessário descrevê-los separadamente. Partindo da taxa de acerto apresentado na Figura 8, pode-se observar o valor médio dessa métricas para cada um dos anos do campeonatos observados.

Figura 8: Gráfico de comparação da Taxa de Acerto dos modelos por ano de competição, expressa em porcentagem.



Fonte: Dos Autores

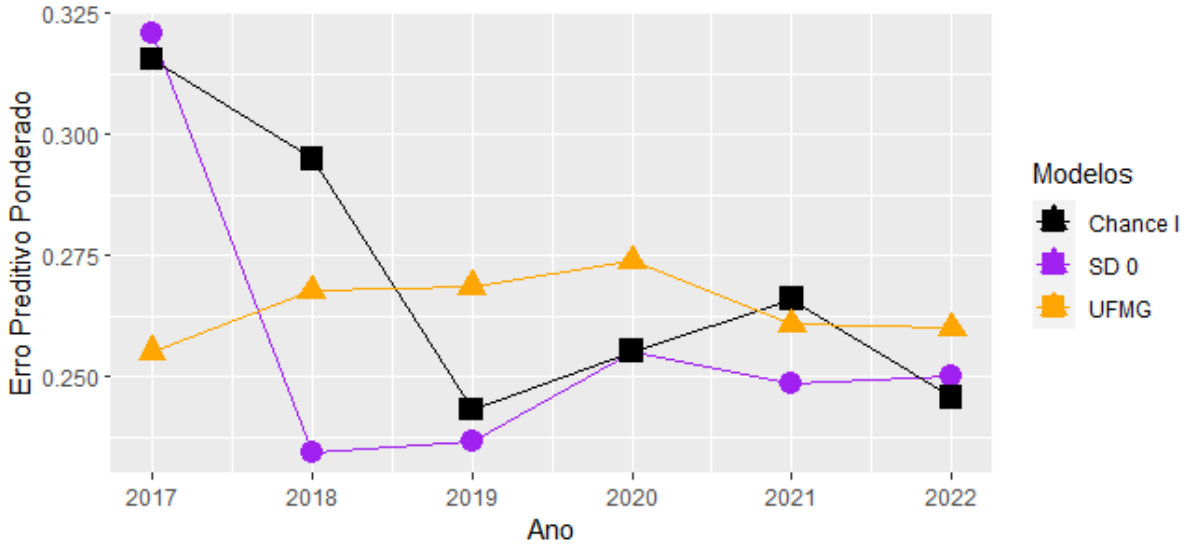
Em relação à taxa de acerto, destaca-se que, com exceção do ano de 2017 no qual as métricas apresentaram valores em torno de 38%, o modelo *SD 0* demonstrou superioridade em todos os outros anos. O ápice dessa métrica foi atingido em 2019, registrando um valor máximo de 51,8%. O modelo *Chance I*, embora não tenha apresentado resultados tão satisfatórios, parece seguir uma tendência semelhante ao modelo superior. Em contraste, o modelo da *UFMG* revelou-se equilibrado, mantendo-se em torno de uma taxa de acerto de 40%.

Na literatura, resultados desse tipo são recorrentes. Por exemplo, Ramos, Fernandes e Batista (2021) obteve uma taxa de acerto de 54,8% em sua aplicação em 2018. Da mesma forma, Suzuki *et al.* (2010) alcançou 57,8% aplicando na Copa do Mundo de 2006, Araújo *et al.* (2015) obteve 52,9% em sua aplicação no Campeonato Brasileiro Série A de

2014, Tavares e Suzuki (2015) registrou 53,0% em sua aplicação no Campeonato Brasileiro Série A de 2013, e Santana *et al.* (2020), com seu modelo estatístico que utiliza machine learning, atingiu uma taxa de 40%. Essa consistência de resultados na literatura reforça a complexidade do desafio de previsão esportiva e sugere que alcançar taxas de acerto substancialmente elevadas é uma tarefa desafiadora, mesmo para modelos estatísticos avançados.

Quanto ao erro preditivo médio ponderado, trazido na Figura 9, vale lembrar que trata-se de uma medida de erro. Ou seja, o melhores valores são aqueles mais próximos de zero. O mesmo vale para as outras 2 métricas restantes, já que também são medidas de erro.

Figura 9: Gráfico de comparação do Erro Preditivo Médio Ponderado.

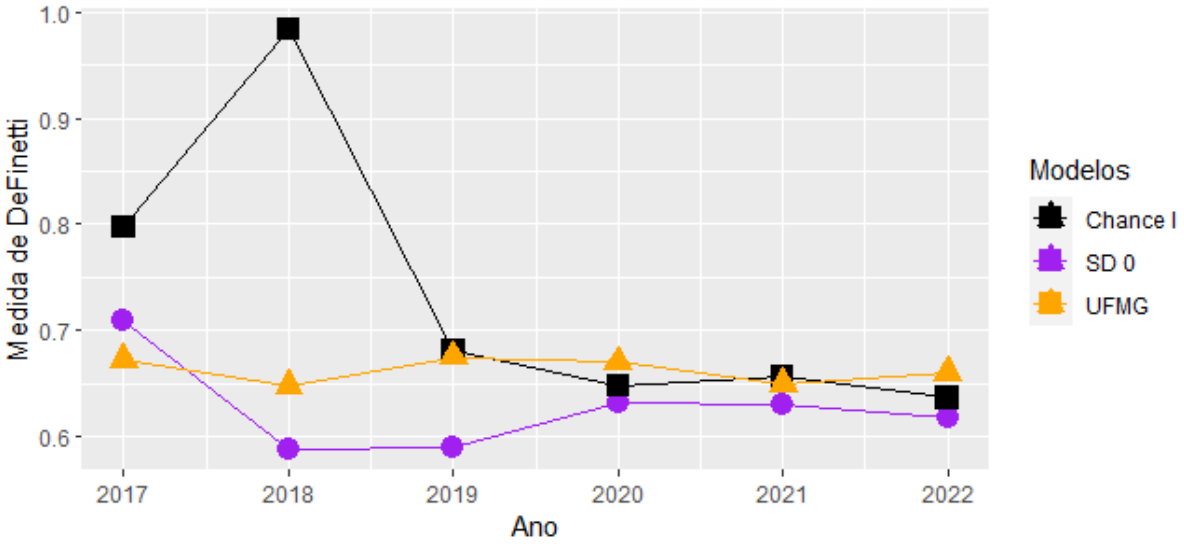


Fonte: Dos Autores

Da Figura 9, observa-se que novamente que o modelo *SD 0*, em geral, apresentou melhores métricas. Em 2017 e 2022 o modelo *Chance I* conseguiu superá-lo ainda que por pouco. Já em 2020, as métricas desses dois modelos se equipararam. O modelo da *UFMG*, saiu na frente em 2017, porém manteve sua característica de estabilidade e permaneceu com métricas em torno de 0.26.

Para a Medida de DeFinetti, Foi possível observar uma mudança em relação ao modelo *Chance I*, como observado na Figura 10, a seguir:

Figura 10: Gráfico de comparação da Medida de DeFinetti.



Fonte: Dos Autores

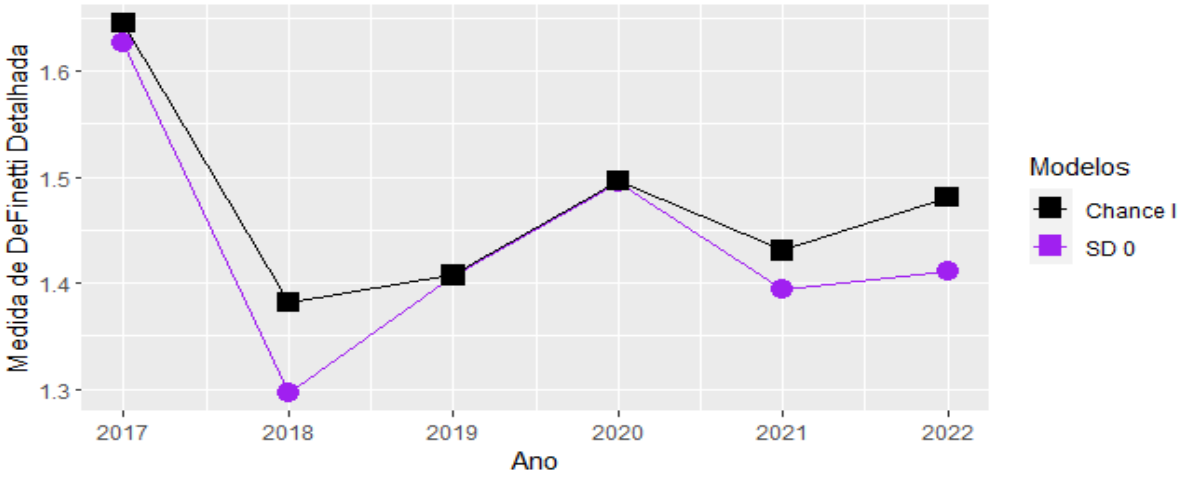
Embora o desempenho do modelo *UFMG* tenha se destacado em comparação com o modelo *Chance I* nesta métrica, ainda assim, as métricas registradas foram inferiores em comparação com o modelo *SD 0*. Contudo, é interessante notar uma peculiaridade em relação ao modelo *chance I* no ano de 2018, no qual ele quase atingiu o valor de $DF \approx 1$. Contrariamente ao que possa parecer, isso não indica que o modelo cometeu erros em todas as previsões nesse ano, mas sim, que em média houve uma considerável disparidade entre as probabilidades, que são componentes do ponto P de previsão, e o vértice V resposta.

O valor médio da métrica para o modelo *SD 0* foi de 0.628, revelando uma consistência notável e valores bastante próximos aos encontrados na literatura. Por exemplo, Filho *et al.* (2017) registrou 0.595, Araújo *et al.* (2015) obteve 0.560, e Tavares e Suzuki (2015) apresentou 0.609. Essa convergência destaca a estabilidade do modelo e a preservação das características fundamentais do campeonato ao longo do tempo. Mesmo com a aplicação em diferentes anos, observa-se a manutenção das características centrais que o modelo utiliza para suas previsões.

O próprio autor dos modelos também registrou medidas semelhantes, com 0.6247 de DF para o modelo *SD 0* e 0.6226 para o modelo *Chance I*. Além disso, Santana *et al.* (2020) apresenta uma medida comparável chamada de Brier Score, a qual atingiu o valor de 0.6618. Esses resultados congruentes sugerem que o modelo *SD 0* não apenas mantém sua estabilidade ao longo do tempo, mas também preserva sua eficácia em termos de previsões, o que é essencial para a confiabilidade e utilidade prática desses modelos estatísticos. Essa coerência com outros estudos valida ainda, a robustez do modelo *SD 0* no contexto da previsão estatística esportiva.

Para a medida de DeFinetti detalhada, é importante lembrar que não foi possível calculá-la para o modelo *UFMG* uma vez que é necessário conhecer o placar mais provável, para compará-lo com o placar observado.

Figura 11: Gráfico de comparação da Medida de DeFinetti Detalhada.



Fonte: Dos Autores

Por fim, a última métrica mostrou o quanto os modelos estatísticos se afastaram do exato resultado do placar. Nesse sentido, foi possível constatar visualmente que ambos os modelos seguiram uma mesma tendencia de valores. Em 2017, 2019 e 2020 os resultados foram bastante próximos, até iguais considerando algum critério de arredondamento menos rigoroso. Nos demais anos, o modelo que sobressaiu foi novamente o modelo *SD 0*.

Além disso, nota-se, de forma geral, que os gráficos tendem a convergir para uma área de menor amplitude a partir de 2019. Trabalhos futuros podem investigar se essa tendência está relacionada à expansão do banco de dados, uma vez que, a cada predição, mais dados foram incorporados na estimação dos parâmetros. Uma pesquisa adicional pode estabelecer um número mínimo para que as previsões tenham uma resposta considerada aceitável, bem como um número máximo ou uma data limite que indique quando essas partidas deixam de ser tão informativas para o processo de predição.

6 CONSIDERAÇÕES FINAIS

Em resumo, a análise do banco de dados revelou informações valiosas, permitindo validar tendências nos modelos de previsão, como a influência do fator campo e a escolha da distribuição de Poisson. Embora as métricas obtidas tenham sido satisfatórias, elas ainda não atingiram os patamares desejados para um modelo de previsão ideal, que na visão dos autores desse estudo significa apresentar uma taxa de acerto médio acima de 75%, além de uma baixa amplitude nas demais métricas observadas.

Os objetivos estabelecidos no início da pesquisa foram integralmente alcançados. Foi possível extrair com base nas estatísticas do banco de dados, métricas importantes sobre os modelos. E ainda, adquirir as previsões do modelo da *UFMG*, implementar os modelos propostos por Arruda (2000) e obter as probabilidades para cada confronto. Além de propor duas métricas para mensurar a precisão dos modelos.

O modelo *UFMG* exibiu uma característica desejável para modelos de predição, destacando-se pela estabilidade evidenciada em suas métricas, as quais demonstraram uma amplitude consideravelmente baixa em comparação com os demais modelos. Uma hipótese plausível para explicar esse padrão é a natureza do cálculo das previsões, que é fechado dentro de cada campeonato. Em outras palavras, de um campeonato para o outro, os valores dos vetores de previsões retornam ao valor inicial. Essa peculiaridade pode contribuir para a estabilidade observada nas métricas do modelo *UFMG*, proporcionando uma consistência notável em suas predições ao longo de diferentes campeonatos.

Apesar do modelo *chance I* utilizar um método de estimação mais popular e privilegiado na literatura, os resultados não foram tão satisfatórios quanto aos do modelo *SD 0*. Entretanto, considerando a sua Medida de DeFinetti média tanto em comparação com outros aplicações desse modelo, quanto em comparação com essa métrica aplicada aos resultados do modelo *SD 0*, observou-se resultados muito próximos.

Ao avaliarmos as métricas apresentadas, é evidente que o modelo *SD 0* se destaca, obtendo resultados superiores em todas as métricas calculadas. Essa constatação não apenas valida a eficácia do modelo *SD 0*, mas também aponta para possíveis áreas de melhoria nos outros modelos, incentivando a busca por refinamentos que possam elevar o desempenho geral das previsões. O trabalho desenvolvido estabeleceu uma base sólida, e as considerações feitas ao longo do estudo ofereceram direcionamentos valiosos para futuras investigações e otimizações nos modelos de previsão estatística, como o aprimoramento da calculo dos λ da distribuição de Poisson Bivariada e os múltiplos significados para de um empate como colocado pelo modelo UFMG.

Referências

ALMEIDA, V. **Brasil lidera crescimento de visitas a sites de apostas esportivas**. Similarweb Blog, 2023. Acesso em: 31 out 2023. Disponível em: ⟨https://www.similarweb.com/blog/pt/insights/brasil-lidera-crescimento-de-visitas-a-sites-de-apostas-esportivas/⟩.

ANDRADA, M. **Audiência global do esporte explode no mundo digital**. Poder360, 2022. Acesso em: 31 out 2023. Disponível em: ⟨https://www.poder360.com.br/opiniao/audiencia-global-do-esporte-explode-no-mundo-digital/⟩.

ARAÚJO, C. T. P. de; TAVARES, L.; ALVARES, L. G.; NETO, F. L.; SUZUKI, A. K. Modelagem estatística para a previsão de jogos de futebol: Uma aplicação no campeonato brasileiro de futebol 2014. **Revista da Estatística da Universidade Federal de Ouro Preto**, Ouro Preto, 2015.

ARRUDA, M. L. d. **Poisson, Bayes, Futebol e DeFinetti**. Tese (Doutorado) — Universidade de São Paulo, São Paulo, 2000.

ASTIVIA, O. L. O. Issues, problems and potential solutions when simulating continuous, non-normal data in the social sciences. **Meta-Psychology**, v. 4, 2020.

BARBANTI, V. O que é esporte? **Revista brasileira de atividade física & saúde**, v. 11, n. 1, p. 54–58, 2006.

CASELLA, G.; BERGER, R. L. **Statistical inference**. Autralia: Cengage Learning, 2021.

Confederação Brasileira de Futebol. **CAMPEONATO BRASILEIRO DA SERIE A TABELA DETALHADA**. 2023. Acesso em: 31 out 2023. Disponível em: ⟨https://conteudo.cbf.com.br/cdn/202212/20181204151220_289.pdf⟩.

DEFINETTI, B. **Probability, Induction and Statistics: The Art of Guessing**. New York: John Wiley, 1972.

DEGAM, P. C. **Distribuição de Poisson Bivariada aplicada a predições de resultados em partidas de futebol do Campeonato Rondoniense de 2019**. Ji-Paraná-RO: [s.n.], 2019. Trabalho de Conclusão de Curso (Bacharel em Estatística).

DRAPER, N. R.; SMITH, H. **Applied regression analysis**. [S.l.]: John Wiley & Sons, 1998. v. 3.

FILHO, C. A. O.; SUZUKI, A. K.; LOUZADA, F.; SARAIVA, E. F.; SALAZAR, L. E. B. Uma abordagem bayesiana para previsão de resultados de jogos de futebol: Uma aplicação ao campeonato inglês. **REVISTA BRASILEIRA DE BIOMETRIA**, v. 35, n. 1, p. 76–97, 2017.

H2 Gambling Capital. **Europe's Gambling Market Revenue (2019-2027E)**. EUROPEAN GAMBLING & BETTING ASSOCIATION, 2022. Acesso em: 31 out 2023. Disponível em: ⟨https://www.egba.eu/resource-post/gambling-market-revenue-in-europe-2019-2026e/⟩.

HORIZM. **The 100m club**: The global reach of sports leagues on socials big 4. 2023. Acesso em: 31 out 2023. Disponível em: ⟨https://horizm.com/reports/100m_Club_report.pdf⟩.

HYNDMAN, R. J.; ATHANASOPOULOS, G. **Forecasting: principles and practice**. [S.l.]: OTexts, 2018.

JUNIOR, O. G. de S.; GAMERMAN, D. Previsão de partidas de futebol usando modelos dinâmicos. In: **XXXVI Simpósio Brasileiro de Pesquisa Operacional**. São João del-Rei: [s.n.], 2004. p. 650 – 659.

LIMA, B. N. B. d.; COSTA, G. N.; NACIFE, R.; MARTINS, R. V.; GUIMARÃES, R. *et al.* Probabilidades no esporte. **TRIM: revista de investigación multidisciplinar**, n. 5, p. 39–53, 2012.

MAGALHÃES, M. N. **Probabilidade e variáveis aleatórias**. São Paulo, SP: Edusp, 2006.

MATOS, B. F. d. Inferência bayesiana aplicada à previsão de resultados do campeonato brasileiro de futebol. 2017.

MOOD, A.; GRAYBILL, F.; BOES, D. Introduction to the theory of statistics. **International Student Edition. McGraw-Hill**, 1974.

PALUDO, G. F.; FIGUEIREDO, N. N. de; FERREIRA, E. B. Proposta de uma métrica para o efeito de casa baseada em pontos ganhos. **Ciência e Natura**, v. 45, 2023.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2023. Disponível em: ⟨https://www.R-project.org/⟩.

RAMOS, L. F. P.; FERNANDES, H. C.; BATISTA, B. D. de O. Modelagem matemática para previsão de jogos de futebol. **ReviSeM**, v. 6, n. 1, p. 46–64, 2021.

ROSS, S. **A first course in probability**. Boston, MA: Pearson, 1976. v. 9.

SANTANA, H.; FERREIRA, P. H.; ARA, A.; NETO, F. L.; SUZUKI, A. K. Modelagem estatística e de aprendizado de máquina: previsão do campeonato brasileiro série a 2017. **Matemática e Estatística em Foco**, v. 7, n. 1, p. 42–66, 2020.

SOMOGGI, A. **Finanças TOP 20 clubes brasileiros em 2022: A consolidação do marketing e novas receitas**. SportsValue Blog, 2023. Acesso em 31 de outubro de 2023. Disponível em: ⟨https://www.sportsvalue.com.br/estudos/financas-clubes-brasileiros-em-2022-a-consolidacao-do-marketing/⟩.

SUZUKI, A. K.; SALASAR, L. E. B.; LEITE, J.; LOUZADA-NETO, F. A bayesian approach for predicting match outcomes: the 2006 (association) football world cup. **Journal of the Operational Research Society**, Taylor & Francis, v. 61, n. 10, p. 1530–1539, 2010.

TAVARES, L.; SUZUKI, A. K. Modelagem estatística para previsão esportiva: Uma aplicação no futebol. **Matemática e Estatística em Foco**, v. 3, n. 1, p. 32–47, 2015.

7 APÊNDICE

Tabela 5: Tabela Descritiva do número de gols por rodada feitos como mandante(M), visitante (V) e o total (T), e o valor total de gols por rodada em relação aos campeonatos analisados (TR).

	2017			2018			2019		
rod	M	V	T	M	V	T	M	V	T
1	26	7	33	20	7	27	25	8	33
2	10	10	20	9	7	16	15	11	26
3	15	11	26	8	9	17	13	11	24
4	12	3	15	19	11	30	11	7	18
5	17	12	29	18	13	31	17	9	26
6	19	11	30	15	3	18	12	9	21
7	11	7	18	12	12	24	17	4	21
8	10	17	27	17	7	24	11	5	16
9	18	9	27	18	13	31	13	8	21
10	9	10	19	6	9	15	15	7	22
11	14	7	21	19	12	31	12	10	22
12	12	14	26	13	5	18	10	11	21
13	7	20	27	14	7	21	19	10	29
14	17	12	29	18	7	25	18	13	31
15	14	8	22	13	5	18	10	8	18
16	10	12	22	19	11	30	14	11	25
17	9	14	23	7	6	13	12	4	16
18	16	16	32	14	9	23	10	16	26
19	12	6	18	11	9	20	10	8	18
20	12	6	18	9	7	16	11	10	21
21	12	6	18	16	5	21	23	10	33
22	15	10	25	11	8	19	11	6	17
23	13	10	23	13	5	18	8	10	18
24	13	8	21	10	3	13	15	5	20
25	10	12	22	14	5	19	9	7	16
26	8	11	19	18	14	32	12	13	25
27	14	13	27	17	9	26	14	10	24
28	12	10	22	14	12	26	16	9	25
29	16	11	27	17	2	19	13	14	27
30	15	12	27	13	12	25	17	8	25
31	12	11	23	18	11	29	12	11	23
32	19	9	28	12	10	22	9	11	20
33	15	15	30	11	9	20	9	5	14
34	15	7	22	11	5	16	17	16	33
35	21	10	31	10	5	15	15	7	22
36	17	4	21	19	4	23	13	11	24
37	11	14	25	10	8	18	22	5	27
38	18	12	30	11	6	17	15	12	27
T	526	397	923	524	302	826	525	350	875

	2020			2021			2022			
rod	M	V	T	M	V	T	M	V	T	TR
1	7	13	20	13	12	25	16	10	26	164
2	17	6	23	19	13	32	20	10	30	147
3	13	7	20	9	6	15	14	9	23	125
4	13	13	26	7	11	18	11	8	19	126
5	10	7	17	12	9	21	11	10	21	145
6	11	11	22	18	12	30	18	9	27	148
7	14	14	28	11	8	19	9	9	18	128
8	14	12	26	14	15	29	10	11	21	143
9	13	14	27	8	10	18	9	12	21	145
10	15	11	26	13	7	20	17	12	29	131
11	16	11	27	13	11	24	9	9	18	143
12	16	8	24	7	18	25	17	11	28	142
13	15	11	26	15	6	21	14	9	23	147
14	18	10	28	9	8	17	17	8	25	155
15	14	10	24	8	13	21	12	13	25	128
16	11	15	26	18	10	28	15	4	19	150
17	14	11	25	8	12	20	9	8	17	114
18	17	12	29	8	8	16	22	11	33	159
19	15	10	25	14	10	24	13	10	23	128
20	13	8	21	10	12	22	17	7	24	122
21	14	15	29	12	9	21	14	11	25	147
22	14	9	23	15	6	21	14	4	18	123
23	14	16	30	12	12	24	13	10	23	136
24	12	9	21	14	13	27	10	5	15	117
25	15	7	22	13	13	26	12	14	26	131
26	16	11	27	15	5	20	13	7	20	143
27	12	8	20	10	8	18	11	10	21	136
28	15	11	26	18	7	25	18	6	24	148
29	12	7	19	10	5	15	17	10	27	134
30	15	13	28	9	7	16	16	13	29	150
31	18	17	35	20	6	26	16	11	27	163
32	19	12	31	17	15	32	5	10	15	148
33	12	9	21	9	7	16	14	12	26	127
34	14	8	22	13	6	19	17	12	29	141
35	16	16	32	18	5	23	14	15	29	152
36	14	10	24	14	9	23	12	10	22	137
37	8	12	20	12	9	21	24	7	31	142
38	19	5	24	18	6	24	14	13	27	149
T	535	409	944	483	359	842	534	370	904	5314

Fonte: Dos autores.