Revista Eletrônica Matemática e Estatística em Foco
Volume 11 - Número 1
Outubro 2024

Artigo de Pesquisa

# Ciência de Dados no Futebol: Uma Análise Estatística das Cotações de Casas de Apostas *Online*

**Paulo Henrique Ferreira**
Universidade Federal da Bahia
paulohenri@ufba.br

**Gregori Ramos**
Universidade Federal da Bahia
gregori_sr@hotmail.com

**Resumo**

O futebol é um dos esportes mais populares do mundo e, por conta disso, não é incomum ouvir falar das estatísticas dessa modalidade esportiva; a cotação ou *odds* é uma delas. O presente trabalho tem como objetivo principal analisar e gerar *insights* das cotações de partidas de futebol em casas de apostas *online*. Dentre as perguntas respondidas neste trabalho, estão: Existe viés de vitória do time mandante? Como as casas de apostas lucram? Quão boas são as cotações das casas de apostas? Qual a relação da sabedoria das multidões com as apostas esportivas? Também será visto como falácias estatísticas podem surgir durante as análises esportivas e como as pessoas (torcedores, apostadores, patrocinadores etc.) estão expostas a eventos aleatórios e de grande impacto. Como exemplo, foi observado durante a pesquisa que se um apostador apostasse apenas no Ituano jogando em casa em 2022, teria lucrado mais do que todos os investimentos tradicionais de janeiro até novembro de 2022. Essa constatação ressalta a importância de analisar cuidadosamente as tendências e padrões dentro das apostas esportivas, além de demonstrar como o conhecimento estatístico pode ser aplicado de forma lucrativa.

**Palavras-chaves:** Análise exploratória e descritiva, Apostas esportivas, Brier Score, Falácias estatísticas, Futebol, *Odds*.

**Abstract**

Football is one of the most popular sports in the world. Because of this, it is not uncommon to hear about the statistics of this sports modality; the quotation or odds is one of them. The main objective of this work is to analyze and generate insights from the odds of football matches in online bookmakers. Among the questions answered in this work are: Is there a winning bias for the home team? How do bookmakers make money? How good are the odds from bookmakers? How is the wisdom of the crowd related to sports betting? It will also be seen how statistical fallacies can arise during sports analysis and how people (fans, bettors and sponsors, among others) are exposed to random and high-impact events. As an example, it was observed during the research that if a bettor only bet on Ituano playing at home in 2022, they would have profited more than all traditional investments from January to November 2022. This finding highlights the importance of carefully analyzing trends and patterns within sports betting, in addition to demonstrating how statistical knowledge can be applied profitably.

**Keywords:** Exploratory and descriptive analysis, Sports betting, Brier Score, Statistical fallacies, soccer, *Odds*.

## 1 Introdução

Nos últimos anos, as apostas esportivas ganharam destaque e começaram a aparecer com frequência no cenário do futebol brasileiro e mundial, desde comerciais de TV em horário nobre, patrocínio de clubes de futebol, programas de televisão e *sites*, até canais do Youtube que falam sobre esse tema.

Em 2022, todos os 20 clubes da Série A do Campeonato Brasileiro foram patrocinados por casas de apostas, incluindo patrocínio máster, que é o espaço publicitário mais caro do uniforme de um clube [1]. Algumas das casas de apostas *online* mais conhecidas são: PixBet, Betsson, Amuleto Bet, Betano, Blaze, NetBet, entre outras.

No ínicio, as apostas esportivas eram feitas em locais físicos, onde era possível fazê-las presencialmente, porém, devido ao crescimento dos *sites* de apostas e sua facilidade de depositar e sacar valores, cada dia mais as apostas *onlines* crescem no meio digital (os locais físicos ainda existem, só que em pequeno número). Por conta

[1] <https://exame.com/casual/todos-os-20-times-da-serie-a-tem-sites-de-apostas-esportivas-como-patrocinadores/>.

disso, surgiram vários *sites* de apostas esportivas na Internet; o mais famoso deles, o Bet365 [2], de acordo com o *site* BNLData [3], chegou a faturar mais de 3 bilhões de dólares em 2021.

Os dados são peças fundamentais no cenário das apostas esportivas, representando um conjunto de informações cruciais que impulsionam todo o ecossistema desse mercado. Quando se fala em dados, está se fazendo referência aqui a informações concretas e mensuráveis que são coletadas, armazenadas e analisadas para fornecer *insights* valiosos. Eles são a matéria-prima que alimenta as decisões dos apostadores, oferecendo uma visão mais clara e informada sobre os eventos esportivos em que desejam investir. Além de movimentar muito dinheiro, as apostas esportivas também movimentam muitos dados, visto que várias casas de apostas atuam há mais de 10 anos, oferecendo cotações para diversos esportes, não só o futebol. As cotações de um evento em uma casa de apostas esportivas são similares às cotações de uma empresa na Bolsa de Valores. Um exemplo simples para entendimento: a casa de apostas precifica que o Vasco da Gama tem 50% de "chances" de vencer determinada partida; neste caso, a cotação do clube seria R$ 2,00, ou seja, o apostador dobraria o valor investido em caso de vitória do Vasco da Gama, caso contrário, perderia todo o valor. Essa é só uma das centenas de possibilidades de apostas que o usuário pode fazer. Por exemplo, pode-se apostar em quantidade de gols, escanteios, laterais, tiros de meta, faltas, cartões, placar exato da partida etc.

Tendo em vista os milhares de dados gerados para esses eventos esportivos, a Ciência de Dados, assim como a Estatística, têm um oceano gigantesco de dados para realizar análises tanto exploratórias quanto preditivas. Por esse motivo, já existem diversos trabalhos na literatura que utilizam procedimentos matemáticos, estatísticos ou computacionais para prever resultados de partidas de futebol; alguns deles utilizam modelos Bayesianos dinâmicos com coeficientes autorregressivos de evolução, regressão trinomial, modelos logísticos ordinais ou Poisson autorregressivo (ver, por exemplo, (SANTANA et al., 2020) e as referências ali mencionadas).

Nas próximas seções são apresentadas algumas análises descritivas e exploratórias desses dados, buscando gerar valor sobre esse mundo de dados relacionados a partidas de futebol.

## 2 Material e Métodos

O presente trabalho faz uso de uma base de dados de mais de 1,1 milhão de registros de partidas de futebol do mundo todo, de diversos campeonatos, tanto de clubes como de seleções, e datados de 2004 até 2022. Os dados foram coletados através de rastreador *web* (*web crawler*).

Um rastreador *web* é um programa ou *script* que, metodicamente, analisa e percorre páginas *web* para criar um índice dos dados de interesse. É considerado um agente de *software* que utiliza uma lista de URLs a serem navegadas. A partir dessa navegação, o rastreador *web* identifica todos os *links* das páginas e adiciona-os na lista de URLs que serão visitadas (DHENAKARAN e SAMBANTHAN, 2011).

O *site* escolhido foi o Odds Portal [4] e o rastreador *web* foi desenvolvido pelo primeiro autor deste trabalho, utilizando a linguagem de programação C# e a biblioteca `Selenium`, muito conhecida para testes de interface de aplicações. Os dados foram armazenados em um banco de dados SQL Server e analisados utilizando Python e suas bibliotecas mais populares, como `Pandas`, `Matplotlib` e `Numpy`.

A Figura 1 apresenta o fluxo desenvolvido para obtenção/coleta dos dados, assim como as tecnologias usadas para este fim.


**Figura 1:** Fluxo da coleta dos dados.
Fonte: Próprio autor (2024).

[2] <www.bet365.com>.
[3] <https://bnldata.com.br>.
[4] <www.oddsportal.com>.

### Dicionário dos Dados

O dicionário de dados da Tabela 1 foi utilizado durante todo o trabalho para se referir, de forma resumida, aos campos da base de dados empregada. No que tange aos campos OM, OE e OV dessa tabela, o rastreador *web* desenvolvido/utilizado não busca as cotações de uma única casa de apostas. Então, considera-se esses valores como sendo de mais de uma casa de apostas (médias). Na Tabela 2, tem-se as estatísticas descritivas dos dados utilizados neste trabalho.

**Tabela 1:** Dicionário dos dados.

| Nome | Significado |
|---|---|
| DT - Data | Data em que ocorreu o evento |
| TM - Time Mandante | Nome do time mandante |
| TV - Time Visitante | Nome do time visitante |
| GM - Gols Mandante | Quantidade de gols do time mandante |
| GV - Gols Visitante | Quantidade de gols do time visitante |
| OM - *Odds* Mandante | Cotação do time mandante |
| OE - *Odds* Empate | Cotação do empate |
| OV - *Odds* Visitante | Cotação do time visitante |
| CAMP - Campeonato | Nome do campeonato |
| P - País | País em que ocorreu o evento |
| C - CCASC | Comissão da casa de apostas sobre as cotações |
| EB - Escore Brier | EB calculado |

**Tabela 2:** Resumo descritivo das variáveis quantitativas envolvidas no estudo.

| Estatística | GM | GV | OM | OE | OV | CCASC | EB |
|---|---|---|---|---|---|---|---|
| $n$ | 1.145.502 | 1.145.502 | 1.145.502 | 1.145.502 | 1.145.502 | 1.145.502 | 1.145.502 |
| Média | 1,55 | 1,23 | 2,69 | 3,87 | 4,11 | 9,25 | 0,57 |
| Desvio-padrão | 1,38 | 1,24 | 2,87 | 2,10 | 3,80 | 2,30 | 0,30 |
| Mínimo | 0 | 0 | 1,01 | 1,01 | 1,01 | -78,18 | 0 |
| $1^o$ Quartil | 1,0 | 0 | 1,67 | 3,22 | 2,37 | 7,83 | 0,32 |
| Mediana | 1,0 | 1,0 | 2,13 | 3,49 | 3,19 | 9,33 | 0,58 |
| $3^o$ Quartil | 2,0 | 2,0 | 2,77 | 3,97 | 4,63 | 10,99 | 0,80 |
| Máximo | 22,0 | 21,0 | 509,25 | 503,38 | 515,0 | 197,03 | 1,86 |

### *Return on Investment* (ROI)

Na área financeira, o retorno sobre o investimento (em inglês, *return on investment* ou ROI), também chamado de taxa de retorno, taxa de lucro ou, simplesmente, retorno, é a relação entre a quantidade de dinheiro ganho (ou perdido) como resultado de um investimento e a quantidade de dinheiro investido (OLDCORN e PARKER, 1998).

A fórmula para o cálculo do ROI é:

$$\frac{\text{Lucro Líquido}}{\text{Total Investido}} \times 100\%.$$

Adaptando isso às apostas esportivas, também pode-se calcular o ROI de uma aposta feita, ou até mesmo de um clube, com base no desempenho em uma determinada temporada ou período. Por exemplo, se as casas de apostas atribuem uma cotação de R$ 1,90 para o Vasco da Gama vencer certo confronto, a fórmula para calcular o ROI dessa aposta seria a seguinte:

- Caso o Vasco da Gama vencesse a partida, o apostador teria 90% de ROI sobre a aposta realizada:

$$(1{,}90/1 - 1) \times 100\% = 90\%;$$

- Em caso de derrota do Vasco da Gama ou empate, o apostador perderia todo o valor apostado, logo, teria -100% de lucro sobre o investimento:

$$(0/1 - 1) \times 100\% = -100\%.$$

**Comissão da Casa de Apostas Sobre as Cotações**

CCASC, também conhecida como House Edge ou Juice, é uma expressão usada para descrever a vantagem matemática que a casa de apostas, assim como os locais de jogos comerciais, têm sobre o apostador ao longo do tempo. Esse benefício proporciona um retorno percentual garantido para a casa de apostas ao longo do tempo e uma perda percentual garantida do valor investido. Existe uma frase célebre - “The house always wins” - para dizer que a casa de apostas ou cassino sempre vence, e isso é fato, justamente por causa da CCASC.

Em todos os eventos que uma casa de apostas publica para o apostador fazer sua aposta, está previamente calculado o quanto ela vai receber daquela operação, independente do resultado da mesma. Observando as seguintes linhas, consegue-se facilmente saber exatamente quanto a casa de apostas ganhará em determinado evento.

Por exemplo, o Vasco da Gama jogará contra o Flamengo e a casa de apostas publica o evento com as seguintes cotações:

- Vitória do Vasco da Gama: R$ 2,71;
- Empate: R$ 3,27;
- Vitória do Flamengo: R$ 2,46.

Observa-se que a soma das probabilidades ultrapassa 100%:

$$(1/2{,}71 + 1/3{,}27 + 1/2{,}46) \times 100\% \approx 108{,}13\%.$$

Ou seja, a casa de apostas tem uma CCASC de pouco mais de 8%. Ela controla as cotações pelos valores investidos em cada linha, logo, no total aportado por todos os apostadores, independente de quem vencer, a casa de apostas lucra.

A subseção seguinte trata do EB, e para sua correta medida, deve-se adicionar nas cotações a CCASC aplicada pelas casas de apostas. Dessa forma, utilizando o exemplo anterior, é preciso adicionar 8,13% nas cotações do mandante, do empate e do visitante.

Então, as novas cotações ajustadas seriam:

- Vitória do Vasco da Gama: R$ 2,93;
- Empate: R$ 3,53;
- Vitória do Flamengo: R$ 2,66.

Observa-se que, agora, a soma das probabilidades não ultrapassa 100%:

$$(1/2{,}93 + 1/3{,}53 + 1/2{,}66) \times 100\% \approx 100\%.$$

Feito isso, considera-se quaisquer resultados de EB deste trabalho, já com as cotações ajustadas pela CCASC.

**Escore Brier**

Este trabalho tem como objetivo responder algumas perguntas sobre as cotações das partidas de futebol, como:

- Quão boas são as cotações das casas de apostas?

Para a análise dessa qualidade, foi utilizada a medida EB.

Proposto por (BRIER, 1950), o EB é uma função (ou regra) de pontuação adequada, que mede a precisão das previsões probabilísticas. Para previsões unidimensionais, que são aquelas que consideram apenas uma variável ou dimensão para a previsão de um determinado evento ou fenômeno, é equivalente ao erro quadrático médio aplicado às probabilidades previstas. Quando aplicado às previsões de várias classes, o EB é definido como (ver, por exemplo, (SANTANA et al., 2020)):

$$\text{EB} = \frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{l}\left(p_{ti} - o_{ti}\right)^2,$$

em que:

- $l$ é o número de classes (aqui, $l = 3$, visto que são três as classificações possíveis para uma partida - Vitória do Mandante, Empate e Vitória do Visitante);
- $n$ é o número de instâncias (partidas);
- $p_{ti}$ é a probabilidade prevista da $t$-ésima instância pertencer à $i$-ésima classe;
- e $o_{ti}$ é 1 se a classe atual $y_t$ é igual a $i$, ou 0 se a classe $y_t$ é diferente de $i$.

Para esta pesquisa, foi utilizado o valor de referência de 2/3 ($\approx$ 0,67), por conta das três classificações possíveis; sendo 0 e 2, respectivamente, os valores mínimo e máximo da medida neste contexto. Quanto menor (mais próximo de 0) o valor do EB, melhor é a qualidade das previsões.

O EB foi calculado a partir de todas as partidas da base de dados, utilizando a seguinte operação (individual):

- Para cada partida que a equipe mandante venceu, sua contribuição ao EB foi dada por:

$$\left(1/\text{OM} - 1\right)^2 + \left(1/\text{OE} - 0\right)^2 + \left(1/\text{OV} - 0\right)^2;$$

- Para cada partida que o empate ocorreu:

$$\left(1/\text{OM} - 0\right)^2 + \left(1/\text{OE} - 1\right)^2 + \left(1/\text{OV} - 0\right)^2;$$

- Por fim, para cada partida que a equipe visitante levou a melhor:

$$\left(1/\text{OM} - 0\right)^2 + \left(1/\text{OE} - 0\right)^2 + \left(1/\text{OV} - 1\right)^2.$$

## 3 Resultados

Nesta seção são apresentados alguns resultados da aplicação dos métodos descritos anteriormente.

### O Viés de Vitória do Mandante

No futebol, muito se fala sobre o fator casa (ou mando de campo), segundo o qual o time, quando joga a partida frente à sua torcida, tem uma vantagem por receber mais apoio. Além disso, em um país de dimensões continentais como o Brasil, a própria viagem para locais distantes pode dar ainda mais vantagem para o time mandante.

Outros fatores como comportamento da torcida local, sua influência sobre visitantes e árbitros, diferentes condições no campo de jogo e, principalmente, os transtornos advindos das distâncias percorridas e diferenças climáticas regionais, parecem ser bons caminhos para ampliar a pesquisa no futebol brasileiro, na tentativa de explicar essa maior taxa de vitórias do mandante (SILVA e MOREIRA, 2008).

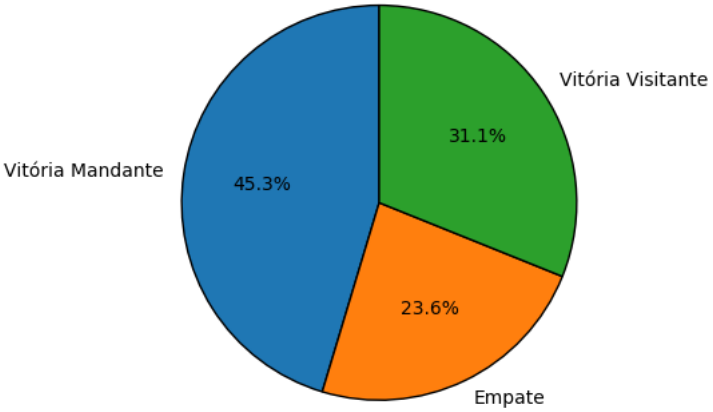


**Figura 2:** Porcentagem de partidas que terminaram com vitória do time mandante, empatadas ou com vitória do time visitante.

Fonte: Próprio autor (2024).

Conforme mostra a Figura 2, os dados coletados confirmam a ideia de que o time mandante realmente vence mais jogos do que o time visitante, ou até mesmo o empate. Isso mostra a parte subjetiva do futebol, em que o apoio da torcida pode sim impactar no resultado de campo. Então, para uma futura análise preditiva, apenas os fatores objetivos não seriam suficientes para estimar um vencedor do confronto, tendo assim que medir o apoio que essa torcida tem no determinado jogo.

A Figura 2, ainda que explicativa, considera todos os jogos da base de dados. Para analisar de forma mais aprofundada, é preciso buscar os jogos em que o time mandante e o time visitante eram favoritos em suas respectivas partidas, ou seja, OM < OV quando o mandante era favorito e OV < OM quando o visitante era favorito.

De acordo com a Figura 3, analisando agora os jogos em que as equipes eram favoritas nos seus confrontos, ainda assim, a equipe mandante abre mais vantagem, enquanto que a equipe visitante praticamente não sofre alteração. Em resumo, a equipe mandante aumentou cerca de 25% de vitórias quando era favorita no confronto (OM $<$ OV). Então, de forma geral, a equipe mandante tem 45,3% de vitórias, ou transformando em cotação, uma *odds* de R$ 2,21; porém, quando favorita, essa vantagem sobe ainda mais, indo para 70,9% de vitórias, percentual este que, transformado em cotação, fica R$ 1,41. Ou seja, é observada uma diferença de + R$ 0,80.

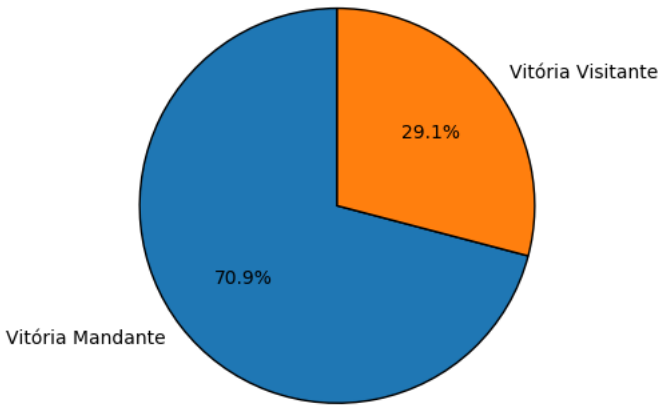


**Figura 3:** Porcentagem de partidas que terminaram com vitória do time mandante ou com vitória do time visitante, quando estes eram favoritos.

Fonte: Próprio autor (2024).

### Apostas Esportivas e Investimentos Tradicionais

O ROI, baseado nas cotações dos jogos, é um indicador interessante para saber a lucratividade que um determinado time teve na temporada. A Figura 4 apresenta os 3 times mandantes mais lucrativos e os 3 times mandantes menos lucrativos do Brasil em 2022, considerando apenas aqueles que tiveram 20 ou mais partidas como mandante neste ano. Por sua vez, a Tabela 3 mostra a rentabilidade dos investimentos tradicionais em 2022.

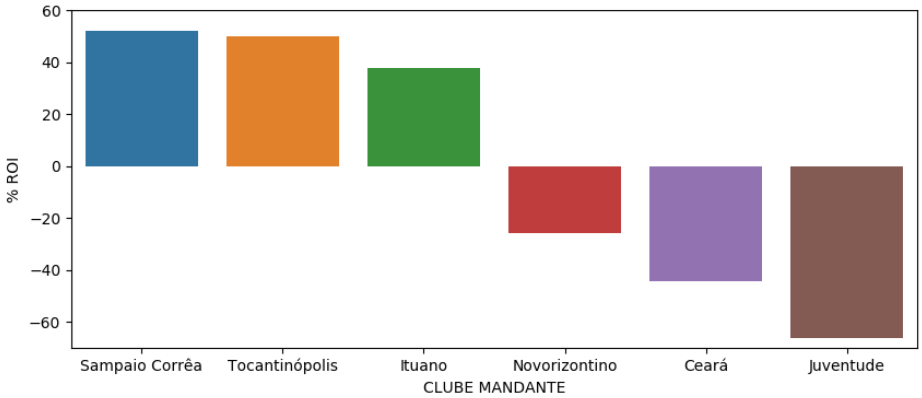


**Figura 4:** Times mandantes mais e menos lucrativos do Brasil em 2022.
Fonte: Próprio autor (2024).

Analisando essa figura e essa tabela, pode-se concluir que se um apostador apostasse apenas no Ituano jogando em casa em 2022, teria lucrado mais do que todos os investimentos tradicionais de janeiro até novembro de 2022.

**Tabela 3:** Rentabilidade de investimentos em 2022.

| Investimento | Rentabilidade | Investimento | Rentabilidade |
|---|---|---|---|
| CDB de banco médio | 8,26% | Tesouro IPCA | 0,48% |
| Tesouro Selic | 7,16% | SMLL | -8,81% |
| CDB de banco grande | 5,50% | Dólar | -10,32% |
| Poupança | 4,95% | Ouro | -12,45% |
| Divi | 3,97% | IVVB11 | -20,01% |
| Tesouro prefixado | 3,79% | BDRX | -21,68% |
| Ibovespa | 3,65% | Bitcoin | -54,32% |
| IFIX | 0,75% | Ethereum | -55,48% |

Fonte: https://www.suno.com.br/noticias/cdb-investimento-mais-rentavel-2022/

### Como as Casas de Apostas Lucram

Como explicado anteriormente (Seção 2.3), a CCASC é a margem de lucro das casas de apostas. Mas, além disso, ela funciona também como uma segurança em jogos que as casas de apostas não têm muita informação para precificar. Sendo assim, elas incorporam uma CCASC mais alta, a fim de se proteger de possíveis erros de precificação.

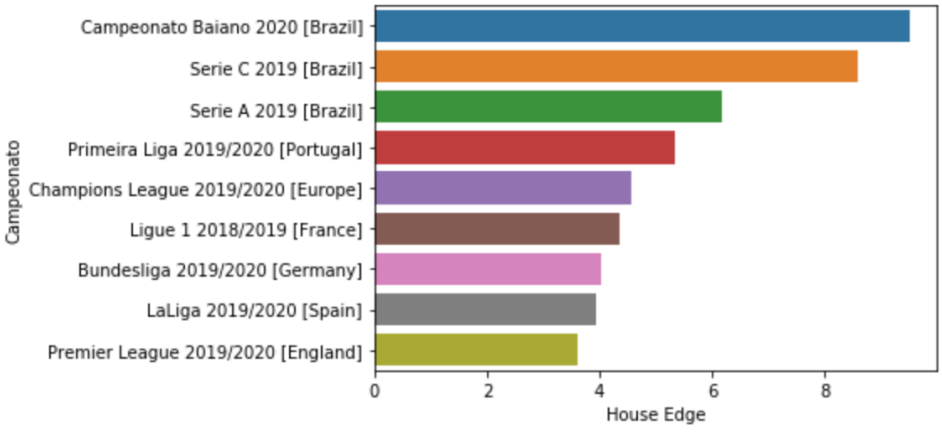


**Figura 5:** Porcentagem da CCASC e a sabedoria das casas de apostas, nos campeonatos de 2019 a 2020.
Fonte: Próprio autor (2024).

Na Figura 5, pode-se comprovar que em campeonatos com pouca liquidez, como o Campeonato Baiano, as casas de apostas se previnem adicionando uma alta comissão nessas linhas. Essa comissão pode aumentar ou diminuir; tudo vai depender da entrada de dinheiro dos apostadores ou do ganho de maiores informações sobre as partidas. Já em campeonatos mais conhecidos, como a Premier League, da Inglaterra, o volume de dinheiro nesses jogos é tão alto que as casas de apostas podem colocar uma comissão menor, pois já possuem informações suficientes sobre os eventos.

### Escore Brier e a Qualidade das Cotações

Nesta seção foi avaliada a qualidade das cotações por meio do EB calculado das partidas. Observou-se a qualidade das cotações durante os anos, assim como a sua qualidade por: Vitória do Mandante, Empate e Vitória do Visitante, entre outros pontos.

Na Figura 6, pode-se observar que as casas de apostas vêm evoluindo sua precificação ao longo do tempo, como mostrado na queda do EB durante os anos. Essa evolução vem do grande número de dados sendo gerados ano após ano, o que faz com que cada vez mais sejam realizadas análises mais apuradas sobre o desempenho de um time em uma partida de futebol.

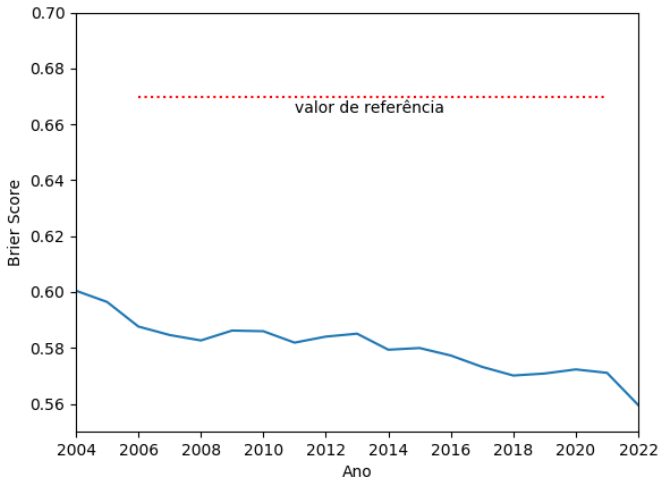


**Figura 6:** Qualidade das cotações ao longo dos anos.

Fonte: Próprio autor (2024).

Na Figura 7, pode-se ver que as casas de apostas têm uma certa dificuldade em atribuir as cotações dos empates, visto que a cotação desse evento (empate) atingiu um EB de 0,85, diferentemente do EB do mandante e do visitante, que foi de 0,40 e 0,60, respectivamente. A qualidade da cotação do empate, inclusive, ultrapassou o valor de referência de 2/3, que é o valor máximo aceitável que este trabalho considerou.

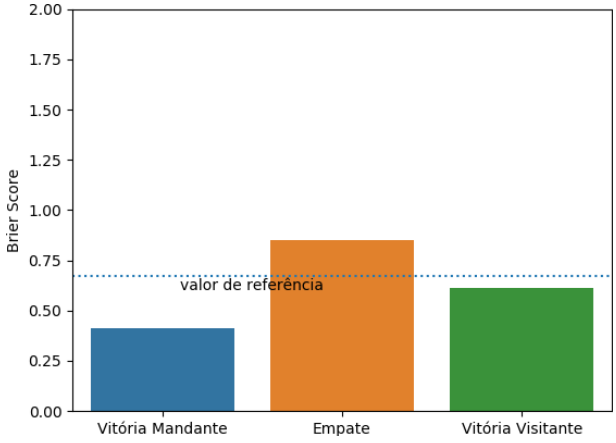


**Figura 7:** Qualidade das cotações por: Vitória do Mandante, Empate e Vitória do Visitante.

Fonte: Próprio autor (2024).

### Escore Brier e a Sabedoria das Multidões

É sabido que o EB informa a qualidade das cotações geradas pelas casas de apostas, mas apenas as casas não saberiam definir tão bem as linhas de um evento se não fossem os próprios apostadores, os principais atores que definem as oscilações dessas linhas. No livro "A Sabedoria das Multidões" (SUROWIECKI, 2005), o autor mostra que "as melhores decisões coletivas são produtos de desacordos e contendas, e não de consenso e compromisso". Em resumo, a análise de um grande grupo de pessoas pode ser, e muitas vezes é, mais relevante do que a opinião de uma só, mesmo essa pessoa sendo um especialista.

A Figura 8 mostra uma seleção de dados, em que as 5 primeiras ligas (Alemanha, Brasil, Espanha, Itália e Inglaterra) são as ligas com melhor qualidade de previsão em termos de EB, enquanto que as 4 últimas ligas (Gâmbia, Líbia, Ilhas Maurício e Laos) são as que apresentam a pior qualidade de EB. Isso demonstra o quanto é significativa a sabedoria das multidões nas principais ligas do planeta. Quanto mais pessoas apostando, mais a precificação desses jogos se ajusta; já nos países africanos, as casas de apostas ainda têm dificuldade de precificar, muito por conta de poucos apostadores colocando dinheiro nesses jogos.

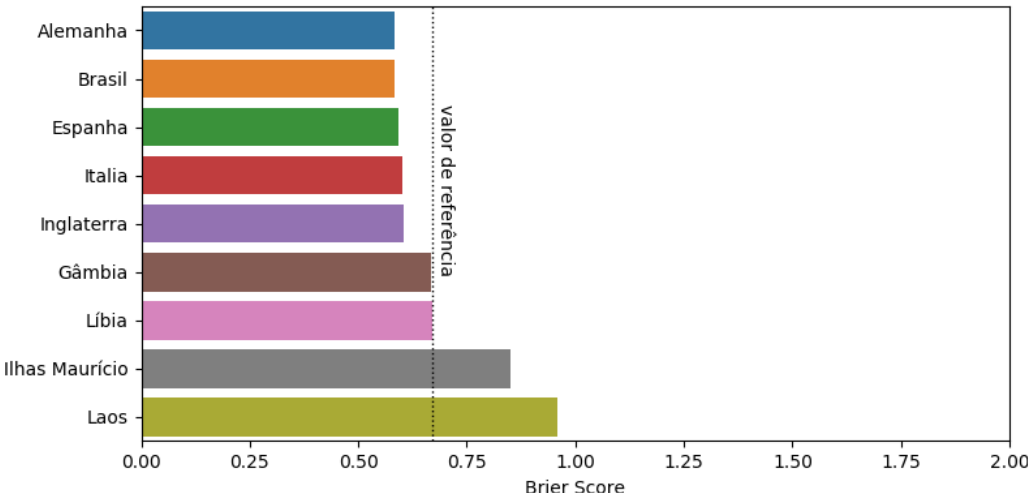


**Figura 8:** EB nas maiores e menores ligas de futebol do mundo.

Fonte: Próprio autor (2024).

### Falácia da Falsa Causalidade

Em uma amostra de 1.000 partidas, escolhidas de forma aleatória, foi observada uma forte correlação linear negativa entre a cotação do mandante e a cotação do visitante. Contudo, antes de entrar nesta análise em si, deve-se deixar claro que essa correlação não implica em causalidade. A correlação é a relação estatística entre duas variáveis, mas nem sempre correlação implica causalidade.

Pode-se utilizar o seguinte exemplo: apesar do galo sempre cantar antes do amanhecer, não é verdade que amanhece por causa do canto do galo. Isso é uma falácia lógica chamada "*Post hoc ergo propter hoc*" ("depois disso, logo, por causa disso") [5].

Para medir essa correlação em números, foi utilizado o coeficente de correlação linear de Pearson ($r$), que é uma medida estatística de quão próximos os dados estão da linha de regressão ajustada.

Na Figura 9, é observado que a cotação do mandante possui forte correlação linear negativa com a cotação do visitante ($r = -0{,}82$). Isso pode ser um indicativo de que, quanto mais favorito o mandante, menos favorito é o visitante, e vice-versa.

Conforme mostra a Figura 10, o mesmo não pode ser dito sobre a correlação do EB com a CCASC ($r = -0{,}05$), indicando que não existe associação entre a margem atribuída antes da partida começar, com a qualidade dessas cotações.

[5] <https://gec.proec.ufabc.edu.br/o-que-que-a-ciencia-tem/qual-a-diferenca-entre-correlacao-e-causalidade/>.

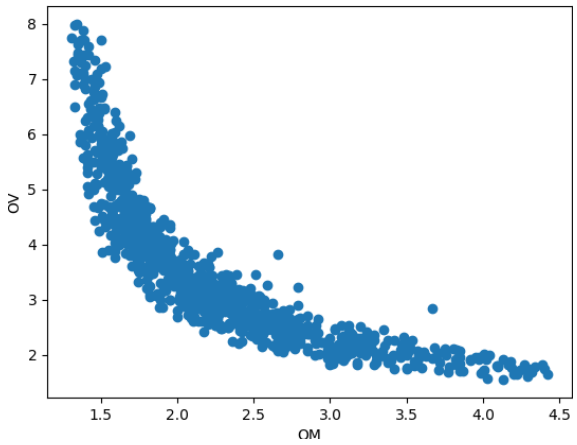


**Figura 9:** Gráfico de dispersão entre OM e OV.

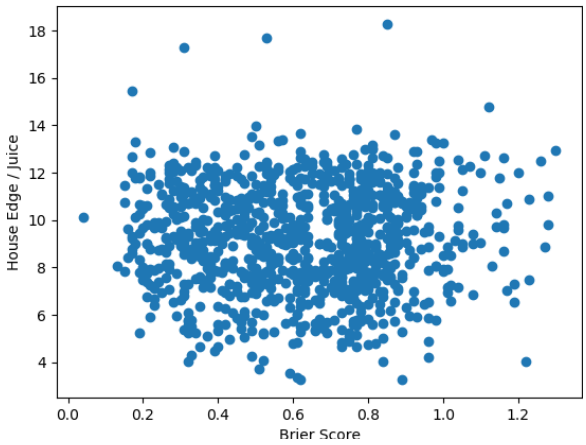


**Figura 10:** Gráfico de dispersão entre Brier Score e House Edge/Juice.

### Falácia do Apostador

Este é um equívoco de que, se algo acontece com mais frequência do que o normal, é menos provável que aconteça agora no futuro e vice-versa. Isso também é conhecido como falácia de Monte Carlo, devido a um exemplo que ocorreu em uma mesa de roleta em 1913. A bola caiu na área preta 26 vezes seguidas e o jogador perdeu milhões de apostas na suposição de que a sequência continuaria. No entanto, não importa o que aconteceu no passado, a probabilidade da parte preta é sempre a mesma da parte vermelha, porque a probabilidade subjacente permanece a mesma. Mais importante ainda, a mesa de roleta não tem memória [6].

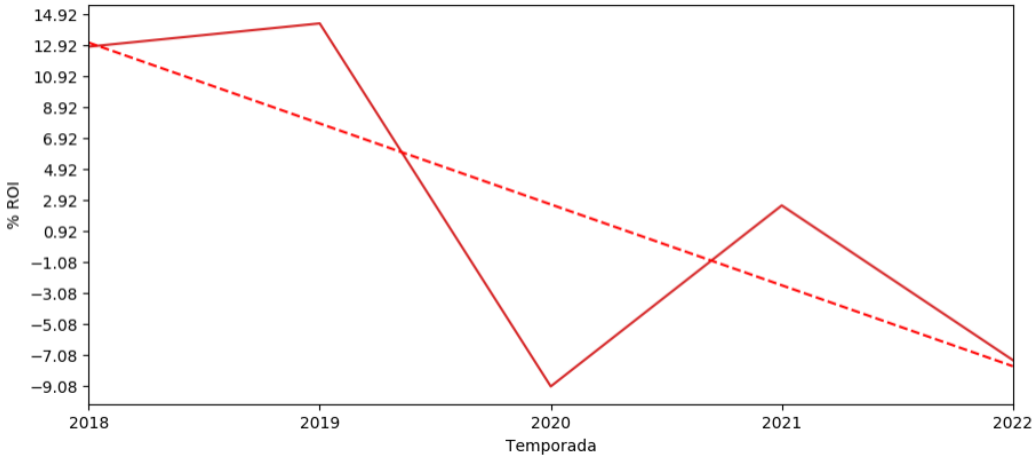


**Figura 11:** Porcentagem de ROI do Flamengo nas últimas 5 temporadas (tendência em linha tracejada).
Fonte: Próprio autor (2024).

Conforme pode ser visto na Figura 11, o Flamengo apresenta uma tendência de queda no seu ROI. Em 2019, quem apostou em todos os jogos do time teve um retorno de 14,33%; porém, logo no ano seguinte, o ROI caiu para -9,08%, o que confirma que não se deve iludir pela falácia do apostador e que a rentabilidade passada não garante rentabilidade futura.

### Lógica do Cisne Negro: O Impacto do Altamente Improvável

Primeiro, o cisne negro é um *outlier*, pois está fora do âmbito das expectativas comuns, já que nada no passado pode apontar convincentemente para a sua possibilidade. Segundo, ele exerce um impacto extremo. Terceiro, apesar de ser um *outlier*, a natureza humana faz com que sejam desenvolvidas explicações para a sua ocorrência após o evento, tornando-o explicável e previsível. Resumidamente, tem-se o terceto: raridade, impacto extremo e previsibilidade retrospectiva (mas não prospectiva) (Tabeb (TALEB, 2015)).

[6] <https://ibpad.com.br/ciencia-dados/15-falacias-estatisticas-que-voce-deve-evitar/>.

Assim, foi observado o impacto desses eventos improváveis no Campeonato Brasileiro da Série A de 2022 (ou Brasileirão 2022). Lembrando que o cisne negro é apenas uma maneira de apresentar o conceito; os eventos apresentados a seguir não foram considerados cisnes negros pelo fato das cotações por si só já serem uma maneira de explicar sua probabilidade e, como visto anteriormente, o cisne negro não pode ser explicado e medido antes do evento ocorrer.

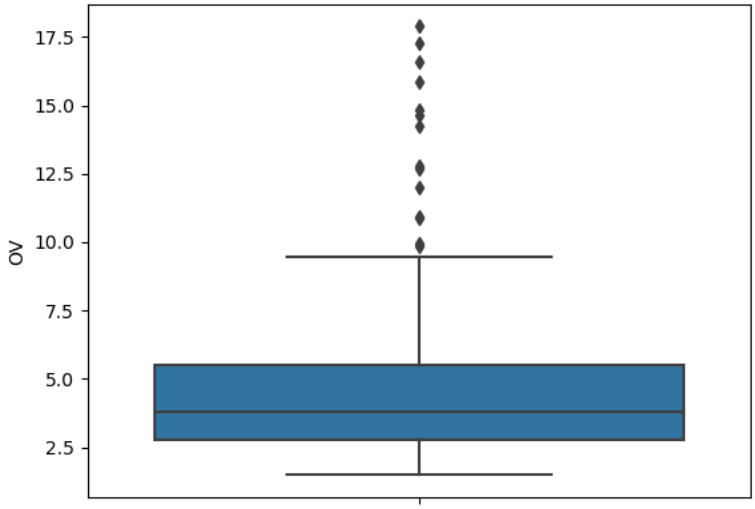


**Figura 12:** *Boxplot* das cotações do time visitante (OV) no Brasileirão 2022.
Fonte: Próprio autor (2024).

A Figura 12 mostra um diagrama de caixa (*boxplot*), de forma que pode-se identificar os visitantes *outliers* e entender os impactos desses eventos. Sendo assim, *outliers* são as partidas em que o visitante tem cotação maior do que 10 (OV $>$ 10), e não *outliers* são os demais jogos (OV $<$ 10).

**Tabela 4:** Partidas com visitantes *outliers* no Brasileirão 2022.

| TM | TV | OM | OV | GM | GV |
|---|---|---|---|---|---|
| Flamengo | Avaí | 1,17 | 15,84 | 1 | 2 |
| Atlético-MG | Juventude | 1,20 | 14,26 | 1 | 0 |
| Palmeiras | Avaí | 1,16 | 17,89 | 3 | 0 |
| Palmeiras | Coritiba | 1,23 | 12,00 | 4 | 0 |
| Palmeiras | Juventude | 1,18 | 17,28 | 2 | 1 |
| Atlético-MG | Goiás | 1,23 | 12,78 | 0 | 1 |
| Palmeiras | Goiás | 1,26 | 10,94 | 3 | 0 |
| Palmeiras | Cuiabá | 1,25 | 12,69 | 1 | 0 |
| Atlético-MG | Avaí | 1,21 | 14,63 | 2 | 1 |
| Flamengo | Goiás | 1,17 | 16,61 | 1 | 0 |
| Atlético-MG | Atlético-GO | 1,26 | 10,90 | 2 | 0 |
| Atlético-MG | Coritiba | 1,24 | 14,81 | 2 | 2 |

Ao observar a Tabela 4, verifica-se que das 12 partidas em questão, 2 foram vencidas pelos *outliers*. Em uma situação hipotética em que o investidor apostou R$ 1,00 em todos os *outliers*, foram gastos R$ 12,00 e obteve-se um lucro de R$ 28,62, o que equivale a um ROI de 138%.

Dessa forma, obteve-se um resultado assimétrico, ou seja, quando o resultado de um evento tem duas possibilidades, ocorrerá um resultado assimétrico se a vantagem (potencial de ganho ou retorno) é muito maior do que a desvantagem.

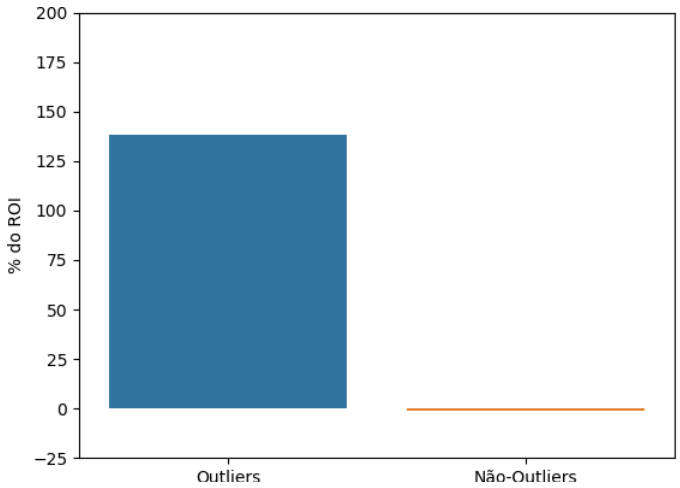


**Figura 13:** Lucro obtido em apostas nos *outliers* e nos outros demais visitantes (não *outliers*).
Fonte: Próprio autor (2024).

A Figura 13 mostra o ROI caso o investidor apostasse apenas nos *outliers* (12 partidas) ou nos demais visitantes (368 partidas) no Brasileirão 2022. Ficou claro como os resultados assimétricos podem trazer retornos muito mais significativos: os *outliers* deram um retorno de 138% e os demais visitantes tiveram um prejuízo de -1,14%.

"Não é a probabilidade de um evento acontecer que importa. O que deve ser considerado é o valor gerado quando o evento acontecer. A frequência do lucro é irrelevante; é a magnitude do resultado que conta." (TALEB, 2019)

## 4 Considerações Finais

Neste trabalho, observou-se que, no futebol, o viés de vitória do mandante existe, seja por apoio da torcida, ou questões geográficas. Ou seja, os times mandantes têm um favoritismo e isso se reflete nas *odds*. Foi visto também que, apesar de assertivas, as casas de apostas *online* ainda têm algumas falhas de precificação, ainda mais quando se refere à cotação do empate, conforme visto pelo seu EB elevado. O grande motivo para elas não perderem dinheiro nessas operações é por conta da CCASC, que garante uma comissão para a casa de apostas em todas as linhas que a mesma publica.

A CCASC também é utilizada como um certo regulador dos preços: quando a casa de apostas ainda não tem certeza sobre uma precificação ou as cotações foram lançadas há pouco tempo e não têm liquidez suficiente, a mesma coloca uma CCASC um pouco mais alta para se proteger de possíveis erros e, após a entrada de dinheiro feita pelos apostadores nas linhas, a casa de apostas consegue ir abaixando a sua comissão gradativamente, mas, obviamente, ganhando mais por ter mais apostadores apostando naquele evento. Além disso, é possível concluir que, no longo prazo, as casas de apostas irão cada vez mais melhorar a sua precificação e, cada vez mais, terão comissões menores e um EB menor também, visto que tal medida vem diminuindo, mesmo que pouco, ao longo dos anos, conforme mostrado na Figura 6. Essa melhora está relacionada tanto com os modelos cada vez mais assertivos das casas de apostas, como também com a sabedoria das multidões.

Observou-se, ainda, como as apostas esportivas, se bem analisadas e exploradas, podem ser um campo lucrativo, até superior a investimentos tradicionais. Algumas falácias estatísticas podem se apresentar e o apostador deve ficar ciente delas para não se deixar enganar por falsas previsões e análises. As apostas esportivas, assim como outros eventos estatísticos, também lidam com a incerteza; por mais que o indivíduo tente prever o futuro, sempre haverá uma dose de incerteza (ou aleatoriedade). Dessa forma, pode-se ficar exposto a cisnes negros.

Um dos próximos passos possíveis, utilizando-se das análises exploratórias e descritivas deste trabalho, seria o desenvolvimento de modelos preditivos considerando as cotações (*odds*) como variáveis preditoras, como visto, por exemplo, em (ODACHOWSKI e GREKOW, 2013).

## Referências

BRIER, G. W. (1950). **Verification of forecasts expressed in terms of probability**. Em: *Theoretical Economics Letters* 78.1, pp. 1–3.

DHENAKARAN, S. S. e SAMBANTHAN, K. Thirugnana (2011). **Web Crawler - An Overview**. Em: *International Journal of Computer Science and Communication* 2.1, pp. 265–267.

ODACHOWSKI, Karol e GREKOW, Jacek (2013). **Using bookmaker odds to predict the final result of football matches**. Em: *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, pp. 196–205.

OLDCORN, Roger e PARKER, David (1998). **Decisão Estratégica para Investidores**. São Paulo: Nobel.

SANTANA, Hugo et al. (2020). **Modelagem Estatística e de Aprendizado de Máquina: Previsão do Campeonato Brasileiro Série A 2017**. Em: *Matemática e Estatística em Foco* 7.1, pp. 42–66.

SILVA, C. D. e MOREIRA, D. G. (2008). **A vantagem em casa no futebol: comparação entre o Campeonato Brasileiro e as principais ligas nacionais do mundo**. Em: *Revista Brasileira de Cineantropometria & Desempenho Humano* 10.8, pp. 184–188.

SUROWIECKI, James (2005). **The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations**. New York: Anchor.

TALEB, Nassim N. (2015). **A lógica do Cisne Negro: O impacto do altamente improvável.** 9ª ed. Rio de Janeiro: Best Seller.

TALEB, Nassim N. (2019). **Iludidos pelo acaso: A influência da sorte nos mercados e na vida.** Objetiva.