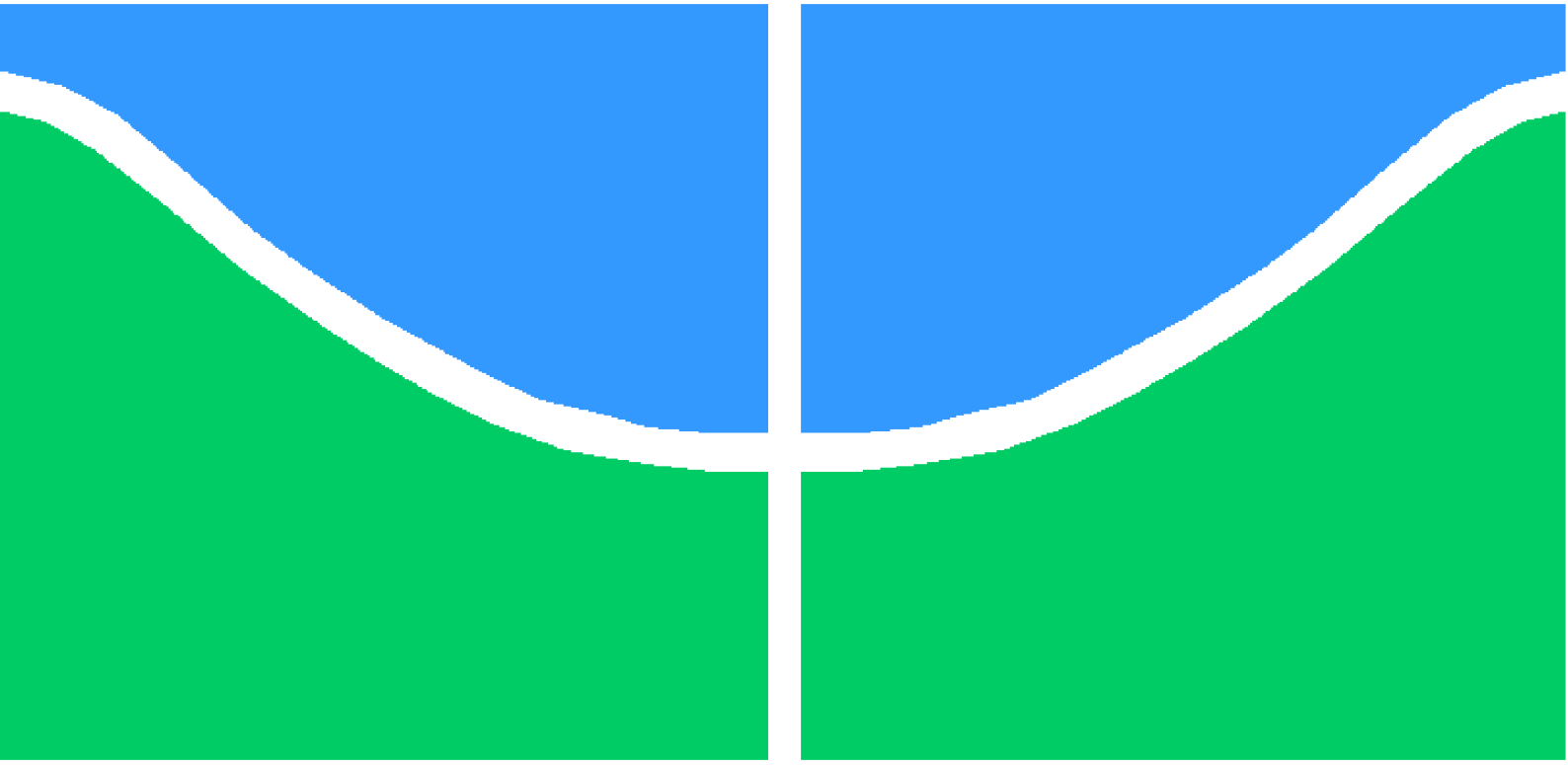


Universidade de Brasília - UnB

Faculdade UnB Gama - FGA

Engenharia de Software

Explorando Ineficiências no Mercado de Apostas Esportivas Com Uso de Modelos Preditivos e Estratégias de Arbitragem Estatística

Autor: Lucas Pimentel Quintão

Orientador: Prof. Dr. Marcelino Monteiro de Andrade

Brasília, DF

2025

Lucas Pimentel Quintão

Explorando Ineficiências no Mercado de Apostas Esportivas Com Uso de Modelos Preditivos e Estratégias de Arbitragem Estatística

Monografia submetida ao curso de graduação em Engenharia de Software da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia de Software .

Universidade de Brasília - UnB

Faculdade UnB Gama - FGA

Orientador: Prof. Dr. Marcelino Monteiro de Andrade

Brasília, DF

2025

Lucas Pimentel Quintão

Explorando Ineficiências no Mercado de Apostas Esportivas Com Uso de Modelos Preditivos e Estratégias de Arbitragem Estatística/ Lucas Pimentel Quintão. – Brasília, DF, 2025-

84 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Marcelino Monteiro de Andrade

Trabalho de Conclusão de Curso – Universidade de Brasília - UnB
Faculdade UnB Gama - FGA , 2025.

1. Apostas esportivas. 2. Arbitragem Estatística. I. Prof. Dr. Marcelino Monteiro de Andrade. II. Universidade de Brasília. III. Faculdade UnB Gama. IV. Explorando Ineficiências no Mercado de Apostas Esportivas Com Uso de Modelos Preditivos e Estratégias de Arbitragem Estatística

CDU 02:141:005.6

Lucas Pimentel Quintão

Explorando Ineficiências no Mercado de Apostas Esportivas Com Uso de Modelos Preditivos e Estratégias de Arbitragem Estatística

Monografia submetida ao curso de graduação em Engenharia de Software da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia de Software .

Trabalho aprovado. Brasília, DF, 20 de fevereiro de 2025:

Prof. Dr. Marcelino Monteiro de Andrade
Orientador

Prof. Dr. Edson Alves da Costa Júnior
Convidado 1

Prof. Dr. Rafael Morgado Silva
Convidado 2

Brasília, DF
2025

Resumo

Com a rápida expansão do mercado de apostas esportivas e o amplo alcance proporcionado pelas tecnologias atuais, esse setor tem se tornado um dos principais tópicos de debate na sociedade brasileira. Paralelamente, o cenário esportivo, especialmente na NBA, tem passado por uma revolução na coleta e aplicação de dados, impulsionando análises estatísticas cada vez mais sofisticadas. Diante desse contexto, este trabalho investiga as ineficiências do mercado de apostas esportivas na NBA e sua exploração financeira por meio de estratégias de arbitragem estatística. O objetivo foi desenvolver e testar algoritmos capazes de identificar e explorar discrepâncias entre as *odds* oferecidas por diferentes casas de apostas, além de avaliar a viabilidade da aplicação de modelos preditivos para aprimorar essa estratégia. Para isso, foi realizada a coleta de dados históricos da NBA por meio de *web scrapping*, seguida do desenvolvimento e teste de modelos de *machine learning* para prever probabilidades de vitória de times em partidas. Além disso, foi implementado um sistema automatizado para identificar oportunidades de apostas com valor esperado de lucro positivo, que operou por 112 dias, armazenando todas as apostas encontradas. Posteriormente, foi realizada uma simulação das apostas detectadas, testando diferentes estratégias de exposição ao risco. Os resultados indicaram que, apesar de os modelos preditivos não apresentarem desempenho superior às *odds* de fechamento, a estratégia de exploração de ineficiências foi altamente lucrativa em simulação, atingindo um ROI máximo de 49.365,4%, ainda que com alta volatilidade. Observou-se que uma única casa de apostas, a 1XBet, foi responsável por grande parte dos lucros, sugerindo forte desregulação em suas *odds*. Mesmo excluindo essa casa da simulação, as estratégias permaneceram lucrativas, com ROI mínimo de 353,94%. No entanto, a aplicabilidade prática dessa abordagem é limitada, uma vez que as casas de apostas impõem restrições a apostadores lucrativos e proíbem o uso de automações.

Palavras-chaves: Apostas Esportivas; Ineficiências; Arbitragem Estatística; Aprendizado de Máquina; NBA; Valor Esperado;

Abstract

With the rapid expansion of the sports betting market and the broad reach provided by current technologies, this sector has become one of the main topics of debate in Brazilian society. At the same time, the sports landscape—especially in the NBA—has undergone a revolution in data collection and application, driving increasingly sophisticated statistical analyses. In light of this context, this work investigates the inefficiencies in the NBA sports betting market and their financial exploitation through statistical arbitrage strategies.The objective was to develop and test algorithms capable of identifying and exploiting discrepancies between the odds offered by different bookmakers, as well as to evaluate the feasibility of applying predictive models to enhance this strategy. To this end, historical NBA data was collected through web scraping, followed by the development and testing of machine learning models to predict teams' win probabilities in matches. Additionally, an automated system was implemented to identify betting opportunities with a positive expected value of profit, which operated for 112 days, storing all the bets found. Subsequently, a simulation of the detected bets was conducted, testing different risk exposure strategies. The results indicated that, although the predictive models did not outperform the closing odds, the strategy of exploiting inefficiencies was highly profitable in simulation, reaching a maximum ROI of 49,365.4%, albeit with high volatility. It was observed that a single betting house, 1XBet, was responsible for a large portion of the profits, suggesting strong deregulation in its odds. Even when excluding this house from the simulation, the strategies remained profitable, with a minimum ROI of 353.94%. However, the practical applicability of this approach is limited, as bookmakers impose restrictions on profitable bettors and prohibit the use of automation.

Key-words: Sports Betting; Inefficiencies; Statistical Arbitrage; Machine Learning; NBA; Expected Value;

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ANN	*Artificial Neural Networks* (Rede Neural Artificial)
MLP	*Multilayer Perceptron* (Perceptron Multicamadas)
DNN	*Deep Neural Networks* (Rede Neural Profunda)
ROI	*Return Over Investiment* (Retorno sobre Investimento)
LSTM	*Long Short-Term Memory*
MSE	*Mean Squared Error* (Erro Quadrático Médio)

Sumário

1 Introdução

No vasto panorama dos esportes profissionais, a *National Basketball Association* (NBA) se destaca como uma liga de renome internacional, reunindo os maiores talentos do basquete mundial. Dessa forma, a NBA não é apenas um símbolo do mais alto nível de excelência atlética, mas também uma área para a inovação e o avanço tecnológico.

Nos últimos anos, o cenário esportivo tem testemunhado uma revolução sem precedentes na forma como os dados são coletados, analisados e aplicados (CAO, 2012). Esse movimento em direção à análise de dados tem se mostrado especialmente relevante na NBA, onde o uso adequado de estatísticas pode ser a diferença entre a vitória e a derrota. Desde métricas básicas, como pontos marcados e rebotes, até indicadores estatísticos avançados, como eficiência de tiro e avaliação defensiva e ofensiva, os dados desempenham um papel crucial na avaliação do desempenho de um jogador e no desenvolvimento de estratégias de equipe (KUBATKO et al., 2007).

Neste contexto, o papel da inteligência artificial se destaca como uma ferramenta poderosa para extrair informações valiosas a partir de conjuntos massivos de dados. O aprendizado de máquina oferece a capacidade de identificar padrões, fazer previsões e tomar decisões de forma automatizada e precisa, o que o torna ideal para lidar com a complexidade e a dimensionalidade dos dados esportivos modernos (BAUMER; ZIMBALIST, 2014).

Aliado a isso, o mercado de apostas esportivas está em rápida expansão e abrange uma variedade de modalidades esportivas. Esse mercado é dinâmico e se distingue essencialmente do mercado tradicional de apostas, alcançando um público muito maior graças ao fácil acesso proporcionado pelas tecnologias de informação e comunicação disponíveis (POVOA et al., 2023). Com essa expansão, muito esforço foi dedicado à previsão dos resultados de eventos esportivos com o objetivo de testar este mercado quanto à sua eficiência econômica (STEKLER; SENDOR; VERLANDER, 2010).

Este trabalho tem como objetivo investigar e explorar as ineficiências nos mercados de apostas esportivas, desenvolvendo algoritmos capazes de identificar e aproveitar oportunidades financeiras. Com foco no desenvolvimento e teste de uma estratégia de apostas que compara as *odds* oferecidas por diferentes casas de apostas com as probabilidades previstas por modelos de *machine learning*, avaliando sua viabilidade econômica.

Este enfoque não só avança no aspecto técnico e acadêmico dentro da engenharia de software e ciência de dados, mas também impulsiona inovações práticas no campo das apostas esportivas, que tem experimentado uma expansão significativa, com seu mercado crescendo mais de 734% desde 2021 (NAKAMURA, 2024).

1.1 Objetivo

O objetivo geral deste trabalho é investigar e explorar as ineficiências dos mercados de apostas esportivas, desenvolvendo algoritmos capazes de identificar e aproveitar oportunidades financeiras.

Os três objetivos específicos são:

1. Realizar uma análise dos diferentes indicadores estatísticos disponíveis para a NBA e selecionar aquele que será alvo dos modelos de predição.

2. Estudar a viabilidade da utilização de modelos de *machine learning* adequados para a previsão de probabilidades associadas a mercados de apostas, avaliando sua capacidade de oferecer estimativas precisas e calibradas para auxiliar na identificação de ineficiências.

3. Desenvolver e testar a viabilidade econômica de uma estratégia de apostas esportivas, utilizando em conjunto as probabilidades previstas pelos modelos preditivos e as *odds* disponibilizadas pelas casas de apostas.

1.2 Metodologia

Para a realização deste trabalho, será utilizado o site *basketball-reference*[1], que reúne dados estatísticos históricos da NBA, permitindo a construção de uma base de dados inicial (KUBATKO et al., 2007). Esses dados incluem estatísticas de jogadores em partidas desde a temporada de 1982-83 até a temporada mais recente (2023-24) e serão extraidos do site por meio de técnicas de *web-scrapping*.

Para os dados das casas de apostas, que serão utilizados para a comparação e avaliação dos modelos de inteligência artificial, serão empregadas as ferramentas disponíveis no site *Sportsbook Reviews Online*[2], que oferece o valor das linhas de abertura e de fechamento de diversas casas de apostas.

Após a coleta, será realizada a definição do indicador estatístico que será o alvo das predições, com foco em selecionar aqueles mais utilizados em estudos semelhantes e com maior potencial econômico, tomando como base as referências bibliográficas utilizadas no estudo.

No contexto dos modelos preditivos, será utilizado como base um artigo que busca explorar mercados de apostas esportivas com o uso de *machine learning* (HUBACEK; SOUREK; ZELEZNY, 2019). Sua estratégia consiste em utilizar um modelo de regressão

[1] https://www.basketball-reference.com/
[2] https://www.sportsbookreviewsonline.com/

logística simples para servir de base na comparação de resultados e, além disso, utilizar modelos baseados em arquiteturas de redes neurais que serão: Redes Neurais Artificiais (ANNs) e Redes Neurais Recorrentes (RNNs).

Para a avaliação dos modelos, a métrica de calibração foi escolhida por sua relevância na precisão das probabilidades preditas, especialmente no contexto de apostas esportivas, superando métricas tradicionais como a acurácia (WALSH; JOSHI, 2024). Finalmente, o treinamento e teste dos modelos serão realizados utilizando ferramentas como Jupyter, TensorFlow e Pandas para desenvolver e avaliar os modelos em um ambiente controlado.

Em seguida, será desenvolvido o algoritmo responsável por coletar as *odds* de diferentes casas de apostas e compará-las, além de analisar as *odds* preditas pelos modelos de inteligência artificial. O objetivo é identificar discrepâncias que possibilitem a execução da estratégia de apostas proposta. A partir dessas discrepâncias, serão realizadas simulações de apostas para avaliar a eficácia da abordagem. Por fim, a avaliação dos resultados será realizada utilizando a métrica de retorno sobre investimento (ROI).

1.3 Organização do Documento

Este trabalho está organizado em capítulos. O primeiro capítulo, que é este, apresenta uma introdução sobre o uso de inteligência artificial na NBA e nos mercados de apostas, bem como uma visão geral do trabalho desenvolvido. Os capítulos subsequentes são os seguintes:

- **Capítulo 2 - Referencial Teórico:** Apresenta os fundamentos conceituais e as pesquisas que embasam este estudo, fornecendo o conhecimento necessário para a compreensão e desenvolvimento do trabalho.
- **Capítulo 3 - Metodologia:** Descreve em detalhes os métodos, técnicas e estratégias adotadas na condução do estudo, desde a coleta de dados até a implementação dos modelos e algoritmos.
- **Capítulo 4 - Resultados:** Expõe de forma estruturada o desenvolvimento do estudo, detalhando as etapas realizadas, os experimentos conduzidos e os principais achados da pesquisa.
- **Capítulo 5 - Conclusão:** Apresenta uma síntese dos principais achados do estudo, discutindo as implicações dos resultados obtidos, as limitações encontradas e sugestões para pesquisas futuras.

2 Referencial Teórico

Este capítulo explica os principais conceitos teóricos que servirão de base para a pesquisa e desenvolvimento deste trabalho. Serão discutidos assuntos relevantes para a compreensão do contexto em que o estudo se encontra, trazendo conceitos teóricos fundamentais para embasar as análises críticas.

2.1 Modelos Preditivos

Modelos Preditivos são métodos estatísticos e algoritmos de aprendizado de máquina usados para prever resultados futuros com base em dados históricos e atuais. Eles se concentram em identificar padrões nos dados e fazer previsões sobre novas observações (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Nesta seção de referencial teórico, serão apresentados os principais conceitos de *machine learning* que fundamentam este trabalho. A abordagem incluirá explicações sobre os métodos e técnicas de aprendizado de máquina relevantes para a predição de eventos esportivos, detalhando os algoritmos, métricas de avaliação, processos de otimização e estratégias de modelagem que serão utilizados ao longo do estudo.

2.1.1 Aprendizado supervisionado

Os modelos preditivos supervisionados são construídos através de um processo de treinamento, onde um algoritmo aprende a partir de um conjunto de dados de treinamento, e são avaliados com um conjunto de dados de teste para medir sua precisão (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Em tais modelos, os *datasets* consistem em uma coleção de exemplos rotulados no formato $[(x_i, y_i)]_{i=1}^{N}$, onde cada elemento x_i é um vetor de características com dimensão $j \in \{1, .., D\}$, representado como:

$$x_i = \begin{bmatrix} x_i^{(D)} \\ x_i^{(D)} \\ \vdots \\ x_i^{(D)} \end{bmatrix} \tag{2.1}$$

onde cada $x^{(j)}$ contém um valor que descreve o exemplo de alguma maneira. O elemento y_i corresponde ao rótulo do exemplo, que pode ser um valor pertencente a um conjunto finito de classes, um número real ou outras estruturas mais complexas como grafos, vetores e matrizes. (BURKOV, 2019).

Dessa forma, o objetivo de um algoritmo de aprendizado supervisionado é usar um conjunto de dados para produzir um modelo que recebe um vetor de características x como entrada e fornece como *output* informações que permitem deduzir o rótulo correspondente a esse vetor de características (BURKOV, 2019).

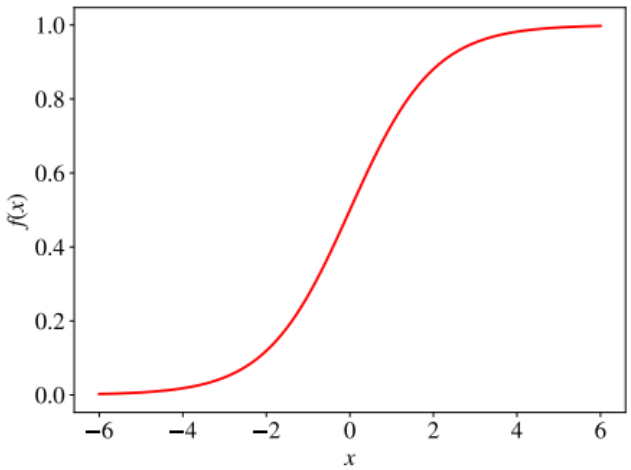
2.1.2 Regressão Logística

A regressão logística é um método linear utilizado principalmente para resolver problemas de classificação. Seu objetivo é modelar y como as probabilidades posteriores das K classes por meio de funções lineares em x, garantindo ao mesmo tempo que elas somem a um e permaneçam no intervalo $[0, 1]$ (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Neste trabalho, focaremos nos casos de classificação binária, onde o conjunto de classes possui apenas dois elementos.

Como a combinação linear das características como $wx_i + b$ é uma função cuja imagem é o cojunto dos números reais, e o rótulo y_i pode assumir apenas dois valores, é necessário utilizar a função sigmoid, definida pela Eq. 2.2. A função sigmoid é contínua e possui contradomínio no intervalo (0,1), como pode ser visto na Fig. 1 (BURKOV, 2019).

$$\sigma(x) = \frac{1}{1 + e^{-z}} \tag{2.2}$$

Figura 1 – Função logística.



Fonte: Burkov (2019)

Assim, se o valor retornado pelo modelo para a entrada x estiver abaixo de um determinado limiar t, atribuímos um rótulo negativo a x; caso contrário, o exemplo é rotulado como positivo. O modelo de regressão logística é definido pela Eq. 2.3, onde w é um vetor de parâmetros com dimensão D e b é um número real. Ao otimizar apropriadamente os valores de x e b, o *output* da função $f(x)$ pode ser interpretado como a probabilidade de y_i ser positivo (BURKOV, 2019).

$$f_{w,b}(x) = \frac{1}{1 + e^{-(wx+b)}} \tag{2.3}$$

2.1.3 Redes Neurais

Redes Neurais Artificiais (ANNs) são modelos computacionais inspirados na estrutura e no funcionamento do cérebro humano. Essas redes consistem em unidades de processamento interconectadas, conhecidas como neurônios. O poder de computação das redes neurais advém, em grande parte, de sua estrutura distribuída massivamente paralela e de sua capacidade de aprender e generalizar. A generalização refere-se à habilidade da rede neural de gerar saídas coerentes para entradas que não foram utilizadas durante o processo de treinamento (HAYKIN, 2009).

Uma ANN, assim como a regressão logistíca, é uma função matemática representada pela Eq. 2.4. No entanto, as redes neurais possuem uma estrutura em camadas aninhadas, onde cada camada da rede representa uma função distinta. Como exemplo, a estrura de uma rede neural com três camadas pode ser definida pela Eq. 2.5 (BURKOV, 2019).

$$y = f_{NN}(x) \tag{2.4}$$

$$y = f_{NN}(x) = f_3(f_2(f_1(X))) \tag{2.5}$$

2.1.3.1 Perceptron Multicamadas (MLP)

O perceptron é a forma mais simples de uma rede neural utilizada para a classificação de padrões considerados linearmente separáveis (ou seja, padrões que estão em lados opostos de um hiperplano). Basicamente, ele consiste em um único neurônio com pesos sinápticos ajustáveis e um viés. Em termos matemáticos, um perceptron k representado pela Fig. 2, pode ser representado pelas Eqs. 2.6 e 2.7 onde x representa as entradas do modelo; w_k, os pesos; u_k, o resultado da combinação linear das entradas; b_k, o viés do modelo; φ, a função de ativação e y_k, o *output* produzido pelo modelo (HAYKIN, 2009).

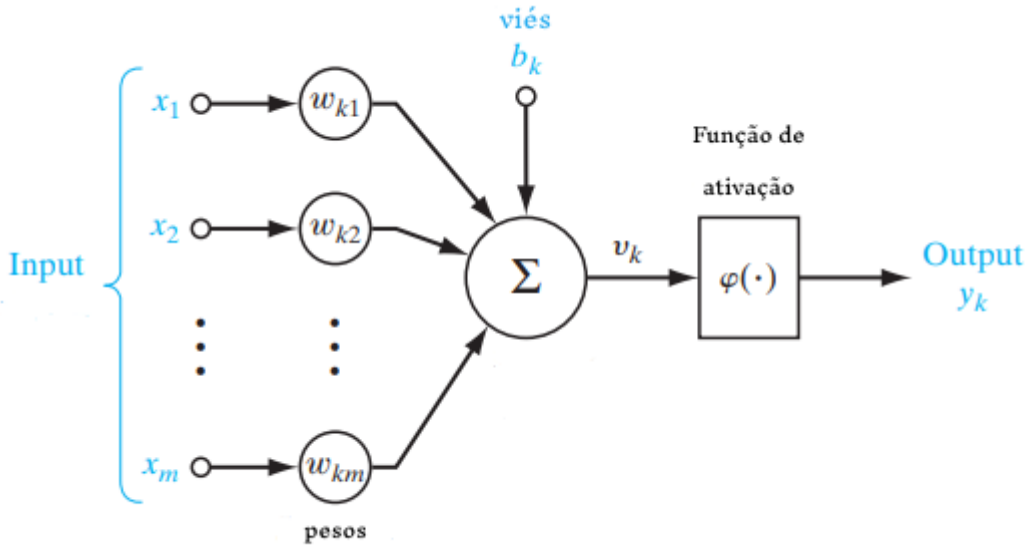
$$u_k = \sum_{j=1}^{m} w_{kj} \cdot x_j \tag{2.6}$$

$$y_k = \varphi(u_k + b_k) \tag{2.7}$$

As funções de ativação determinam a saída de um perceptron e são definidas por $\varphi(v)$, onde:

$$v_k = u_k + b_k. \tag{2.8}$$

Figura 2 – Perceptron.



Fonte: Haykin (2009), com modificações

Existem diversas funções de ativação possíveis para uso, entre as quais destacam-se no contexto deste estudo: a função sigmoid, já apresentada anteriormente, definida pela Eq. 2.2; a função ReLU (unidade linear retificada), que retorna zero quando a entrada v é negativa e v caso contrário, conforme definido na Eq. 2.9; e a função TanH (tangente hiperbólica), que é similar à regressão logística, mas com contradomínio no intervalo $(-1, 1)$, sendo definida pela Eq. 2.10 (BURKOV, 2019).

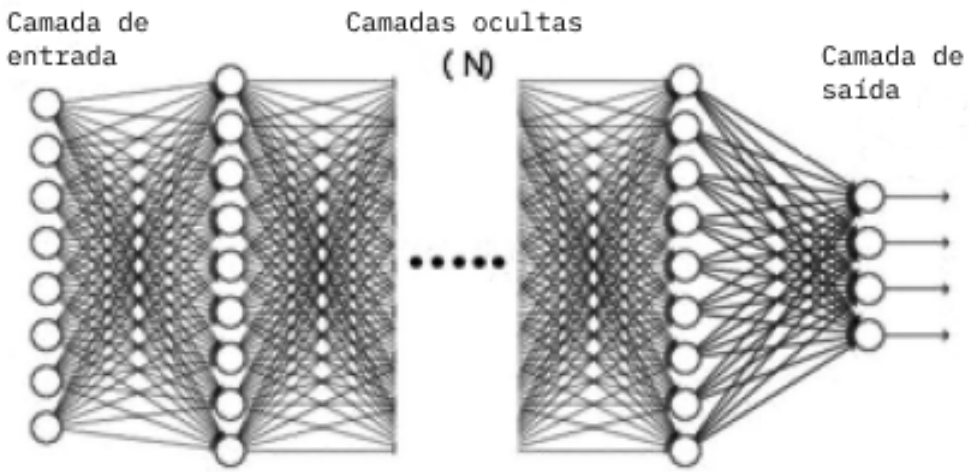
$$\text{relu}(v) = \begin{cases} 0, & \text{se } v < 0 \\ v, & \text{caso contrário} \end{cases} \tag{2.9}$$

$$\tanh(v) = \frac{e^{v} - e^{-v}}{e^{v} + e^{-v}} \tag{2.10}$$

O perceptron pode ser visto como uma rede neural de camada única com apenas um neurônio. Ao adicionar mais camadas e neurônios à estrutura da rede neural, é possível criar um Perceptron Multicamadas (MLP), conforme ilustrado na Fig. 3. O MLP é reconhecido como a arquitetura base das redes neurais profundas (DNN). Um MLP típico é uma rede *feedfoward* totalmente conectada, composta por uma camada de entrada, que recebe os dados; uma camada de saída, que faz decisões ou previsões sobre o sinal de entrada; e uma ou mais camadas ocultas entre elas, que atuam como o núcleo computacional da rede (SARKER, 2021).

Para resolver problemas de classificação, como os abordados ao longo deste trabalho, a camada de saída da rede neural geralmente possui apenas uma unidade. Se essa

Figura 3 – Perceptron Multicamadas.



Fonte: Sarker (2021), com modificações

unidade tiver uma função de ativação logística, a rede neural é considerada um modelo de classificação (BURKOV, 2019).

Para treinar o MLP, utiliza-se o algoritmo de Retropropagação, a técnica de aprendizado supervisionado mais amplamente empregada, considerada o bloco de construção mais fundamental de uma rede neural. Durante a etapa de treinamento, são aplicadas diversas abordagens de otimização, como o Gradiente Descendente Estocástico, o algoritmo de Broyden-Fletcher-Goldfarb-Shanno com memória limitada (L-BFGS) e o otimizador Adam. Além disso, o perceptron multicamadas requer o ajuste de vários hiperparâmetros, incluindo o número de camadas ocultas, neurônios e iterações (SARKER, 2021).

2.1.3.2 *Long Short-Term Memory* (LSTM)

Uma Rede Neural Recorrente (RNN) é uma arquitetura neural amplamente utilizada, especialmente adequada para dados sequenciais ou séries temporais. Como seu diferencial, as RNNs utilizam a saída da etapa anterior como entrada para a etapa atual. (SARKER, 2021). Esse *loop*, faz com que as RNNs não possuam uma arquitetura *feedfoward*, dessa forma, cada unidade u de uma camada recorrente l, possui um estado denominado por $h_{l,u}$. Esse estado pode ser visto como a memória da unidade. Assim, em uma RNN cada unidade u da camada recorrente l deve receber duas entradas: um vetor de saídas da camada $l-1$ e um vetor de estados dessa mesma camada l de um passo temporal t anterior (BURKOV, 2019).

Para treinar modelos RNNs é utilizada uma versão específica do algoritmo de retropropagação, chamado de retropropagação no tempo (BURKOV, 2019). Assim, um problema que requer atenção nas aplicações práticas de redes recorrentes é o problema dos gradientes de desaparecimento, que surge no treinamento da rede para produzir uma

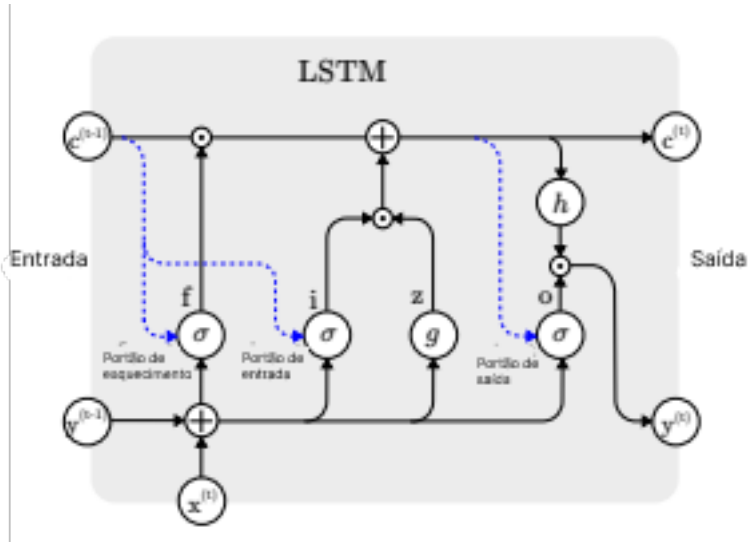
resposta desejada no tempo atual que depende de dados de entrada no passado distante. Devido às não linearidades combinadas, uma mudança infinitesimal em uma entrada distante no tempo pode não ter praticamente nenhum efeito no treinamento da rede. Esse problema dos gradientes de desaparecimento torna o aprendizado de dependências de longo prazo em algoritmos de treinamento baseados em gradiente difícil, senão virtualmente impossível, em certos casos (HAYKIN, 2009).

Os modelos de rede neural recorrente mais eficazes usados na prática são as RNNs com portões. Isso inclui as redes *Long Short-Term Memory* (LSTM) e as redes baseadas na unidade recorrente fechada (GRU) (BURKOV, 2019). Neste estudo, focaremos especificamente nos modelos *Long Short-Term Memory*.

O modelo LSTM é uma poderosa rede neural recorrente projetada para superar o problema do gradiente de desaparecimento. Neste modelo, cada unidade u é uma rede neural recorrente que possui um carrossel de erro constante (CEC), criado para evitar a propagração de erros, mantendo-os dentro da própria célula. Assim, uma unidade LSTM básica é composta por uma célula, um portão de entrada, um portão de saída e um portão de esquecimento (HOUDT; MOSQUERA; NáPOLES, 2020).

A arquitetura do LSTM consiste em blocos de memória, um conjunto de sub-redes conectadas recorrentemente. A Figura 4 representa a arquitetura de um bloco LSTM, possuindo a entrada $x^{(t)}$, a saída $y^{(t)}$. Nesse modelo, σ representa a função de ativação sigmoid, enquanto a função de ativação da entrada g e a função de ativação da saída h, normalmente utilizam a função tangente hiperbólica (HOUDT; MOSQUERA; NáPOLES, 2020).

Figura 4 – Bloco LSTM.



Fonte: Houdt, Mosquera e Nápoles (2020), com modificações

A seguir será feita a descrição de cada portão, assim como da entrada e da saída

da célula, considerando uma rede composta por N blocos e M entradas:

- **Entrada do bloco**: Neste passo a entrada do bloco $x^{(t)}$ é combinada com o output do último bloco $y^{(t-1)}$. Esta combinação é feita por meio da Eq. 2.11, onde W_z e R_z são os pesos de $x^{(t)}$ e $y^{(t-1)}$, respectivamente, e b_z é o vetor de viés (HOUDT; MOSQUERA; NáPOLES, 2020).

$$z^{(t)} = g(W_z x^{(t)} + R_z y^{(t-1)} + b_z) \tag{2.11}$$

- **Portão de entrada**: Nesta etapa o valor do portão de entrada é atualizado com a combinação de $x^{(t)}$, $y^{(t-1)}$ e o valor da célula $c^{(t-1)}$ na última iteração. Isto é feito através da Eq. 2.12, na qual $\odot$ representa a multiplicação elemento a elemento de dois vetores e p_i é o peso associado a $c^{(t-1)}$ (HOUDT; MOSQUERA; NáPOLES, 2020).

$$i^{(t)} = \sigma(W_i x^{(t)} + R_i y^{(t-1)} + p_i \odot c^{(t-1)} + b_i) \tag{2.12}$$

- **Portão de esquecimento**: Nesse momento o valor de ativação $f^{(t)}$ é calculado a partir de $x^{(t)}$, $y^{(t-1)}$, $c^{(t-1)}$ e pelo viés do portão b_f a partir da Eq. 2.13. Nesta etapa é determinada qual informação deve ser removida do estado da última célula $c^{(t-1)}$ (HOUDT; MOSQUERA; NáPOLES, 2020).

$$f^{(t)} = \sigma(W_f x^{(t)} + R_f y^{(t-1)} + p_f \odot c^{(t-1)} + b_f) \tag{2.13}$$

- **Célula**: Esta etapa calcula o valor da célula $c^{(t)}$, por meio da Eq. 2.14, combinando a entrada do bloco $z^{(t)}$, o valor do portão de entrada $i^{(t)}$ e o do portão de esquecimento $f^{(t)}$, com o valor da última célula $c^{(t-1)}$

$$c^{(t)} = z^{(t)} \odot i^{(t)} + c^{(t-1)} \odot f^{(t)} \tag{2.14}$$

- **Portão de saída**: Já o portão de saída é calculado a partir da Eq. 2.15, com os pesos e viéses do portão.

$$o^{(t)} = \sigma(W_o x^{(t)} + R_o y^{(t-1)} + p_o \odot c^{(t-1)} + b_o) \tag{2.15}$$

- **Saída do bloco**: Por fim, a saída do bloco é calulada pela Eq. 2.16, combinando o valor da célula atual $c^{(t)}$ com o do portão de saída $o^{(t)}$.

$$y^{(t)} = g(c^{(t)}) \odot o^{(t)} \tag{2.16}$$

2.1.4 Métricas para avaliação de modelos

A avaliação de modelos preditivos tem o objetivo de descartar métodos que apresentem desempenho insatisfatório e continuar a otimização daqueles que se mostram promissores. No aprendizado de máquina supervisionado, os datasets são divididos em dados de treino e dados de teste. Os dados de treino são utilizados para o treinamento e validação do modelo, enquanto os dados de teste são usados para comparar os resultados preditos com os valores reais do conjunto de informações. Dessa forma, é possível avaliar o desempenho do modelo e compará-lo com outros modelos existentes ou com predições feitas por humanos (RAINIO; TEUHO; KLéN, 2024). Assim, existem diversas métricas utilizadas para avaliar o desempenho desses modelos. Nesta seção, serão detalhadas as principais métricas empregadas na avaliação de modelos de classificação binária.

Na classificação binária, um rótulo y_i pode ter dois possíveis resultados: positivo ou negativo. Com isso, cada valor de rótulo predito $\hat{y}_i$ pode ser classificado em uma das quatro categorias: verdadeiro positivo (VP), quando um rótulo positivo é corretamente predito como positivo; falso positivo (FP), quando um rótulo negativo é erroneamente predito como positivo; verdadeiro negativo (VN), quando um rótulo negativo é corretamente predito; e falso negativo (FN), quando um rótulo positivo é incorretamente predito como negativo. (RAINIO; TEUHO; KLéN, 2024). Assim, uma matriz de confusão contendo os valores de VP, FP, VN e FN pode ser elaborada, como mostra o Quadro 1. A partir

Quadro 1 – Matriz de confusão de rótulos

Predito/Classe verdadeira	**Positivo**	**Negativo**
Positivo	VP	FP
Negativo	FN	VN

Fonte: Rainio, Teuho e Klén (2024).

da matriz de confusão, podemos definir as principais métricas utilizadas na avaliação de modelos de classificação binária:

- **Acurácia:** expressa a porcentagem de rótulos corretamente preditos em relação ao total de rótulos preditos. A acurárcia (acc) é representada pela Eq. 2.17.

$$\text{acc} = \frac{VP + VN}{VP + FP + FN + VN} \tag{2.17}$$

- ***Recall*:** também chamado de sensibilidade, o recall (rec) expressa a taxa de verdadeiros positivos preditos pelo modelo e é definido pela Eq. 2.18

$$\text{rec} = \frac{VP}{VP + FN} \tag{2.18}$$

- **Especificidade:** calcula a porcentagem de verdadeiros negativos preditos pelo modelo e é calculada utilizando a Eq. 2.19.

$$\text{esp} = \frac{VN}{VN + FP} \tag{2.19}$$

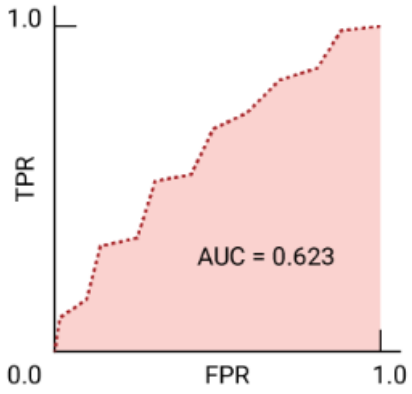
- **Precisão:** a precisão (pre) calcula a porcentagem de acertos entre as instâncias classificadas como positivo e é representada pela Eq. 2.20.

$$\text{pre} = \frac{VP}{VP + FP} \tag{2.20}$$

Outra métrica utilizada para a avaliação de modelos preditivos é a curva característica de operação do receptor (ROC), que utiliza as previsões numéricas antes da conversão para valores binários para plotar a sensibilidade contra a taxa de falso positivo (FPR) para todos os limiares possíveis. A curva ROC é uma função monotonicamente crescente dentro do quadrado unitário delimitado pelos pontos (0,0) e (1,1). Quanto mais próxima a curva ROC estiver do ponto (0,1), melhores são as previsões do modelo (RAINIO; TEUHO; KLéN, 2024).

Além disso, a área sob a curva ROC (AUC) é outra métrica de avaliação com valores no intervalo [0,1], sendo que um modelo com AUC igual a 1 representa um modelo perfeito (RAINIO; TEUHO; KLéN, 2024). A Figura 5, mostra um exemplo de curva ROC de um modelo preditivo e sua respectiva AUC, nela FPR indica a taxa de falsos positivos e TPR a taxa de verdadeiros positivos.

Figura 5 – Exemplo de curva ROC com seu valor de AUC.



Fonte: *Google for Developers*

A calibração é uma métrica utilizada para estimar quão próximas as probabilidades preditas por um modelo estão das probabilidades reais. Para aplicações de modelos de classificação em cenários que exigem probabilidades que reflitam a verdadeira probabilidade de uma previsão, essa métrica se destaca como uma métrica essencial (KUMAR; LIANG; MA, 2019).

Supondo que o modelo de classificação possui a forma apresentada pela Eq.2.21, onde $\hat{y}_i$ é o rótulo predito para um exemplo x_i e $\hat{p}_i$ a probabilidade associada a este rótulo. O modelo de predição poderá ser considerado um modelo bem calibrado se, a distribuição das classes seja aproximadamente distribuida como $\hat{p}$, o vetor de probabilidades previstas entre os exemplos de teste (WALSH; JOSHI, 2024).

$$f(x) = (\hat{y}_x, \hat{p}_x) \tag{2.21}$$

Como a maioria dos modelos modernos de aprendizado de máquina, como redes neurais, não produzem probabilidades calibradas de forma nativa, pesquisadores recorrem a métodos de recalibração que ajustam as saídas de um modelo não calibrado, transformando-as em probabilidades calibradas. Abordagens de escalonamento para recalibração, como o *Platt scaling*, a regressão isotônica e o escalonamento de temperatura, são amplamente utilizadas e exigem muito poucas amostras (KUMAR; LIANG; MA, 2019).

Para um classificador binário definido por $f : X \to [0, 1]$, onde a saída representa a confiança do modelo que o rotúlo é 1. O erro de calibração (CE) examina a diferença entre a probabilidade dada pelo modelo e a probabilidade real, e é representado pela Eq. 2.22, onde E representa o valor esperado das variáveis aleatórias. (KUMAR; LIANG; MA, 2019).

$$\mathrm{CE}(f) = (E[|f(x) - E[y \mid f(x)]|^2])^{\frac{1}{2}} \tag{2.22}$$

Além disso, para complementar o cálculo do erro de calibração, é calculado o erro quadrático médio (MSE) das probabilidades preditas pelo modelo. Para isso, é utilizada a Eq. 2.23.

$$\mathrm{MSE}(f) = E[(f(x) - y)^2] \tag{2.23}$$

2.2 Dados

Nesta seção, será abordada a fundamentação teórica relacionada aos diferentes tipos de dados, destacando suas características e a importância de sua organização para a realização de análises eficazes. Os dados são a base de qualquer estudo ou pesquisa, e compreendê-los em suas diversas formas é essencial para a escolha das técnicas de análise e tratamento adequadas.

2.2.1 Tipos de Dados

A organização dos dados em diferentes estruturas é fundamental para sua análise e utilização eficaz. Nesta seção, serão explorados os principais tipos de dados, destacando suas características e aplicações específicas.

Dados estruturados referem-se a informações organizadas em um formato rígido, fixo e padronizado. Este tipo de dado é comumente utilizado para armazenar registros em tabelas de banco de dados relacionais. Nessas bases, esses dados são organizados em linhas e colunas, onde cada coluna representa um atributo específico e cada linha corresponde a uma entrada única, garantindo consistência e integridade dos dados. A estrutura rígida permite a utilização de linguagens de consulta como SQL para realizar operações complexas, como filtragem e agregação de dados (ELMASRI; NAVATHE, 2004).

Dados semiestruturados são informações que podem possuir uma estrutura, porém, diferente dos dados estruturados, não possuem uma estrutura rígida. Dessa forma, nem toda informação coletada deverá ter a mesma estrutura. Neste tipo de dado não existem esquemas predefinidos, podendo adicionar atributos novos em algumas entidades a qualquer momento. Esses dados não são armazenados em tabelas convencionais de banco de dados relacionais, mas utilizam estruturas flexíveis, como XML ou JSON, que permitem a inclusão de atributos variados e formatos hierárquicos (ELMASRI; NAVATHE, 2004).

Os dados não estruturados são as informações que não seguem nenhum tipo de padrão ou estrutura predefinida. Esses dados são geralmente compostos por textos, imagens, vídeos, áudios e outros tipos de conteúdo multimídia que não podem ser facilmente categorizados em tabelas ou bancos de dados relacionais (ELMASRI; NAVATHE, 2004).

2.2.2 Coleta de Dados

A coleta de dados é uma etapa crucial em qualquer projeto de análise de informações, pois é por meio dela que se obtêm as bases sobre as quais se construirão as análises e conclusões subsequentes. Dessa forma, a seguir será detalhada a técnica utilizada para a coleta de dados para o desenvolvimento do trabalho.

Web Scraping é o processo automatizado de extração de informações e dados de sites da internet, utilizando softwares ou scripts especificamente desenvolvidos para navegar e coletar conteúdo de páginas web de forma eficiente. Essa técnica permite a compilação de grandes volumes de dados não estruturados presentes em páginas web, que podem ser transformados para outras formas de dados estruturados ou semiestruturados (KHDER, 2021).

2.3 A Matemática das Apostas Esportivas

Nesta seção, serão apresentados os conceitos fundamentais da probabilidade que embasam o funcionamento do mercado de apostas esportivas, proporcionando uma compreensão clara de suas dinâmicas e estruturas.

2.3.1 *Odds*

As *odds* podem ser definidas como o multiplicador do valor apostado em caso de uma aposta vencedora. Elas refletem a probabilidade dos diferentes resultados e, em vez de serem expressas como porcentagens, são apresentadas como um indicador do retorno potencial sobre a aposta, facilitando a compreensão do ganho financeiro esperado (CORTIS, 2015). Embora existam várias formas de representar as *odds*, no Brasil elas são comumente expressas em formato decimal.

Quando definida em seu formato decimal, ao inverter as *odds*, obtém-se a probabilidade implícita de o resultado ocorrer (CORTIS, 2015). Então, assumindo O_i como a *odd* dada a um resultado $i \in \{1, ..., n\}$ de um evento por uma casa de apostas, temos que a probabilidade implícita deste evento π_i calculada pela casa de apostas é definida pela Eq.(2.24) .

$$\pi_i = \frac{1}{O_i} \tag{2.24}$$

Além disso, dentro de um mercado de apostas, a linha refere-se ao valor de referência estabelecido para uma determinada estatística ou condição de um evento esportivo. Por exemplo, no mercado de total de pontos, a linha pode ser fixada em 210.5 pontos em um jogo da NBA, significando que o apostador pode apostar se o total de pontos combinados das equipes será maior ou menor que esse valor. Da mesma forma, para apostas em desempenho de jogadores, como pontos de um jogador, a linha pode ser definida em 25.5 pontos, indicando se um determinado jogador marcará mais ou menos do que esse número. A definição das linhas influencia diretamente as *odds* oferecidas pelas casas de apostas, uma vez que ajustes são feitos com base no comportamento do mercado e na expectativa do desempenho do evento.

2.3.1.1 Margem da Casa de Apostas

Outro fator relevante a ser considerado em relação às *odds* é a margem aplicada pelas casas de apostas para garantir seus lucros. Em um cenário ideal, onde as *odds* oferecidas são justas (ou seja, sem a adição de margens), a soma das probabilidades implícitas r de um evento esportivo com k possíveis resultados seria igual a 100%. Assim,

podemos definir a margem (v) aplicada pela casa de apostas, conforme apresentado na Eq. (2.26) (WHEATCROFT, 2020).

$$r = \sum_{i=1}^{k} \frac{1}{O_i} \tag{2.25}$$

$$v = r - 1 \tag{2.26}$$

Assim, diversas metodologias foram desenvolvidas para normalizar as probabilidades implícitas das *odds* de modo que sua soma seja igual a 1. Neste trabalho, adotaremos a abordagem de normalização apresentada na Eq. (2.27) (WHEATCROFT, 2020).

$$\tilde{\pi}_i = \frac{\pi_i}{r} \tag{2.27}$$

Para ilustrar um caso real, a Fig. (6) mostra as *odds* de um evento esportivo com 2 possíveis resultados do site de apostas Bet365[1]. Ao somar as probabilidades implícitas desses resultados, obtém-se um valor aproximado de 104,39%. Dessa forma, aplicando a Eq. (2.25), a margem da casa de apostas é definida como 4,39%. Seguindo adiante, ao utilizarmos a Eq. (2.24) e Eq. (2.25) temos que a probabilidade implícita do Time A (com *odd* de 2.75) vencer a partida é de 36,36% , equanto a probabilidade implícita normalizada da vitória é de 34,83%.

Figura 6 – Exemplo de *odds* em casas de apostas.

Fonte: Bet365

2.3.1.2 Valor de Fechamento e Valor de Abertura

Dentro de uma casa de apostas, as *odds* não são valores fixos. Desde o momento de sua abertura até o início do evento esportivo, essas *odds* passam por frequentes variações, influenciadas por notícias sobre o evento, informações privilegiadas, movimentações do mercado e outros fatores que podem impactar diretamente o cálculo das probabilidades associadas aos resultados do evento (GANDAR et al., 1998). Nesse contexto, podemos definir o Valor de Abertura como a odd disponível nos primeiros momentos em que a casa abre o mercado, e o Valor de Fechamento como a odd disponível minutos antes do início do evento.

Stekler, Sendor e Verlander (2010) mostram em seu estudo que frequentemente há grandes variações entre as cotações de abertura e fechamento em jogos da NBA, concluindo

[1] https://www.bet365.com/

que as *odds* de fechamento mostram ser melhores preditores do resultado do jogo do que as *odds* de abertura. Portanto, o Valor de Fechamento pode ser considerado como uma estimativa mais precisa das probabilidades dos resultados do evento em comparação com o Valor de Abertura.

2.3.2 Apostas de valor

O Valor Esperado, ou esperança matemática, $E(X)$ de uma variárel aleatória X é a média dos valores assumidos por X, ponderada pelas probabilidades correspondentes a esses valores (ROLLA; LIMA, 2024). No caso de X ser uma variável aleatória discreta, o Valor Esperado de X é definido pela Eq. (2.28),

$$E(X) = \sum_{j=1}^{\infty} x_j \cdot p(x_j) \tag{2.28}$$

desde que a série seja convergente. Caso contrário, E(X) não é definido.

Uma interpretação que pode ser dada ao valor esperado é tê-lo como o valor médio após muitas realizações de um mesmo experimento (ROLLA; LIMA, 2024). Como exemplo, representado pela Eq. (2.29), consideremos o lançamento de uma moeda 2 vezes, sendo X a variável aleatória que conta a quantidade de vezes que o resultado é coroa.

$$E(X) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{2}{4} + 2 \cdot \frac{1}{4} = 1 \tag{2.29}$$

No contexto das apostas esportivas, o valor esperado é fundamental para o cálculo do lucro esperado, tanto para as casas de apostas quanto para os apostadores. No artigo de Cortis (2015), é feita uma análise das restrições que as casas de apostas precisam impor sobre suas *odds* para assegurar um valor esperado positivo nos retornos financeiros. Uma das restrições demonstradas é que a probabilidade implícita nas *odds* deve ser maior do que a probabilidade real do evento, garantindo assim que a casa de apostas mantenha um lucro esperado positivo.

Para a demonstração feita por Cortis (2015), identificada pela Eq. (2.30) e pela Eq. (2.31), é utilizada a prova por contradição da proposição: uma casa de apostas corre o risco de ter um lucro esperado negativo se a probabilidade implícita de qualquer resultado for menor do que a probabilidade real. Assumindo $\tilde{L}$ como o lucro da casa de apostas, w_i o número total de unidades apostadas no resultado i e p_i como a probabilidade real do evento.

$$E(\tilde{L}) = E(\text{apostas}) - E(\text{pagamentos}) = \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} \frac{w_i p_i}{\pi_i} \tag{2.30}$$

Sendo $p_k > \pi_k$ para um evento $k \in \{1, ..., n\}$ e assumindo uma única aposta de 1 unidade no evento.

$$E(\tilde{L}) = E(\text{apostas}) - E(\text{pagamentos}) = 1 - \frac{p_k}{\pi_k} < 0 \tag{2.31}$$

Thorp (2008) define o problema fundamental das apostas como a busca por oportunidades com valor esperado de retorno financeiro positivo. Assim, ao contrário da casa de apostas, o apostador deve buscar oportunidades onde a probabilidade implícita das *odds* seja menor que a probabilidade real do evento para obter resultados lucrativos. Essas oportunidades são definidas como Apostas de Valor (WALSH; JOSHI, 2024). Nessa situação, a Eq. (2.32) define o valor esperado do lucro do apostador (L) em um evento esportivo com 2 possíveis resultados (HUBACEK; SOUREK; ZELEZNY, 2019).

$$E(L) = (O_i - 1)p_i + (-1)(1 - p_i) = p_i O_i - 1 \tag{2.32}$$

2.3.3 Critério de Kelly

Kelly (1956) utiliza o exemplo de um apostador que, com base no conhecimento dos símbolos recebidos por um canal de comunicação, faz apostas lucrativas nos símbolos transmitidos com o objetivo de alcançar um crescimento exponencial de seu capital inicial. Ele apresenta essa situação como um exemplo do mundo real que incorpora as características fundamentais de um canal de comunicação. A partir disso, o autor traça paralelos entre a teoria da comunicação e essa situação prática, para demonstrar uma teoria de alocação de capital otimizada que maximiza o valor esperado do logaritmo do capital do apostador $\mathrm{E}(\log X)$ a cada aposta.

Para demonstrar a utilidade da estratégia proposta por Kelly (1956), consideremos um apostador com capital inicial X_0 que participa de n apostas repetidas, todas com valor esperado positivo, ou seja, $p > \pi$, onde $q = 1 - p$. Suponha também que a *odd* para essas apostas seja igual a 2, de modo que, a cada aposta vencedora, o retorno para o apostador seja $2 \cdot B_i$, onde B_i é o valor apostado na entrada $i \in \{1, ..., n\}$. O lucro do apostador pode então ser representado por $B_i \cdot T_i$ onde $T_i = 1$ em casos de vitória e $T_i = -1$ em casos de derrota. Assim, podemos partir para as equações Eq. (2.33), Eq. (2.34) e Eq. (2.35) (THORP, 2008).

$$X_i = X_{i-1} + T_i B_i \tag{2.33}$$

$$X_n = X_0 + \sum_{i=1}^{n} T_i B_i \tag{2.34}$$

$$E(X_n) = X_0 + \sum_{i=1}^{n} E(T_i B_i) = X_0 + \sum_{i=1}^{n} (p - q) E(B_i) \tag{2.35}$$

Assim, considerando que estamos lidando com apostas de valor com odd 2, temos que $\pi \geq 0.5$, que implica que $p - q > 0$. Para maximizar $E(X_n)$, é necessário maximizar $E(B_i)$ em cada aposta. Isso nos leva à estratégia de apostar todo o capital disponível em cada entrada, levando a um ganho de capital exponencial. Porém, com $p < 1$, a falência ($X_i = 0$) é praticamente garantida. Outra abordagem seria adotar um modelo que visa minimizar o risco de falência, apostando o menor valor possível em cada entrada. Embora essa estratégia reduza significativamente o risco, ela também resulta na minimização do valor esperado de lucro (THORP, 2008).

Então, para maximizar $E(X_n)$ e minimizar a probabilidade de falência, o critério de Kelly se apresenta como uma alternativa ótima. Para aplicar esta teoria de alocação de capital, a cada aposta iremos alocar uma fração f do capital disponível, de modo que $B_i = fX_{i-1}$, sendo $0 \leq f \leq 1$. Nesta situação, a falência ($X_i = 0$) não ocorrerá, porém para adequar a cenários reais deve-se considerar falência quando o valor de X_i for pequeno o suficiente para ser um valor econômico não significante para uma aposta (THORP, 2008).

Na estratégia de Kelly (1956), o objetivo é maximizar o valor esperado do coeficiente de taxa de crescimento $g(f)$, conforme definido na Eq. 2.36. Para um número fixo de entradas n, isso equivale a maximizar $E(logX_n)$. O valor ideal para alcançar essa maximização é dado por $f_i = p_i - q_i$, em situações onde o lucro é igual ao valor apostado, ou seja, para apostas com *odd* igual a 2. (THORP, 2008).

$$g(f) = p \cdot \log(1 + f) + q \cdot \log(1 - f) \tag{2.36}$$

Thorp (2008) expande a aplicação do critério de Kelly para situações em que o lucro gerado pela vitória da aposta não é igual ao valor apostado. Com isso, ele chega na Eq. 2.37, que define a fração f do capital do apostador que deve ser alocada em cada aposta para maximizar $E(\log X_n)$. Na Equação 2.37, b é o fator multiplicador que gera o lucro do apostador na entrada, definido por $b_i = O_i - 1$.

$$f = \frac{(bp - q)}{b} \tag{2.37}$$

Embora seja considerada uma estratégia ideal de alocação de capital para cenários com apostas repetidas de valor esperado positivo, o critério de Kelly apresenta algumas limitações, sendo a principal delas as apostas de alto valor que o modelo pode sugerir. Em contextos esportivos, onde as probabilidades frequentemente não são bem definidas,

a alocação de uma parcela significativa do capital em uma única aposta pode expor o apostador ao risco de falência (HSIEH; BARMISH, 2015).

Assim, a estratégia fracionada de Kelly surge como uma alternativa mais adequada. Nesse modelo, representado pela Eq. 2.38, é introduzida um constante k, onde $0 < k < 1$. Essa constante tem o objetivo de fracionar o valor sugerido pelo critério de Kelly, reduzindo assim o risco de falência ao diminuir a quantidade de capital alocada em cada aposta (HSIEH; BARMISH, 2015).

$$f = \frac{(bp - q)}{b} \cdot k \tag{2.38}$$

2.4 Engenharia Financeira

A Engenharia Financeira desempenha um papel crucial no desenvolvimento de estratégias que combinam matemática, estatística e computação para explorar ineficiências de mercado. Independentemente do foco da aplicação, o objetivo principal dessa disciplina é extrair informações relevantes de sinais observados e coletados, que, apesar de conterem ruídos e aparentarem ser aleatórios, carregam dados valiosos (AKANSU; TORUN, 2015).

Um conceito fundamental na engenharia financeira é a Hipótese dos Mercados Eficientes, que afirma que o preço atual de um ativo reflete todas as informações disponíveis sobre seu valor e os fatores que o influenciam. Dessa forma, segundo essa teoria, não seria possível obter lucro consistentemente com base em novas informações que venham a surgir, pois estas já estariam instantaneamente incorporadas ao preço do ativo (CLARKE; JANDIK; MANDELKER, 2001).

Considera-se que um mercado é eficiente quando seus preços se ajustam rapidamente e de forma imparcial às novas informações, refletindo, a qualquer momento, todos os dados disponíveis sobre os ativos (CLARKE; JANDIK; MANDELKER, 2001). Diante desse cenário, a ineficiência, no contexto deste estudo, é definida como a incapacidade dos preços de incorporar de maneira completa e precisa todas as informações relevantes, gerando oportunidades para que investidores explorem essas discrepâncias. Em mercados ineficientes, as informações podem ser interpretadas erroneamente, processadas com lentidão ou não serem incorporadas de forma imediata aos preços.

A Hipótese dos Mercados Eficientes propõe que, caso o preço de um ativo se torne hipervalorizado, possivelmente devido a compras sucessivas de investidores não sofisticados ou irracionais, o ativo deixará de ser uma boa opção de compra, pois seu preço não refletirá mais seu valor e risco reais. Investidores bem informados reconheceriam essa discrepância e poderiam realizar operações de arbitragem, explorando a diferença entre o preço atual e o valor fundamental do ativo. As vendas decorrentes dessas operações de

arbitragem contribuiriam para corrigir o preço, ajustando-o ao seu valor real (SHLEIFER, 2000).

Nesse contexto, a arbitragem pode ser definida como a compra e venda simultânea de ativos em diferentes mercados, visando aproveitar discrepâncias de preços para o mesmo ativo. Já na arbitragem estatística, as decisões de compra e venda são fundamentadas no cálculo do *spread* do preço do ativo, utilizando modelos matemáticos, medições precisas e análises estatísticas avançadas. Entre as principais características da arbitragem estatística estão as operações de curto prazo, o alto volume de ativos negociados, os períodos reduzidos de retenção e a necessidade de uma infraestrutura computacional robusta (AKANSU; TORUN, 2015).

3 Metodologia

Este capítulo tem como objetivo detalhar em seções, a metodologia utilizada para o desenvolvimento da pesquisa. Desde a aquisição de dados até a validação dos resultados, e também as tecnologias e ferramentas utilizadas para a realização do trabalho.

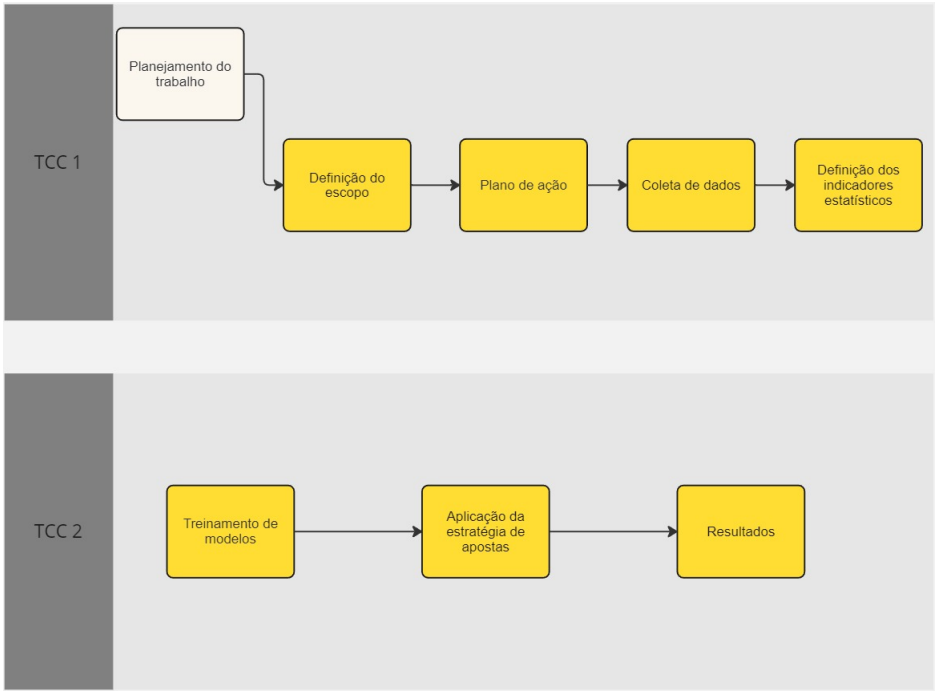
As etapas da metodologia estão organizadas em:

- Planejamento da Pesquisa;
- Referências Bibliográficas;
- Plano de ação.

3.1 Planejamento da Pesquisa

Sendo este um trabalho de conclusão de curso, seu planejamento foi alinhado com as principais entregas ao longo do desenvolvimento da pesquisa. A sequência das etapas planejadas para esta pesquisa está ilustrada na Figura 7.

Figura 7 – Fluxograma de planejamento.



Fonte: Autor.

A primeira etapa do projeto é focada em definir os pontos fundamentais do trabalho, como o planejamento, a definição do escopo e do fluxo de ação, a coleta inicial dos

dados que serão utilizados e, por fim, a definição do indicador estatístico que servirá como alvo das predições.

A segunda etapa do estudo abrange a criação das bases de dados que serão utilizadas no treinamento dos modelos preditivos, o treinamento dos próprios modelos, a análise e aplicação dos resultados obtidos nas estratégias de apostas e, por fim, a apresentação dos resultados alcançados.

3.2 Referências Bibliográficas

Para estabelecer uma base teórica sólida para este estudo, foi realizada uma revisão bibliográfica abrangente, com o objetivo de fundamentar as discussões e análises que serão apresentadas. Além disso, foram pesquisados artigos e estudos que adotam metodologias similares, permitindo a comparação de resultados e propostas desenvolvidas em trabalhos anteriores.

Inspirado pela pesquisa de Hubacek, Sourek e Zelezny (2019), este estudo adota uma abordagem semelhante, porém com um foco específico na comparação das probabilidades oferecidas por diferentes casas de apostas e na integração dessas probabilidades com as estimativas geradas pelos modelos preditivos. Três artigos foram essenciais para servir de alicerce na condução desta pesquisa, sendo eles:

- Hubacek, Sourek e Zelezny (2019) onde é realizado um estudo sobre sistemas de predição projetados para lucrar com o mercado de apostas esportivas utilizando *machine learning*. Além disso, são utilizados elementos da teoria moderna de portfólio para projetar uma estratégia de distribuição de apostas. Este artigo serve como a principal fonte de inspiração para este estudo e fornece a base para grande parte da metodologia aplicada neste trabalho.
- Stekler, Sendor e Verlander (2010), neste trabalho os autores buscam extrair informações sobre as previsões e o processo de previsão a partir de estudos que, testaram a eficiência econômica do mercado de apostas esprotivas. Dessa forma, é analisada a precisão de diversos tipos de modelos de predição de eventos esportivos, incluido: modelos de aprendizado de máquina, opinião de especialistas e o próprio mercado de apostas.
- Walsh e Joshi (2024), propõem o uso da calibração como métrica para a avaliação e otimização de modelos preditivos voltados à previsão de resultados de partidas da NBA. O estudo conclui que modelos otimizados por meio da calibração produzem probabilidades mais precisas para identificar apostas de valor, resultando em estratégias de apostas mais lucrativas. Esse achado é extremamente relevante para o método de avaliação dos modelos preditivos analisados ao longo desta pesquisa.

Os três artigos mencionados anteriormente foram fundamentais para a construção e inspiração deste trabalho. No entanto, outros artigos, não destacados aqui, também serviram como base para a fundamentação teórica, além de terem contribuído para a metodologia, as ideias e a organização do estudo.

Sendo assim, o diferencial deste trabalho está na comparação, em tempo real, das *odds* disponibilizadas por diferentes casas de apostas com as probabilidades preditas pelos modelos de inteligência artificial, visando à elaboração de uma estratégia de apostas otimizada. Essa estratégia tem como objetivo identificar oportunidades com valor esperado positivo e, em seguida, avaliar os resultados financeiros gerados por sua aplicação.

3.3 Plano de ação

A partir das referências estudadas e das informações obtidas, um plano de ação com as seguintes etapas foi elaborado:

1. Coleta de dados
2. Definição do indicador estatístico alvo das predições.
3. Desenvolvimento do algoritmo reponsável pela estratégia de apostas
4. Implementação dos modelos preditivos
5. Aplicação da estratégia de apostas e avaliação de resultados

3.3.1 Coleta de dados

O primeiro passo do plano de ação é a coleta dos dados que serão utilizados ao longo do trabalho. Para isso, foram obtidos dados estatísticos dos jogadores e times da NBA a partir do site *stathead* [1], do provedor *Basketball-Reference*[2], uma enciclopédia de basquete que compila dados históricos das principais ligas do mundo. Esse site disponibiliza informações detalhadas sobre jogadores, técnicos e equipes ao longo de toda a história da NBA (KUBATKO et al., 2007). A fim de obter o maior *dataset* possível, foram coletados dados de times e jogadores de todas as temporadas disponíveis da NBA, de 1983 até 2024.

Para a coleta destes dados foi utilizada a técnica de *web-sraping* com auxílio de *scripts Python* e da biblioteca *Beautiful Soup*, assim sendo possível extrair os dados do site em formato HTML e armazená-los em formato CSV. Dessa forma, já com os dados em um formato semiestruturado se tornou viável fazer a limpeza e adequação destes dados,

[1] https://stathead.com/
[2] https://www.basketball-reference.com/

utilizando a biblioteca *Python* de manipulação de dados: Pandas. Nesta etapa foram eliminados do *data set* colunas que não possuiam uma aplicação para o estudo e foram adicionadas colunas que facilitariam a manipulação destes dados em momentos futuros do trabalho. Em seguida, foi criada uma base de dados SQL com auxílio da ferramenta SQLite, e o arquivo CSV foi convertido em uma tabela nesta base de dados.

Já para os dados referentes às *odds* disponibilizadas pelas casas de aposta online referentes as partidas da NBA, foi utilizado o site *Sportsbook Reviews Online* [3]. As técnicas e estratégias de extração de dados empregadas foram as mesmas utilizadas na coleta dos dados estatísticos históricos da NBA.

3.3.2 Definição do indicador estatístico alvo das predições

Após a coleta dos dados, o próximo passo necessário para o andamento do estudo foi a definição do indicador estatístico que será o alvo das predições feitas pelos modelos de *machine learning*. Nesse contexto, indicadores estatísticos podem ser entendidos como as estatísticas de times e jogadores em uma determinada partida, como por exemplo: quantidade de pontos marcados por um jogador em uma partida, quantidade total de pontos marcados pelas duas equipes em uma partida, vitória de uma determinada equipe e outros.

Assim, tornou-se fundamental realizar uma análise aprofundada das referências bibliográficas nesta etapa, com o objetivo de identificar os indicadores mais utilizados em trabalhos semelhantes, a fim de utilizá-los como comparativos para os resultados obtidos. Além disso, é importante que sejam selecionados os indicadores com maior potencial econômico, garantindo que as análises sejam não apenas precisas, mas também relevantes do ponto de vista dos resultados financeiros.

3.3.3 Desenvolvimento do algoritmo reponsável pela estratégia de apostas

Será desenvolvida uma estratégia de apostas voltada para identificar e explorar ineficiências do mercado. Essa estratégia consistirá em comparar as *odds* de diferentes casas de apostas para detectar discrepâncias nas probabilidades implícitas, além de combinar as *odds* de determinadas casas com as probabilidades $\hat{p}_i$ geradas pelos modelos de *machine learning*. O objetivo é calcular uma probabilidade ajustada $\hat{P}_i$ que, idealmente, esteja mais próxima do valor de fechamento do evento i. Com base nessa probabilidade $\hat{P}_i$ e nas discrepâncias identificadas entre as casas de apostas, serão determinadas oportunidades de apostas de valor, conforme apresentado na Eq. 3.1, onde L_i representa o lucro do apostador no evento.

[3] https://www.sportsbookreviewsonline.com/

$$E(L_i) = \hat{P}_i O_i - 1 \tag{3.1}$$

A descrição da metodologia de todas as etapas da aplicação da estratégia esta definida nas seguintes seções:

3.3.3.1 Definição das casas de apostas

Uma etapa fundamental da estratégia definida é a seleção das casas de apostas, tanto aquelas que fornecerão as *odds* de abertura para o cálculo de $\hat{P}$, quanto aquelas que serão utilizadas para identificar oportunidades de apostas de valor. Em cada um desses casos, diferentes características das casas de apostas são relevantes.

Para as casas utilizadas no cálculo de $\hat{P}$, é crucial que as *odds* sejam precisas e que os valores de abertura representem uma boa estimativa da probabilidade real do resultado do evento. Essas casas geralmente aplicam margens menores em suas *odds*, reconhecendo o poder de precisão das probabilidades implícitas e, assim, conseguem atrair clientes com *odds* mais competitivas.

Por outro lado, para as casas onde serão buscadas as oportunidades de valor, o oposto é desejável. Quanto mais distante a odd oferecida estiver da probabilidade real do evento, maior será a chance de identificar uma aposta de valor.

A metodologia adotada para essa escolha incluirá a revisão de referências bibliográficas, pesquisas em sites de apostas e a verificação da disponibilidade das *odds* para coleta. Com base nesses critérios, as casas de apostas serão separadas em dois grupos: as precisas, que serão utilizadas para o cálculo da probabilidade, e as não precisas, onde serão buscadas as oportunidades de valor esperado positivo.

3.3.3.2 Coleta das *odds*

A coleta das *odds* para a estratégia será realizada por meio de um código hospedado na nuvem, utilizando a plataforma Heroku. Este *software* será responsável por executar constantemente uma tarefa para coletar as *odds* de todas as casas de apostas definidas na etapa anterior. Essa coleta ocorrerá ao longo dos meses de outubro a fevereiro, correspondendo, respectivamente, ao início da temporada 2024-2025 da NBA e ao prazo final para a conclusão deste trabalho.

Após a coleta, as *odds* serão classificadas em precisas e não precisas, conforme estabelecido na primeira etapa, com o objetivo de armazenar esses dados em uma base de dados.

3.3.4 Implementação dos modelos preditivos

A implementação dos modelos preditivos é um passo crucial para o desenvolvimento do trabalho. Assim, esta etapa foi subdividida três passos:

3.3.4.1 Seleção dos modelos utilizados

Para a implementação dos modelos preditivos, foi essencial iniciar pela seleção dos modelos a serem utilizados. Considerando que o objetivo da predição é gerar valores contínuos no intervalo [0, 1], representando a probabilidade de ocorrência de um determinado evento, optou-se pela classe de modelos de regressão logística como a solução mais adequada para o problema (HUBACEK; SOUREK; ZELEZNY, 2019).

Utilizando como base metodológica o artigo de Hubacek, Sourek e Zelezny (2019), foi selecionada uma rede neural artificial com apenas um neurônio, fazendo uso da função sigmoide como função de ativação, para servir como linha de base das previsões. Além disso, devido à natureza cronológica dos dados utilizados no treinamento, optou-se por uma Rede Neural Recorrente (RNN) para a predição dos resultados, especificamente o modelo *Long Short-Term Memory* (LSTM) (SARKER, 2021).

3.3.4.2 Escolha da métrica de avaliação

A escolha da métrica de avaliação foi orientada pelo objetivo final das predições dos modelos selecionados. Considerando o contexto deste estudo, ficou evidente que o foco principal deve ser nas probabilidades previstas pelos modelos, que serão comparadas com as probabilidades oferecidas pelas casas de apostas.

Dessa forma, uma alta acurácia ou uma boa análise da área sob a curva ROC podem não ser suficientes para garantir a precisão das probabilidades previstas pelo modelo. Neste contexto, a calibração se mostra uma métrica mais adequada (NICULESCU-MIZIL; CARUANA, 2005). A aplicação dessa métrica no cenário das apostas esportivas foi avaliada no estudo de Walsh e Joshi (2024), onde os modelos que utilizaram a calibração como métrica de avaliação apresentaram resultados com retornos econômicos 664% melhores em comparação aos modelos que se basearam apenas na acurácia.

Diante das informações apresentadas, este estudo optou por utilizar a métrica de calibração para a avaliação dos modelos preditivos, considerando sua importância tanto para o treinamento quanto para a otimização dos modelos, por meio da seleção de características e do ajuste fino dos hiperparâmetros.

3.3.4.3 Treinamento e teste dos modelos

Essa etapa consistirá da implementação dos modelos preditivos em si, com o desenvolvimento de um código com axílio das ferramentas: *Jupyter Notebooks*, como ambiente

de desenvolvimento python; *Tensor Flow*, biblioteca Python que facilita a criação de modelos de *machine learning* com funções e arquiteturas pré-definidas; *Pandas*, biblioteca python para manipulação dos datasets de treino e de teste.

O desenvolvimento dos modelos preditivos envolve uma série de passos que serão seguidos, incluindo:

- **Definição do *dataset* utilizado**: será necessário definir a quantidade de temporadas que serão utilizadas para o treinamento e teste do modelo;
- **Subdivisão da base de dados**: a base de dados deverá ser subdividida em dados para treinamento, validação e teste;
- **Escolha das características**: as características que servirão como entrada para os modelos preditivos deverão ser selecionadas;
- **Ajuste fino dos hiperparâmetros**: deverão ser escolhidas estratégias e algoritmos para o ajuste dos hiperparâmetros dos modelos;
- **Análise dos resultados**: as saídas dos modelos deverão ser analisadas tendo como base as métricas de avalição selecionadas;

3.3.5 Aplicação da estratégia de apostas e avaliação de resultados

Por fim, com as *odds* das casas de apostas precisas armazenadas e as probabilidades geradas pelos modelos preditivos definidas, será possível calcular a probabilidade $\hat{P}$. Para a obtenção da probabilidade implicíta normalizada $\tilde{\pi}$ a partir das *odds* armzenadas será utilizada a Eq. 3.2, onde π é a probabilidade implícita da odd e r a soma das probabilidades implícitas dos possíveis resultados do evento.

$$\tilde{\pi} = \frac{\pi}{r} \tag{3.2}$$

Assim, com o cálculo de $\tilde{\pi}$, teremos duas probabilidades para o mesmo evento esportivo: $\tilde{\pi}$, obtida a partir das *odds* das casas de apostas precisas, e $\hat{p}$, gerada pelos modelos preditivos. O objetivo desta estapa é explorar métodos para combinar essas duas probabilidades em uma única $\hat{P}$.

3.3.5.1 Buscas por oportunidades de apostas

Com a probabilidade $\hat{P}$ definida, será realizada uma busca entre as *odds* das casas de apostas não precisas utilizando a Eq. 3.1 para encontrar apostas de valor. Quando uma aposta apresentar $EV(L)$ positivo, os dados referentes a esta aposta serão armazenados.

Também será criado um *dataset* de controle com resultados baseados em apostas realizadas exclusivamente com o uso de $\tilde{\pi}$ para identificar oportunidades de valor, definindo o valor esperado conforme a Eq. 3.3. Esse *dataset* permitirá a comparação entre os resultados obtidos com as probabilidades preditas pelos modelos de *machine learning* e aqueles obtidos utilizando apenas as probabilidades calculadas pelas casas de apostas.

$$E(L_{c_i}) = \tilde{\pi}_i O_i - 1 \tag{3.3}$$

Além disso, será calculada a quantidade de capital a ser alocada em cada aposta utilizando o critério de Kelly fracionado identificado pela Eq. 3.4. Durante a avaliação do resultado da estratégia, o desempenho econômico dos seguintes valores de k serão testados e comparados: 1, 0.5, 0.25 e 0.125 (DOTAN; LEW; LIU, 2020).

$$f = \frac{(bp - q)}{b} \cdot k \tag{3.4}$$

Para avaliar os resultados obtidos, serão utilizadas duas métricas fundamentais:

- **Retorno sobre Investimento (ROI)**: definido pela Eq. 3.5, onde X_0 representa o capital inicial utilizado para realizar n apostas e X_n corresponde ao capital final após as apostas, permitindo quantificar a rentabilidade da estratégia aplicada.

$$ROI = \frac{X_n}{X_0} \cdot 100 \tag{3.5}$$

- **Volatilidade** (*vol*): métrica que mede a dispersão do capital ao longo do tempo, indicando a magnitude das flutuações, tanto para cima quanto para baixo. Essa medida é calculada conforme a Eq. 3.6, onde σ representa o desvio padrão das variações do capital a cada dia e T corresponde ao número de dias apostados.

$$vol = \sigma\sqrt{T} \tag{3.6}$$

4 Resultados

Este capítulo apresenta a aplicação do plano de ação da pesquisa aos resultados obtidos.

4.1 Coleta de dados

Para a coleta dos dados estatísticos históricos da NBA, necessários para o treinamento dos modelos preditivos, foi primeiramente necessário definir a fonte desses dados. Para isso, foi realizada uma pesquisa nas referências bibliográficas coletadas, focando em estudos com escopo semelhante a este.

Entre os artigos selecionados, Walsh e Joshi (2024) e Kubatko et al. (2007) destacam o portal *Basketball-Reference*[1], um site que compila dados estatísticos detalhados de jogadores em partidas da NBA desde a temporada de 1983. Ao analisar o site, foi encontrada uma seção específica de dados de jogadores em partidas sob o nome *Stat Head*[2], lá observou-se que esta seção apresenta características que facilitam a coleta de dados por meio de *web scraping*, como a organização dos dados em tabelas sequenciais. Portanto, por se tratar de uma fonte confiável e por permitir a coleta automatizada de dados, o site *Stat Head*, do provedor *Basketball-Reference*, foi escolhido como a fonte dos dados históricos da NBA (KUBATKO et al., 2007).

Com o provedor de dados definido, foi possível desenvolver o *web scraper* responsável por coletar os dados do site e armazená-los em uma base de dados apropriada. Para isso, foi criado um código em Python que envia uma série de requisições ao site *Basketball-Reference*, obtendo a estrutura HTML do site como resposta, conforme ilustrado na Fig. 8. Em seguida, utilizando a biblioteca *Beautiful Soup*, foi possível extrair e armazenar as linhas das tabelas de estatísticas, identificando-as pelo seu identificador comum "*stats*", em uma estrutura de dados iterável, conforme ilustrado na Fig.9. Assim, foi possível armazenar os dados presentes nestas tabelas em um arquivo CSV, com auxílio da biblioteca Pandas.

A partir disso, um novo código em Python foi desenvolvido para organizar e limpar as informações armazenadas no arquivo CSV. Utilizando novamente a biblioteca *Pandas* como ferramenta, colunas foram renomeadas, novas colunas foram adicionadas e os valores foram normalizados, com o objetivo de facilitar a manipulação dos dados para o treinamento dos modelos preditivos. Para exemplificar algumas destas mudanças, a coluna com

[1] https://www.basketball-reference.com/
[2] https://stathead.com/

Figura 8 – Requisição para o site *Stat Head*.

```
url = f"https://stathead.com/basketball/player-game-finder.cgi
↪ request=1&player_game_min=1&team_game_min=1&comp_type=E&
↪ order_by=date&match=player_game&season_start=1&player_game_max
↪ =9999&year_max=2024&team_game_max=84&season_end=-1&comp_id=NBA
↪ &year_min=1947&offset{offSet}"

response = requests.get(url)
```

Fonte: Autor

Figura 9 – Utilização da biblioteca *Beautiful Soup*.

```
soup = BeautifulSoup(response.text, 'html.parser')

rows = soup.find('table', {'id': 'stats'}).find('tbody').find_all('tr')
```

Fonte: Autor

nome `+/-` foi renomeada para `plusMinus`, os nomes dos times foram substituidos por identificadores únicos, e foram criadas as colunas `homeTeam` e `awayTeam` para identificar time da casa e o time visitante, respectivamente.

Após a organização dos dados coletados, foi necessária a criação de uma base de dados para armazenar essas informações. Para esse propósito, foi utilizada a ferramenta SQLite. Com o auxílio de um script Python, foi criada a base de dados denominada `dadosnba`, e a partir do arquivo CSV, foi criada uma tabela chamada `dadosJogadoresPartidas`. Para coletar metadados sobre a base criada, algumas consultas SQL foram executadas, cujos resultados estão apresentados na Tab. 2.

Tabela 2 – Metadados obtidos após consultas

Nome da Tabela	Número de Linhas	Número de Colunas
dadosJogadoresPartidas	1.036.489	40

Fonte: Autor.

4.2 Definição do indicador estatístico alvo das predições

No contexto deste trabalho, indicadores estatísticos são medidas acumuladas por jogadores e times durante uma partida, como: quantidade de pontos marcados, número de passes, equipe vencedora do confronto, entre outros. Para a escolha do indicador estatístico

a ser previsto pelos modelos de *machine learning*, uma das restrições estabelecidas foi que o indicador deveria ter um mercado de apostas associado. Além disso, um fator determinante para a seleção do indicador foi a existência de estudos anteriores a este que utilizassem o mesmo indicador, permitindo assim a comparação dos resultados ao final do trabalho.

Nesse contexto,Walsh e Joshi (2024) destacam o mercado de vencedor da partida como a forma mais simples de testar suas hipóteses e modelos preditivos. Este mercado também é foco dos estudos de Hubacek, Sourek e Zelezny (2019), Cao (2012), Torres (2013) e Dotan, Lew e Liu (2020), o que o torna o indicador mais comum nas referências analisadas para este trabalho.

Com base nessa análise, optou-se por utilizar o indicador de vencedor da partida como o alvo das predições dos modelos de aprendizado de máquina neste estudo. Assim, o objetivo das predições será determinar a probabilidade de vitória de um time em uma partida de basquete da NBA. No contexto do mercado de apostas, este indicador é estabelecido como o mercado de vencedor em que, para ganhar uma aposta o apostador deve prever o vencedor do jogo. Se o apostador prever corretamente o vencedor, ele recebe de volta o valor apostado juntamente com o lucro; caso contrário, ele perde a aposta para a casa de apostas (HUBACEK; SOUREK; ZELEZNY, 2019).

4.3 Desenvolvimento do algoritmo de apostas

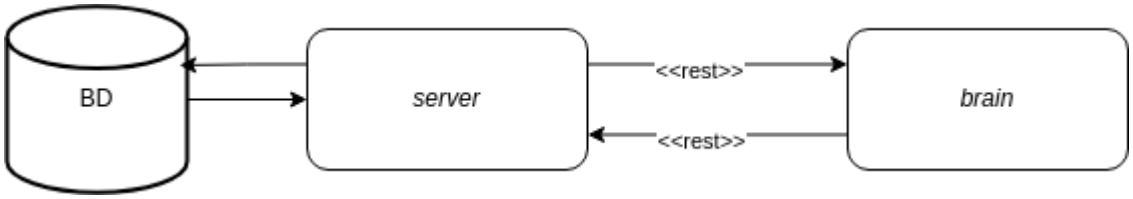
Nesta seção, será detalhado o processo de criação do algoritmo responsável por identificar e explorar ineficiências no mercado de apostas esportivas. Serão abordadas as etapas de estruturação do sistema, a escolha da arquitetura, os desafios encontrados e as soluções implementadas para garantir a eficiência na coleta, análise e comparação das *odds* e probabilidades geradas pelos modelos preditivos.

4.3.1 Arquitetura da solução

O primeiro passo para o desenvolvimento do algoritmo foi a estruturação dos serviços necessários para a execução das tarefas envolvidas. Dessa forma, a etapa inicial consistiu na definição dos serviços, suas respectivas responsabilidades e a forma como interagiriam entre si, garantindo um fluxo eficiente e organizado do sistema.

Nesse contexto, optou-se pela arquitetura de microsserviços, conforme ilustrado na Fig. 10, composta por dois serviços principais: Server, responsável pela coleta das *odds* e pelo armazenamento das apostas de valor no banco de dados; e Brain, desenvolvido em Python, encarregado de analisar as *odds* coletadas e identificar oportunidades de apostas com valor esperado positivo. Essa abordagem arquitetural foi escolhida para modularizar as diferentes lógicas do sistema, promovendo maior escalabilidade e permitindo a adição de novas funcionalidades de forma eficiente e estruturada em trabalhos futuros.

Figura 10 – Diagrama relacional.



Fonte: Autor.

Com o objetivo de evitar a ociosidade do serviço Brain e maximizar a identificação de oportunidades de apostas, foi implementado um sistema de *webhooks* para a comunicação entre os serviços. Nesse sistema, o Server busca as *odds* nas casas de apostas e as envia para o Brain, que processa os dados, identifica oportunidades de apostas de valor e retorna uma resposta ao servidor. Assim que recebe essa resposta, o Server armazena as informações no banco de dados e reinicia o ciclo, buscando novas *odds* para enviar novamente ao Brain.

Ambos os serviços foram implantados na plataforma de computação em *cloud* Heroku, permitindo que a comunicação entre eles ocorra de forma contínua por meio de requisições https no padrão REST.

4.3.2 Serviço Server

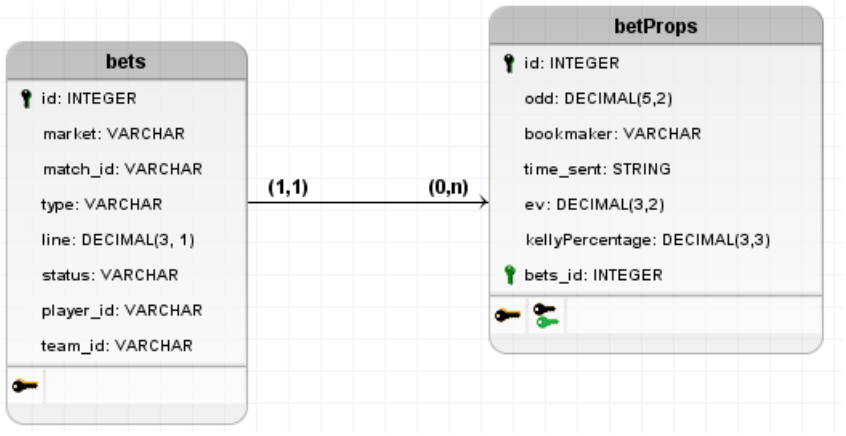
Como mencionado anteriormente, este serviço é responsável por coletar as *odds* de todas as casas de apostas utilizadas no estudo, enviá-las ao *Brain* para processamento e armazenar as respostas recebidas em um banco de dados, garantindo a persistência das informações. Além disso, desempenha um papel fundamental na classificação das apostas, categorizando-as como ganha, perdida ou valor retornado, este último aplicável em casos de empate.

Para a implementação dessa solução, foram escolhidas as seguintes tecnologias: *Node.js*, juntamente com a linguagem *TypeScript*, para o desenvolvimento do servidor encarregado do envio e recebimento de requisições, e *PostgreSQL* como banco de dados relacional, responsável pelo armazenamento das apostas de valor identificadas.

Para o armazenamento dessas apostas no banco de dados SQL, foi necessária a criação de duas tabelas: `Bets` e `BetProps`. Essas tabelas apresentam um relacionamento do tipo um-para-muitos (1:n), onde uma `Bet` pode estar associada a uma ou mais `BetProps`, enquanto cada `BetProp` está vinculada a uma única `Bet`, conforme ilustrado no diagrama lógico de dados representado na Figura 11.

Essa estrutura relacional foi definida devido à possibilidade de existência de múltiplas oportunidades de apostas de valor em diferentes casas de apostas para um mesmo mercado de jogador ou time. Essa modelagem possibilita uma organização mais eficiente

Figura 11 – Diagrama lógico de dados.



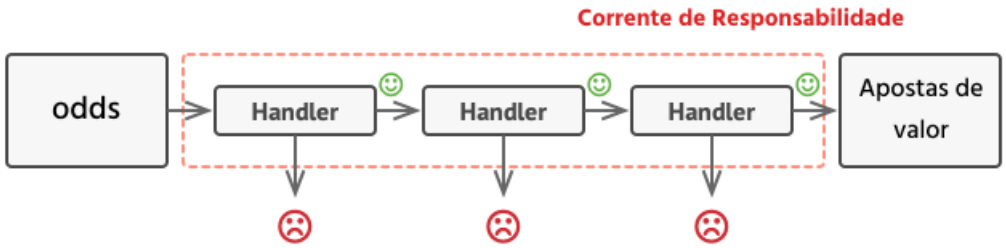
Fonte: Autor.

e garante maior rastreabilidade das apostas de valor armazenadas, permitindo consultas otimizadas e uma melhor integridade dos dados ao longo do estudo.

4.3.3 Serviço Brain

Para o desenvolvimento do algoritmo do Brain, responsável por comparar as *odds* das casas de apostas e identificar oportunidades de valor, foi adotado o padrão de projeto Corrente de Responsabilidade, como mostra a figura 12. Esse padrão comportamental permite que uma requisição seja processada por uma sequência de manipuladores (*handlers*), onde cada um processa a solicitação e decide se retorna um valor vazio ou a encaminha para o próximo na cadeia. Isso garante maior flexibilidade e modularidade no processamento dos dados (SHVETS, 2021).

Figura 12 – Diagrama da corrente de responsabilidade.



Fonte: Shvets (2021), com modificações.

Dessa forma, definiu-se a necessidade de cinco *handlers*, cada um com a seguinte finalidade:

1. Determinar a probabilidade implícita $\tilde{\pi}_i$ dos eventos, removendo a margem (v) das casas de apostas classificadas como precisas.

2. Calcular a probabilidade ajustada $\hat{P}_i$, combinando $\tilde{\pi}_i$ com a probabilidade $\hat{p}_i$ predita pelos modelos de *machine learning*.
3. Comparar as probabilidades $\tilde{\pi}_i$ e $\hat{P}_i$ com as *odds* das demais casas de apostas não precisas, identificando e filtrando ineficiências no mercado.
4. Calcular o valor esperado de lucro do apostador $EV(L)$.
5. Determinar a porcentagem do capital total do apostador (f) a ser alocada na aposta, utilizando o Critério de Kelly.

O desenvolvimento de cada um desses manipuladores envolveu a superação de diversos desafios. Os principais obstáculos enfrentados foram: o tempo de execução do algoritmo em dias com um grande volume de jogos da NBA, onde a alta quantidade de *odds* coletadas gerava gargalos na execução do serviço; e a complexidade na comparação de probabilidades entre diferentes casas de apostas, especialmente ao comparar linhas distintas para um mesmo mercado.

O primeiro desafio se tornava crítico ao comparar as *odds* de diferentes mercados em distintas casas de apostas com as probabilidades $\tilde{\pi}_i$ e $\hat{P}_i$. Para superar esse problema, foram implementadas técnicas de processamento em paralelo, utilizando *threads* e estruturas de dados otimizadas, como *hash maps*.

A solução adotada consistiu em processar simultaneamente as *odds* das casas precisas e não precisas assim que a requisição fosse recebida do *Server*. Durante esse processamento, cada mercado de cada jogo do dia na NBA recebia uma chave única. Para as casas de apostas precisas, foi criado um *hash map* onde a chave de acesso correspondia à chave única do mercado no jogo, e o valor armazenado representava a probabilidade $\tilde{\pi}_i$ no *dataset* de controle e $\hat{P}_i$ no *dataset* que utilizava *machine learning*.

Já para as casas não precisas, foi criado um segundo *hash map*, onde a chave de acesso também correspondia à chave única do mercado para o jogo, porém o valor armazenado era um dicionário Python contendo as *odds* oferecidas por diferentes casas de apostas para aquele mercado. Dessa forma, para cada mercado de cada jogo, torna-se possível um acesso imediato tanto às probabilidades quanto às *odds* que serão comparadas, otimizando a eficiência do processo.

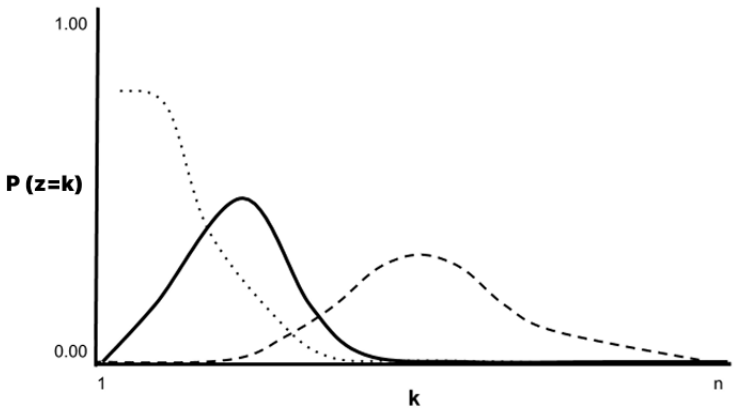
O segundo desafio também surgiu durante o desenvolvimento do *handler* responsável pela comparação das probabilidades com as *odds*, mais especificamente na execução do *dataset* de controle, onde a probabilidade $\tilde{\pi}_i$ era comparada com as *odds* das casas de apostas não precisas. Nesse caso, diferentemente do *dataset* baseado em $\hat{P}_i$, gerado pelos modelos de inteligência artificial e restrito ao mercado de vencedor da partida, não havia essa limitação. Isso permitiu a inclusão de mercados adicionais, como linhas estatísticas de jogadores e times da NBA.

O problema ocorria quando, para um mesmo mercado de um jogador ou time, as casas de apostas precisas utilizavam linhas diferentes das casas não precisas. Isso impossibilitava a comparação direta entre $\tilde{\pi}_i$ e as *odds* das casas não precisas, pois representavam probabilidades associadas a eventos distintos.

Para ilustrar esse cenário, podemos considerar o mercado de pontos de um jogador em uma partida. Enquanto uma casa de apostas precisa calcula as *odds* para o jogador marcar mais ou menos de x pontos, uma casa de apostas não precisa pode oferecer *odds* para esse mesmo jogador marcar mais ou menos de y pontos, onde $x \neq y$. Essa discrepância nas linhas inviabiliza uma comparação direta e exige um ajuste metodológico para garantir a consistência da análise.

Esse ajuste foi realizado assumindo que o número de estatísticas acumuladas por um jogador ou time de basquete ao longo de uma partida segue uma distribuição de Poisson, representada graficamente pela Fig. 13 (GONZáLEZ et al., 2016). Embora essa seja uma simplificação, pois desconsidera pequenas correlações entre estatísticas e interações entre jogadores em quadra, a distribuição de Poisson é amplamente aceita como uma abordagem eficaz e simplificada para modelar estatísticas de desempenho em partidas (KARLIS; NTZOUFRAS, 2003).

Figura 13 – Representação gráfica da distribuição de Poisson.



Fonte: Cowan (2016), com modificações.

Dessa forma, na distribuição de Poisson uma variável aleatória z com parâmetro λ, que representa a média de ocorrências de z em um determinado intervalo de tempo, tem sua distribuição dada pela Eq. (4.1) (ROLLA; LIMA, 2024).

$$pz(k) = \frac{e^{-\lambda}\lambda^k}{k!} \ , \ k \in Z^+ \tag{4.1}$$

No cenário em que diferentes casas de apostas definem linhas distintas para um mesmo mercado, consideremos j como um mercado onde um jogador ou time precisa

atingir uma estatística abaixo da linha x estabelecida por uma casa precisa. Dessa forma, podemos definir $\tilde{\pi}_j$ conforme a Eq. (4.2).

$$\tilde{\pi}_j = e^{\lambda} \sum_{i=0}^{x} \frac{\lambda^i}{i!} \tag{4.2}$$

Como $\tilde{\pi}_j$ e x são valores conhecidos, foi implementada no algoritmo uma função que estima o parâmetro λ por meio de aproximação iterativa, utilizando um método de tentativa e erro até atingir uma precisão mínima de três casas decimais. Com o valor de λ determinado, torna-se possível calcular a probabilidade correspondente à linha y definida pela casa de apostas não precisa e, assim, compará-la com sua *odd* associada.

4.3.4 Definição das casas de apostas

Para selecionar as casas de apostas para este estudo, foram aplicados dois filtros iniciais e essenciais: a casa deve disponibilizar *odds* em tempo real para coleta e operar no Brasil. Com base nesses critérios, as casas foram classificadas em dois grupos: precisas e não precisas.

As casas de apostas precisas são aquelas que apresentam probabilidades altamente eficientes, baseando-se fortemente em ciência de dados e modelos matemáticos avançados. Seu objetivo é oferecer *odds* competitivas com margens reduzidas, garantindo, ainda assim, lucratividade por meio do alto volume de apostas e da precisão na precificação dos eventos (BUCHDAHL, 2021).

Nesse grupo, a principal referência do mercado é a casa Pinnacle, reconhecida por aceitar apostadores lucrativos sem impor restrições, desafiando-os a superar suas precificações. Além disso, a Pinnacle se destaca por aplicar margens extremamente baixas em suas probabilidades, reforçando seu compromisso com a eficiência do mercado (BUCHDAHL, 2020). Por esse motivo, foi escolhida como a casa de apostas precisa de referência, cujas *odds* servirão como base para a identificação de ineficiências no mercado.

As casas de apostas não precisas, também conhecidas como recreativas, têm como principal característica a limitação do valor apostado por usuários que, no longo prazo, demonstram ser lucrativos, conseguindo consistentemente superar a precificação da casa e identificar apostas de valor. Diferentemente das casas precisas, essas plataformas priorizam estratégias de *marketing*, a ampla oferta de mercados e a concessão de bônus para novos apostadores. Para garantir a lucratividade a longo prazo, mesmo com suas ineficiências, essas casas aplicam margens significativamente mais altas em suas probabilidades, explorando a falta de conhecimento dos apostadores menos experientes (BUCHDAHL, 2020).

Dessa forma, para esse grupo, a seleção foi baseada apenas nos filtros iniciais. As

casas de apostas escolhidas para o estudo foram: Bet365, 1XBet, Betano, Sportingbet, Stake, Bwin, Betsson, Superbet, Betnacional e Betfair. Essas serão as plataformas onde as ineficiências de mercado serão buscadas e as apostas serão realizadas.

4.4 Implementação dos modelos preditivos

Nesta seção, são detalhados os principais passos para a implementação dos modelos preditivos utilizados neste estudo. Primeiramente, são abordados os critérios adotados para a seleção dos modelos. Em seguida, descreve-se a definição do *dataset* de treinamento, incluindo a forma como os dados foram estruturados e divididos. Além disso, apresenta-se o processo de seleção das *features*, destacando as variáveis estatísticas utilizadas para alimentar os modelos. Por fim, são discutidos os resultados obtidos, avaliando a eficácia das predições e a adequação dos modelos propostos ao problema analisado.

4.4.1 Seleção de modelos

O primeiro passo para o desenvolvimento dos modelos preditivos foi a seleção dos modelos a serem utilizados no estudo. Para isso, foi realizada uma revisão bibliográfica aprofundada, buscando um ponto de partida já validado em pesquisas anteriores com objetivos semelhantes. Como o objetivo final do modelo é prever probabilidades de vitória no intervalo [0,1], optou-se por modelos de regressão logística, amplamente utilizados para esse tipo de previsão (HUBACEK; SOUREK; ZELEZNY, 2019).

Além dos modelos de aprendizado de máquina, foram avaliadas abordagens alternativas, como Correntes de Markov, que nos trabalhos de Štrumbelj e Vračar (2012) e Vračar, Štrumbelj e Kononenko (2016) demonstraram bons resultados na previsão de partidas de basquete. No entanto, esses estudos utilizam dados jogada-a-jogada, com um nível de detalhamento muito superior às estatísticas agregadas de jogadores por partida disponíveis para este estudo. Dado que não foram encontradas abordagens robustas utilizando estatísticas semelhantes sem recorrer a modelos de aprendizado de máquina, decidiu-se seguir exclusivamente com essa classe de modelos.

Assim, os modelos de *machine learning* selecionados para este estudo foram uma rede neural artificial com um único neurônio e o modelo *Long Short-Term Memory* (LSTM). A escolha metodológica foi baseada no trabalho de Hubacek, Sourek e Zelezny (2019), onde uma rede neural simples é utilizada como linha de base para comparar as probabilidades preditas por um modelo mais avançado. Seguindo essa abordagem, a rede neural com um único neurônio foi implementada com a função de ativação sigmoide, servindo como referência inicial para as predições. Já o modelo LSTM foi escolhido por sua capacidade de capturar padrões temporais, sendo mais adequado para explorar a natureza sequencial dos dados utilizados no treinamento (SARKER, 2021).

4.4.2 Treinamento e testes dos modelos

4.4.2.1 Definição do *dataset* utilizado no treinamento

Durante as etapas anteriores deste trabalho, foram coletadas estatísticas de jogadores em partidas da NBA desde a temporada de 1983 até 2024. Dessa forma, tornou-se necessário definir quais dados seriam utilizados e de que maneira seriam aplicados durante o treinamento e teste dos modelos preditivos. Conforme destacado por Hubacek, Sourek e Zelezny (2019), a temporada atual em relação à predição é a janela de tempo mais relevante para fornecer os dados necessários à previsão de resultados.

Dessa forma, optou-se pelo treinamento do modelo por temporadas, onde, para cada temporada, 80% dos dados foram destinados ao treinamento e 20% à validação do modelo. Essa abordagem garante que apenas os dados da temporada vigente à predição sejam utilizados como referência para os resultados do modelo.

Por fim, para obter uma avaliação mais robusta, as métricas de desempenho dos modelos foram calculadas como a média dos valores obtidos em todas as janelas de tempo testadas.

4.4.2.2 Escolha das características

A abordagem mais comum para a seleção de *features* em modelos desse contexto é o uso direto das estatísticas individuais dos jogadores em partidas. Essas estatísticas são agregadas ao longo da temporada e utilizadas para treinar os modelos, considerando as médias acumuladas até o momento do jogo a ser previsto (HUBACEK; SOUREK; ZELEZNY, 2019). Como o objetivo deste estudo é prever as probabilidades de vitória das equipes na NBA, a primeira etapa essencial foi transformar os dados de jogadores em estatísticas agregadas de times. Para isso, foram utilizadas as funções de manipulação de dados da biblioteca Python Pandas, conforme ilustrado na Figura. 14.

Além disso, a utilização exclusiva de estatísticas básicas de jogadores e times pode não capturar toda a complexidade envolvida em uma partida da NBA. Para mitigar essa limitação, foram incorporadas ao *dataset* estatísticas avançadas, que são métricas derivadas das estatísticas básicas e baseadas no número de posses de bola defensivas, ofensivas e totais de cada equipe na partida. Essas estatísticas oferecem uma visão mais aprofundada do desempenho das equipes e dos jogadores, permitindo uma modelagem mais precisa (KUBATKO et al., 2007). A relação completa das siglas e explicações das estatísticas simples e avançadas utilizadas neste estudo pode ser encontrada no Apêndice A.

Com a base de dados já populada com estatísticas simples e avançadas dos times da NBA, foi possível prosseguir com a manipulação dos dados para a extração das características utilizadas no treinamento dos modelos preditivos. Nesse contexto, conforme

Figura 14 – Função para transformação para estatísticas de equipe.

```
1
2           def transformToTeamStats(df):
3               df['Team_PTS'] = df.groupby(['team_id',
                ↪  'match_id'])['PTS'].transform('sum')
4
5               df['Team_FGA'] = df.groupby(['team_id',
                ↪  'match_id'])['FGA'].transform('sum')
6
7               df['Team_MP'] = df.groupby(['team_id',
                ↪  'match_id'])['MP'].transform('sum')
8
9               df['Team_AST'] = df.groupby(['team_id',
                ↪  'match_id'])['AST'].transform('sum')
10
11              df['Team_FG'] = df.groupby(['team_id',
                ↪  'match_id'])['FG'].transform('sum')
12
13              df['Team_FT'] = df.groupby(['team_id',
                ↪  'match_id'])['FT'].transform('sum')
14
15              df['Team_FTA'] = df.groupby(['team_id',
                ↪  'match_id'])['FTA'].transform('sum')
16
17              df['Team_TOV'] = df.groupby(['team_id',
                ↪  'match_id'])['TOV'].transform('sum')
18
19              df['Team_3P'] = df.groupby(['team_id',
                ↪  'match_id'])['3P'].transform('sum')
```

Fonte: Autor

destacado por Walsh e Joshi (2024), o uso de medidas absolutas das estatísticas pode introduzir um viés nos resultados, tornando o modelo incapaz de se adaptar a mudanças nas dinâmicas da liga ao longo do tempo. Para mitigar esse efeito, foi adotada uma abordagem onde é calculada a diferença entre as estatísticas do time e as de seu adversário em cada partida. Além disso, a média dessas diferenças foi computada ao longo da temporada, permitindo que o modelo utilizasse informações contextualizadas e relativas ao desempenho das equipes até o confronto a ser analisado.

Além das estatísticas simples e avançadas de times, foi incorporada ao modelo a métrica Elo, com o objetivo de fornecer um indicador que ranqueia as equipes ao longo da temporada. O Elo (r), representado pela Eq. (4.3), é uma medida que busca quantificar a performance de um time ao longo das partidas. Sua premissa baseia-se na suposição

de que o desempenho de uma equipe segue uma distribuição normal, cuja média (μ) se ajusta gradualmente ao longo do tempo, refletindo mudanças no nível de competitividade sem sofrer variações abruptas (LANGVILLE; MEYER, 2012).

$$r_{novo} = r_{antigo} + K(S - \mu) \tag{4.3}$$

Para o cálculo do Elo, assume-se que todas as equipes iniciam a temporada com um valor padrão de 1.500, e suas performances são ajustadas progressivamente ao longo das partidas. Após cada jogo, o Elo de um time é atualizado conforme a Eq. (4.3).

No cálculo dessa métrica, considerando um confronto entre o time i e o time j, define-se $S_{ij} = 1$ caso i vença j, e $S_{ij} = 0$ caso i seja derrotado por j. Além disso, a variável μ é determinada pela Eq. (4.4), de forma que o Elo recompensa de maneira mais significativa uma vitória de um time mais fraco sobre um adversário mais forte do que o contrário, garantindo uma atualização proporcional à dificuldade do confronto. Por fim, o valor de K na Eq. (4.3) é definido como uma constante e para esportes entre equipes é convencionado em 32 (LANGVILLE; MEYER, 2012).

$$\mu_{ij} = \frac{1}{1 + 10^{\frac{-d_{ij}}{400}}}, \text{ onde } d_{ij} = r_{i(antigo)} - r_{j(antigo)} \tag{4.4}$$

Dessa forma, foi possível chegar nas *features* finais que foram utilizadas para treinar os modelos preditivos conforme mostra o Quadro 3.

4.4.3 Análise dos resultados

A implementação dos modelos teve início com o desenvolvimento e avaliação de uma rede neural artificial composta por um único neurônio, utilizada como linha de base para a comparação dos resultados obtidos pelo modelo subsequente. Essa rede foi desenvolvida com o auxílio da biblioteca *Keras*, que facilitou a construção, treinamento e avaliação do modelo.

Para isso, o *dataset* foi segmentado por temporadas, e os dados de cada temporada foram divididos em conjuntos de treino e teste. As *features* selecionadas foram alocadas em um vetor X, enquanto a variável alvo das predições foi alocada em um vetor y. Dessa forma, para cada temporada, o conjunto de dados foi separado em `X_train`, `X_test`, `y_train` e `y_test`, garantindo uma estrutura organizada para o treinamento e validação dos modelos.

Com o conjunto de dados selecionado e as *features* devidamente definidas, foram calculadas as métricas de avaliação do modelo. A média da acurácia e do erro quadrático médio (MSE) foram, respectivamente, 56,65% e 0,2458. Esses resultados indicam um desempenho insatisfatório, pois um modelo trivial que atribuísse probabilidades fixas de

Quadro 3 – Lista de *features* utilizadas no modelo

Feature
Diferença média de pontos até a partida
Diferença média de arremessos de quadra convertidos até a partida
Diferença média de arremessos de quadra tentados até a partida
Diferença média do percentual de acerto de arremessos de quadra até a partida
Diferença média de cestas de 2 pontos convertidas até a partida
Diferença média de cestas de 2 pontos tentadas até a partida
Diferença média do percentual de acerto de cestas de 2 pontos até a partida
Diferença média de cestas de 3 pontos convertidas até a partida
Diferença média de cestas de 3 pontos tentadas até a partida
Diferença média do percentual de acerto de cestas de 3 pontos até a partida
Diferença média de lances livres convertidos até a partida
Diferença média de lances livres tentados até a partida
Diferença média do percentual de acerto de lances livres até a partida
Diferença média do percentual de arremessos verdadeiros até a partida
Diferença média de rebotes ofensivos até a partida
Diferença média de rebotes defensivos até a partida
Diferença média de rebotes totais até a partida
Diferença média de assistências até a partida
Diferença média de roubos de bola até a partida
Diferença média de bloqueios até a partida
Diferença média de desperdícios de posse até a partida
Diferença média de faltas pessoais até a partida
Elo do time na partida

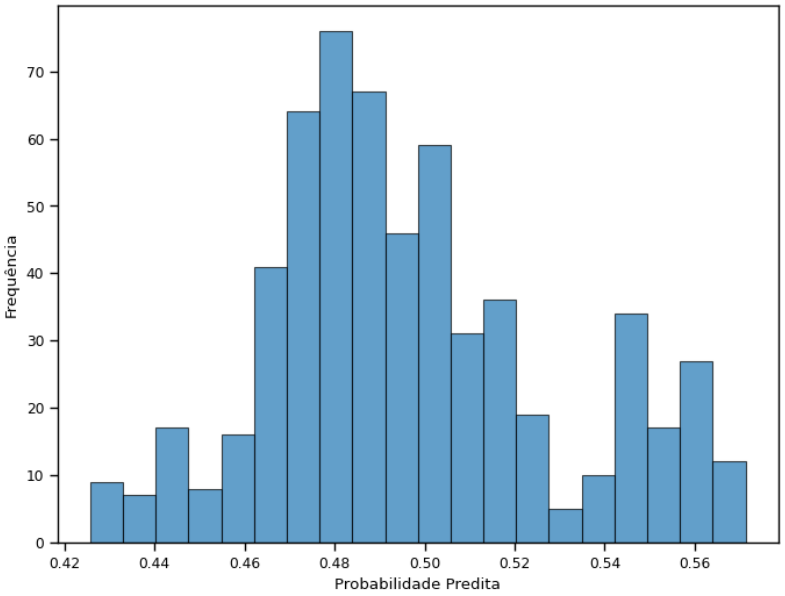
Fonte: Autor.

0.5 para todos os eventos apresentaria um MSE esperado de 0.25. Assim, ao analisar esses valores juntamente com o histograma das probabilidades preditas para a temporada de 2022, representado na Fig. 15, observa-se que a maioria das previsões permaneceu próxima de 50%. Isso sugere que o modelo teve dificuldades em capturar padrões relevantes nos dados, limitando sua capacidade preditiva.

Diante desses resultados, foi aplicada a técnica de recalibração pós-processamento *Platt Scaling*, que ajusta as probabilidades preditas por meio de uma regressão logística, mapeando-as em probabilidades calibradas. No entanto, mesmo após esse ajuste, o MSE das probabilidades preditas permaneceu em 0,2441, um valor ainda muito próximo de 0,25, que seria esperado para um modelo trivial sem capacidade preditiva significativa.

Para avaliar o desempenho do modelo no contexto das apostas esportivas, as probabilidades preditas e calibradas foram comparadas com as *odds* de fechamento coletadas na etapa de obtenção dos dados. A análise revelou que a diferença média entre as probabilidades preditas e as *odds* de fechamento foi de 13%. Esse valor é considerado elevado, especialmente no contexto de busca por estimativas mais precisas, uma vez que as *odds*

Figura 15 – Histograma de probabilidades preditas para temporada de 2022.



Fonte: Autor.

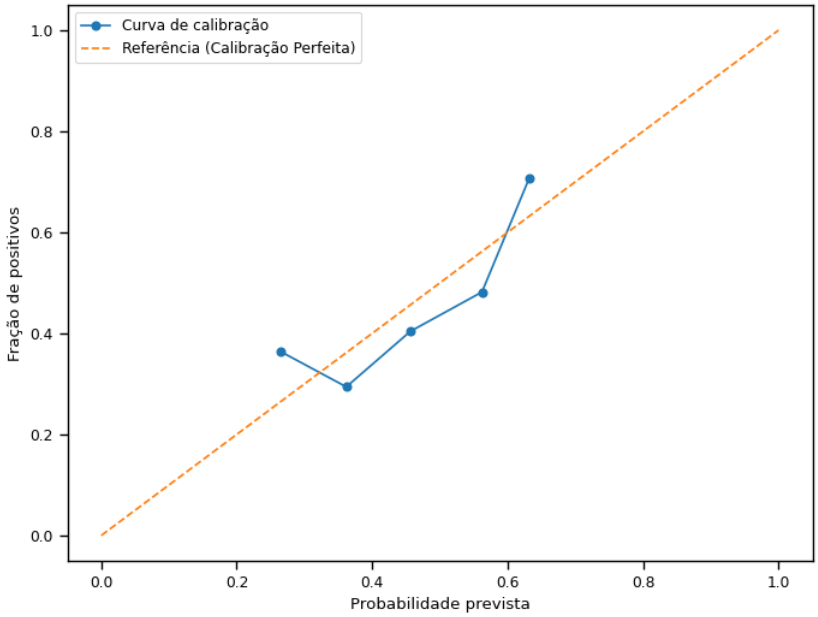
de fechamento são amplamente reconhecidas como boas aproximações da probabilidade real dos eventos.

Após a implementação da rede neural utilizada como linha de base para os resultados, foi desenvolvido o modelo preditivo LSTM. A separação do conjunto de dados seguiu a mesma estrutura da primeira implementação, garantindo que a ordem cronológica dos dados fosse preservada. Assim, foi definida uma janela temporal de cinco partidas, representando o comprimento da sequência utilizada pelo modelo para a criação de seus blocos de memória.

Com essa abordagem, o modelo foi executado para obtenção das probabilidades preditas. A avaliação das métricas de desempenho revelou que a média dos resultados por temporada para acurácia e erro quadrático médio (MSE) foi de 55,6% e 0,2435, respectivamente. Esses valores indicam um desempenho superior ao modelo de rede neural simples, especialmente no que diz respeito ao MSE, mesmo antes da aplicação de técnicas de calibração. O gráfico da curva de calibração para a temporada de 2022 pode ser encontrado na Figura 16

Dessa forma, a mesma técnica de calibração pós-processamento foi aplicada às probabilidades preditas pelo modelo LSTM. Com essa recalibração, o erro quadrático médio (MSE) foi reduzido para 0,2427. Além disso, ao comparar essas probabilidades calibradas com as *odds* de fechamento presentes no *dataset*, observou-se uma diferença média de 10%, indicando uma maior aproximação das probabilidades preditas em relação às estimativas finais do mercado. Porém, ainda existindo uma grande diferença entre estas probabilidades.

Figura 16 – Curva de calibração modelo LSTM temporada 2022.



Fonte: Autor.

A partir dos resultados obtidos, observa-se que os modelos preditivos implementados não demonstraram um desempenho conclusivo na geração de probabilidades mais precisas do que as *odds* de fechamento. A significativa diferença média encontrada entre as probabilidades preditas e as *odds* de fechamento sugere uma alta inconsistência nos valores estimados pelos modelos, mesmo após a aplicação das técnicas de calibração.

Portanto, sem uma calibração adequada e uma redução significativa dessa discrepância, a combinação das probabilidades preditas com as *odds* de fechamento não resultaria em previsões mais precisas. Ademais, devido à subestimação do tempo necessário para desenvolver o algoritmo de apostas e realizar a engenharia de *features*, os resultados do treinamento e dos testes dos modelos, bem como a integração das probabilidades preditas com as *odds* das casas de apostas precisas, foram considerados inconclusivos.

4.5 Aplicação da estratégia de apostas e avaliação de resultados

Com base nos resultados obtidos nas etapas anteriores, foi possível implementar a estratégia de apostas. Diante da ausência de conclusões definitivas na etapa de implementação dos modelos preditivos, optou-se por utilizar exclusivamente o *dataset* de controle para a aplicação da estratégia. Nesse contexto, a probabilidade $\tilde{\pi}_i$, calculada a partir das *odds* da casa de apostas Pinnacle, foi comparada com as *odds* das demais casas de apostas selecionadas para o estudo.

A partir de 22 de outubro de 2024, data de início da temporada 2024-2025 da NBA, o serviço *Server* iniciou a coleta diária das *odds* das partidas em todas as casas

de apostas selecionadas. As *odds* coletadas eram então enviadas ao serviço *Brain*, que processava as informações e retornava as oportunidades de apostas com valor esperado positivo. As respostas recebidas eram armazenadas na base de dados para posterior análise. Esse processo foi mantido continuamente até 9 de fevereiro de 2025, quando a coleta foi interrompida para a análise dos resultados.

Durante esse período, foram identificadas 150.448 apostas com valor esperado de lucro positivo. Dentre elas, 97.444 eram apostas únicas, ou seja, oportunidades distintas de aposta, considerando que uma mesma aposta de valor pode estar disponível em múltiplas casas de apostas. Sendo assim, a Tab. 4 apresenta esses dados em relação às tabelas SQL utilizadas para o armazenamento das informações.

Tabela 4 – Metadados tabelas `bets` e `betProps`

Nome da Tabela	**Número de Linhas**
`bets`	97.444
`betProps`	150.448

Fonte: Autor.

Essas apostas abrangeram um total de 30 mercados diferentes, cada um representando um tipo específico de evento dentro das partidas analisadas. A Tab. 12 detalha todos os mercados considerados no estudo e o número de apostas encontradas em cada um deles e pode ser encontrada no Apêndice B.

Para a simulação das apostas e avaliação dos resultados, foi desenvolvido um *script Python*, que converteu as tabelas `bets` e `betProps` em dois *dataframes* da biblioteca Pandas. Dessa forma, foi possível iterar desde o primeiro dia em que as apostas foram registradas até o último, como representado na Fig. 17. A cada dia, todas as apostas registradas na tabela `bets` eram simuladas, porém, para minimizar o risco de múltiplas entradas na mesma aposta em diferentes casas de apostas, apenas um registro da tabela `betProps` era selecionado para a simulação.

O critério adotado para a escolha da `betProp` foi a ordem cronológica de detecção da aposta, priorizando aquelas identificadas primeiro, de forma a simular o comportamento natural de um apostador. Caso mais de uma aposta fosse registrada simultaneamente, a seleção priorizou aquela com o maior valor esperado de lucro calculado.

Além disso, para a simulação, considerou-se um capital inicial de R$100,00 em cada uma das onze casas de apostas, totalizando um capital inicial de R$1.100,00. Essa abordagem permitiu uma avaliação mais realista da evolução do capital ao longo do período analisado.

A cada dia da iteração, além de registrar as novas entradas, o *script* também processava os resultados das apostas realizadas no dia anterior. Em caso de aposta ganha, o

Figura 17 – *Script* simulação de entradas.

```
start_date = date(2024, 10, 22)
end_date = max_date

current_date = start_date
delta = timedelta(days=1)

while current_date <= end_date:
  process_yesterday_result(current_date)

  buildResultsDataframes(current_date, results_by_date_df,
  ↪  results_boomaker_by_date_df)

  enter_in_today_bets(current_date)

  current_date += delta
```

Fonte: Autor

valor apostado era devolvido juntamente com o lucro correspondente; em caso de empate, apenas o valor apostado era restituído; e, em caso de derrota, não havia devolução de dinheiro. Esse fluxo permitiu acompanhar a evolução do capital ao longo do período analisado, simulando a gestão financeira de um apostador no mercado de apostas esportivas.

Durante o período analisado, foram realizadas um total de 97.444 apostas, das quais 58.133 resultaram em vitórias, 37.140 em derrotas e 2.171 terminaram empatadas, conforme apresentado na Tab. 5. Além disso, a *odd* média das apostas foi de 1.78, enquanto a moda da amostra foi 1.2, indicando uma maior concentração de apostas com este valor de *odd*.

Tabela 5 – Desempenho geral das apostas

Apostas Ganhas	Apostas Perdidas	Apostas Empatadas	Porcentagem de Vitória (%)
58.133	37.140	2.171	59,65%

Fonte: Autor.

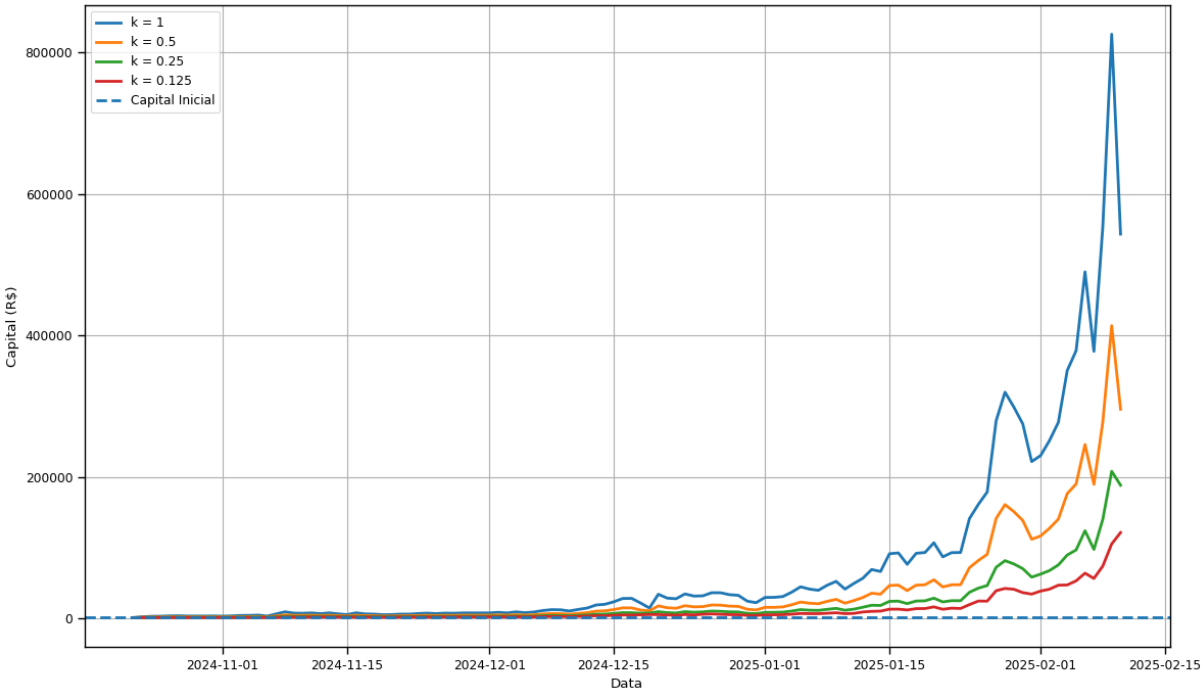
4.5.1 Resultados

A partir da simulação das apostas por meio do *script*, foi possível coletar e analisar os resultados obtidos. Para essa avaliação, foram consideradas quatro abordagens distintas para determinar o valor de entrada do apostador em cada aposta. Como mencionado anteriormente, o montante a ser apostado foi calculado com base no Critério de Kelly,

conforme definido pela Eq. 3.4, esse valor determinado para cada aposta foi armazenado na coluna `kellyPercentage` da tabela `betProps`, garantindo a rastreabilidade e permitindo análises posteriores.

Para avaliar diferentes níveis de exposição ao risco, foram testados quatro valores distintos para a constante k, responsável por fracionar o valor sugerido pelo Critério de Kelly. Os valores analisados foram: 1, 0.5, 0.25 e 0.125, permitindo a comparação do impacto dessa variação sobre os retornos financeiros da estratégia adotada. A Fig. 18 apresenta a evolução do capital do apostador ao longo do tempo para cada uma das estratégias, possibilitando uma análise visual do desempenho financeiro resultante de cada abordagem.

Figura 18 – Evolução do capital do apostador ao longo do tempo.



Fonte: Autor.

A partir dos resultados obtidos, observa-se que a estratégia utilizando $k = 1$ proporcionou os maiores retornos financeiros, encerrando o período com um ROI de 49.365,4%. No entanto, essa mesma abordagem também apresentou a maior volatilidade atingindo 267,16%.

Esse comportamento pode ser explicado pelo fator exponencial do Critério de Kelly, que, por sempre alocar uma fração f do capital em cada aposta, amplifica os ganhos, mas também aumenta significativamente os riscos. Como consequência, essa estratégia registrou um *drawdown* máximo de 48,2%, evidenciando a alta exposição a oscilações no capital investido.

As estratégias utilizando $k = 0.5$ e $k = 0.25$ apresentaram retornos intermediá-

rios em relação às demais abordagens, com ROI de 26.861,0% e 17.102,2%, respectivamente. Ambas registraram grandes oscilações, refletidas em suas volatilidades de 191,63% e 148,44%.

Por outro lado, a estratégia mais conservadora, que utilizou um oitavo do valor sugerido pelo Critério de Kelly ($k = 0.125$), resultou em um retorno sobre investimento de 11.025,9%, encerrando o período com um capital final de R$ 121.284,97. Sua volatilidade ao longo do período foi de 126,4%, e o maior *drawdown* registrado foi de 34,2%. Apesar de ainda apresentar um risco considerável, essa abordagem demonstrou uma exposição muito menor quando comparada à estratégia com $k = 1$. O resumo dos resultados obtidos para cada valor de k está representado na Tabela 6.

Tabela 6 – Resultados financeiros para diferentes valores de k

k	**Capital Inicial (R$)**	**Capital Final (R$)**	**ROI**	*vol*
1	1.100,00	543.020,06	49.365,4%	267,16%
0.5	1.100,00	295.471,15	26.861,0%	191,63%
0.25	1.100,00	188.124,61	17.102,2%	148,44%
0.125	1.100,00	121.284,97	11.025,9%	126,4%

Fonte: Autor.

Após a análise dos resultados das diferentes estratégias de alocação de capital por aposta, foi realizada uma investigação mais aprofundada sobre o desempenho das casas de apostas utilizadas no estudo. Para essa análise, utilizou-se como referência os retornos obtidos pela abordagem mais conservadora, com $k = 0.125$, permitindo uma avaliação menos suscetível a grandes variações de capital.

Dessa forma, a Tabela 7 apresenta um panorama das casas de apostas incluídas no estudo, destacando o número total de apostas realizadas em cada uma delas ao longo do período analisado. Esse levantamento possibilita compreender quais plataformas ofereceram mais oportunidades de apostas de valor e avaliar o impacto dessas casas na estratégia geral.

Os resultados finais de capital para cada casa de apostas foram representados graficamente na Fig. 19. Através dessa representação, foi possível observar uma grande discrepância entre os retornos obtidos nas diferentes casas de apostas, com a 1XBet respondendo por 96,9% do capital final total, mesmo sem ser a casa com o maior número de apostas realizadas.

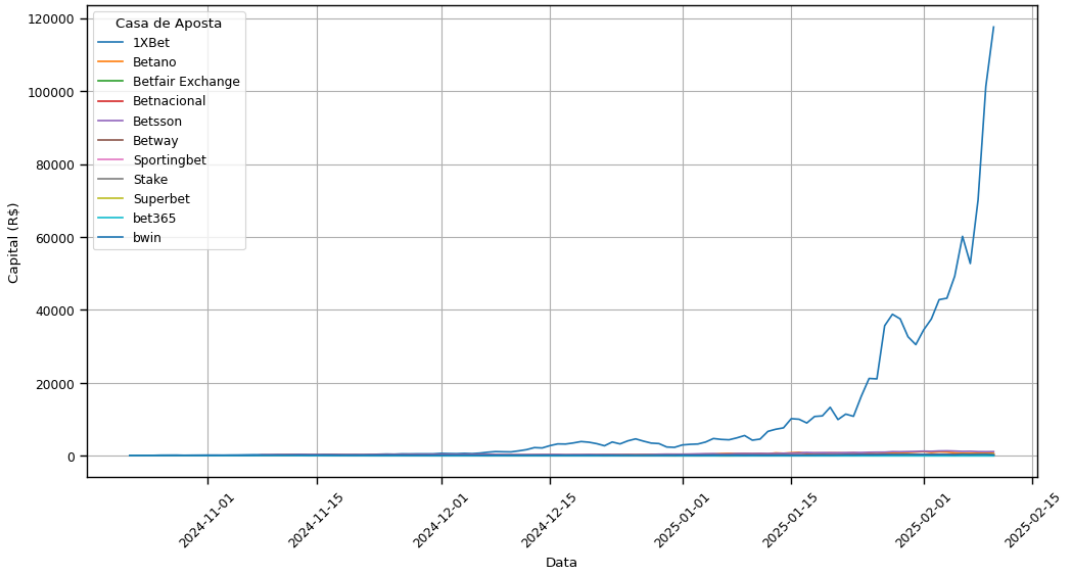
Essa disparidade sugere que as *odds* oferecidas por essa casa frequentemente apresentam desvios significativos em relação às demais, evidenciando ineficiências de mercado que podem ser exploradas por estratégias de apostas. A Tab. 8 apresenta um resumo dos resultados financeiros, segmentados por casa de apostas.

Tabela 7 – Número de apostas realizadas por casa de apostas

Casa de Apostas	**Número de Apostas**
Superbet	26.931
1XBet	26.632
Bet365	20.297
Betano	11.807
Bwin	6.179
Betsson	2.466
Betnacional	1.191
Stake	915
Betfair Exchange	531
Sportingbet	321
Betway	174

Fonte: Autor.

Figura 19 – Evolução do capital do apostador por casa de aposta ao longo do tempo.



Fonte: Autor.

A partir dessa análise, ficou evidente a influência das *odds* da 1XBet nos altos valores de ROI obtidos na avaliação das estratégias de apostas. Dessa forma, para investigar se a lucratividade das estratégias estava unicamente atrelada às *odds* de uma casa de apostas, foi realizada uma nova simulação, dessa vez desconsiderando as apostas de valor salvas da 1XBet. Além disso, para manter a equivalência com a simulação original, foi considerado um capital inicial de R$110,00 por casa de aposta, totalizando R$1.100,00. Essa abordagem permitiu avaliar a robustez da estratégia e verificar se os retornos financeiros

Tabela 8 – Capital inicial e final por casa de apostas

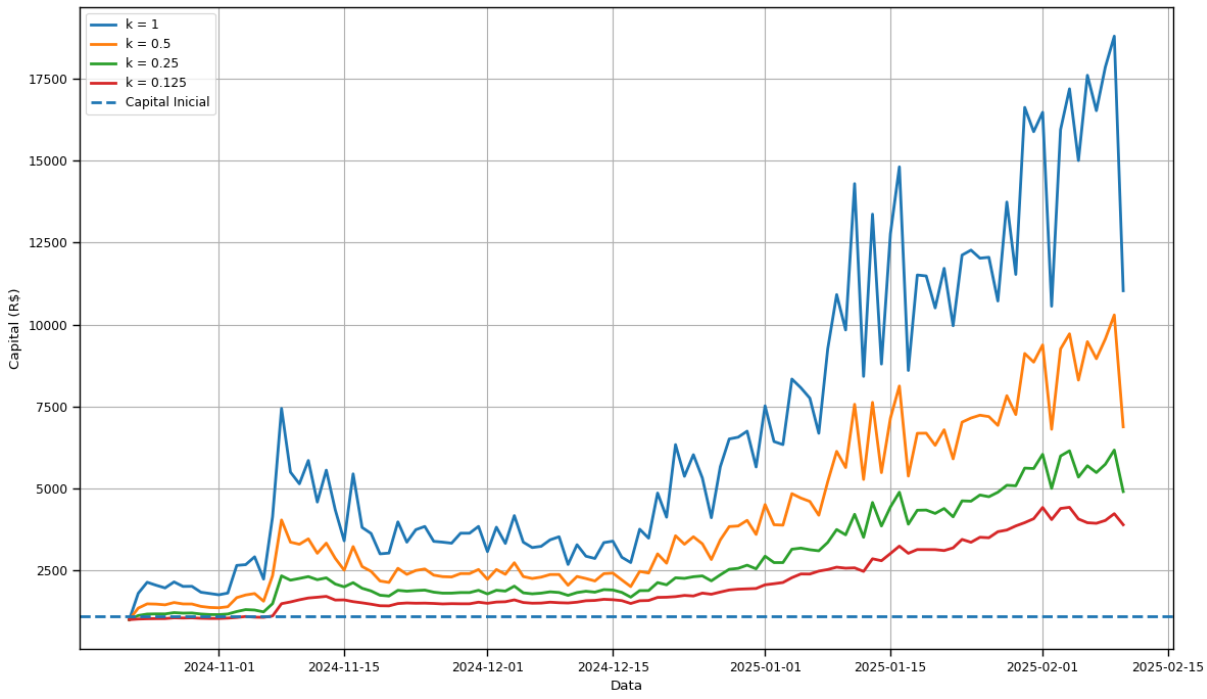
Casa de Apostas	**Capital Inicial (R$)**	**Capital Final (R$)**
1XBet	100,00	117.554,36
Betsson	100,00	1.168,06
Betano	100,00	667,92
Superbet	100,00	419,33
Betnacional	100,00	414,56
bwin	100,00	299,47
Stake	100,00	237,09
Betway	100,00	199,87
Sportingbet	100,00	195,67
Betfair Exchange	100,00	121,83
bet365	100,00	6,81

Fonte: Autor.

permaneciam positivos mesmo sem a contribuição dessa casa de apostas.

Com essa nova simulação, um total de 76.892 apostas foram realizadas, das quais 47.281 resultaram em vitória, correspondendo a uma taxa de acerto de 61,4%. A Fig. 20 ilustra graficamente a evolução do capital do apostador ao longo do tempo para cada um dos diferentes valores de k testados nesta simulação.

Figura 20 – Evolução do capital do apostador ao longo do tempo sem 1XBet.



Fonte: Autor.

Imediatamente, observa-se uma redução expressiva nos ganhos em todas as quatro estratégias de apostas. Na simulação com $k = 1$, o capital final atingiu R$ 12.131,71, re-

presentando uma redução de 97,7% em relação à simulação que considerava as apostas na casa 1XBet. Da mesma forma, as simulações com $k = 0.5$, $k = 0.25$ e $k = 0.125$ apresentaram reduções no capital final de 97,4%, 97,1% e 96,4%, respectivamente, evidenciando o impacto significativo da exclusão dessa casa de apostas nos resultados obtidos.

Ainda assim, todas as estratégias mantiveram um ROI positivo, sendo 1.102,8% para $k = 1$, 687,9% para $k = 0.5$, 490,5% para $k = 0.25$ e 389,3% para $k = 0.125$, conforme demonstrado na Tab. 9. Esses resultados evidenciam que, embora uma parcela significativa dos lucros da simulação inicial tenha se originado das *odds* desreguladas da 1XBet, a estratégia de apostas não depende exclusivamente das ineficiências dessa casa. A abordagem ainda identificou diversas oportunidades de valor esperado positivo em outras casas de apostas.

Tabela 9 – Resultados financeiros para diferentes valores de k sem 1XBet

k	**Capital Inicial (R$)**	**Capital Final (R$)**	**ROI**	*vol*
1	1.100,00	12.131,71	1.102,8%	269,2%
0.5	1.100,00	7.567,38	687,9%	179,7%
0.25	1.100,00	5.396,21	490,5%	101,8%
0.125	1.100,00	4.282,78	389,3%	50,4%

Fonte: Autor.

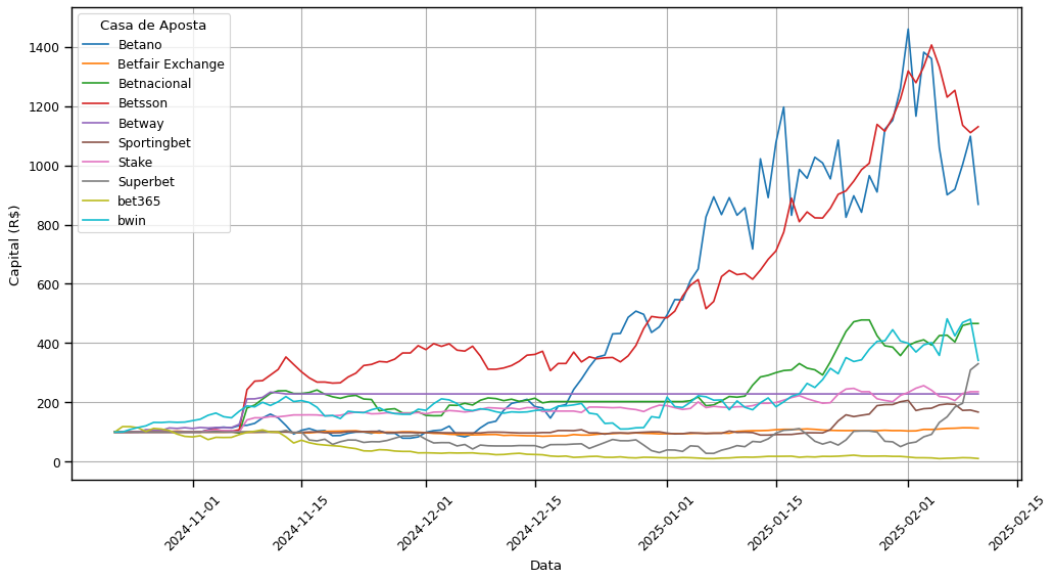
Ao analisar o capital final do apostador em cada casa de aposta, considerando a estratégia com $k = 0.125$, foi possível observar uma distribuição mais equilibrada do capital total, conforme ilustrado na Tab. 10. Nesta simulação, a casa de apostas Betsson apresentou o maior capital final, representando 29% do total acumulado ao longo do período analisado. A Fig. 21 ilustra graficamente a evolução do capital ao longo do tempo para cada uma das casas de apostas.

Tabela 10 – Capital inicial e final por casa de apostas sem 1XBet

Casa de Apostas	**Capital Inicial (R$)**	**Capital Final (R$)**
Betsson	110,00	1.243,29
Betano	110,00	955,31
Betnacional	110,00	512,87
bwin	110,00	376,05
Superbet	110,00	363,56
Stake	110,00	259,78
Betway	110,00	251,52
Sportingbet	110,00	184,19
Betfair Exchange	110,00	124,32
bet365	110,00	11,86

Fonte: Autor.

Figura 21 – Evolução do capital por casa de aposta ao longo do tempo sem 1XBet.



Fonte: Autor.

5 Conclusão

Neste trabalho, foi realizado um estudo sobre as ineficiências do mercado de apostas esportivas da NBA e seu potencial de exploração financeira por meio de estratégias de arbitragem estatística. Além disso, investigou-se a viabilidade da aplicação de modelos preditivos como uma ferramenta para aprimorar essas estratégias, permitindo a identificação e exploração mais eficiente de oportunidades com valor esperado positivo. Com base em artigos e estudos anteriores, foi possível desenvolver uma metodologia estruturada para a implementação da estratégia de apostas e implementação de modelos de *machine learning* para a predição de vencedores em partidas da liga de basquete norte-americana.

Para conduzir este estudo, foram definidos três objetivos principais. O primeiro consistiu na análise dos diferentes indicadores estatísticos disponíveis para a NBA, com o intuito de selecionar aquele que serviria como alvo dos modelos preditivos. Para isso, foi realizada uma revisão bibliográfica, identificando os indicadores mais utilizados em estudos semelhantes, a fim de estabelecer uma base metodológica sólida. O segundo objetivo foi a implementação dos modelos preditivos e a avaliação da viabilidade de sua aplicação na previsão de probabilidades associadas aos mercados de apostas. Por fim, o terceiro objetivo visou o desenvolvimento e a análise da viabilidade econômica de uma estratégia de apostas esportivas, explorando possíveis ineficiências de mercado.

Com os objetivos definidos, foi realizada uma pesquisa para estabelecer a metodologia a ser empregada no estudo, determinando as principais referências que fundamentariam o trabalho. A partir dessa definição, iniciou-se a etapa de coleta de dados, na qual foi aplicada a técnica de *web scraping* para construir uma base de dados estatísticos históricos de jogadores da NBA. Essa base de dados foi essencial para o treinamento e teste dos modelos preditivos desenvolvidos, garantindo uma abordagem fundamentada na análise de dados reais do cenário esportivo.

A partir disso, foi definido o indicador estatístico que serviria como alvo das predições. Para essa escolha, foram novamente consultadas as referências identificadas na pesquisa inicial, conduzindo à seleção do indicador de vencedor da partida. Dessa forma, o alvo das predições foi determinado como a probabilidade de vitória de um time em uma determinada partida, permitindo a modelagem e análise do desempenho das equipes com base em estatísticas históricas. Assim, foram implementados dois modelos distintos de *machine learning*, com foco na métrica de calibração para avaliar seu desempenho. No entanto, os resultados obtidos foram inconclusivos, indicando uma grande inconsistência nas probabilidades preditas. Diante disso, decidiu-se não utilizar estes resultados nas etapas subsequentes do estudo.

Para a execução da estratégia de apostas, foi desenvolvido um algoritmo capaz de coletar e comparar *odds* de diferentes casas de apostas, buscando identificar ineficiências no mercado. Para isso, as casas de apostas foram classificadas em dois grupos: as casas precisas, cujas *odds* serviriam como referência para estimar a probabilidade real dos eventos esportivos, e as casas não precisas, onde seriam buscadas as ineficiências e onde as apostas seriam simuladas.

O algoritmo foi hospedado na nuvem e operou continuamente entre outubro de 2024 e fevereiro de 2025, armazenando todas as apostas com valor esperado de lucro positivo em uma base de dados. A partir dessas informações, foi realizada uma simulação de entradas nas apostas, considerando um capital inicial previamente definido e testando quatro níveis diferentes de exposição ao risco, variando a alocação de capital em cada aposta.

Os resultados obtidos demonstram que mais de 150.000 oportunidades de exploração financeira foram identificadas, sendo essas oportunidades mais frequentes em determinadas casas de apostas, com a 1XBet se destacando nesse aspecto. Além disso, os testes indicaram que todas as estratégias avaliadas apresentaram lucro, com um ROI mínimo de 389,3%. No entanto, os altos índices de volatilidade registrados, chegando a 267,16%, evidenciam o nível de risco associado à estratégia.

Um ponto crucial a ser destacado é a inviabilidade da execução desta estratégia de apostas em um ambiente real fora das simulações. Embora os resultados obtidos possam sugerir um investimento financeiramente atrativo, na prática, esses retornos são inatingíveis. Isso ocorre porque as casas de apostas, cientes das ineficiências exploradas, impõem restrições rigorosas para evitar que apostadores tirem proveito dessas oportunidades. Em particular, essas casas proíbem apostas realizadas por algoritmos ou qualquer forma de automação e, além disso, limitam ou até bloqueiam as operações de usuários que demonstram lucratividade consistente a longo prazo ao explorar tais ineficiências.

Outro ponto relevante a ser destacado é o crescente problema do vício em apostas esportivas, que tem se tornado cada vez mais frequente na sociedade brasileira (ALVES, 2024). Por se tratar de um mercado relativamente novo e, muitas vezes, impulsionado por promessas irreais de ganhos com o objetivo de atrair novos usuários, é fundamental que as apostas esportivas sejam encaradas apenas como uma forma de entretenimento destinada a maiores de idade. Para mitigar os riscos associados ao vício em jogos, torna-se essencial a conscientização do público brasileiro. Essa conscientização pode ser promovida por meio de um entendimento sobre o funcionamento das apostas e a matemática por trás delas, permitindo que os apostadores compreendam que, na maioria das vezes, suas disputas contra as casas de apostas são desvantajosas, caracterizadas por um valor esperado de lucro negativo.

Para pesquisas futuras, pode-se explorar diferentes abordagens para o uso de mo-

delos preditivos na identificação de ineficiências no mercado de apostas esportivas. Além disso, um aprimoramento relevante seria a aplicação de métodos estatísticos e estratégias de alocação de capital que considerem as correlações de dependência entre probabilidades de eventos esportivos simultâneos. Dessa forma, poderia ser investigado como o desempenho de determinados jogadores e times influencia o rendimento de seus companheiros e adversários, proporcionando uma análise mais sofisticada e precisa das oportunidades de apostas.

Referências

AKANSU, A. N.; TORUN, M. U. *A Primer for Financial Engineering*. 1. ed. [S.l.]: Elsevier, 2015. Citado 2 vezes nas páginas 37 e 38.

ALVES, I. *Os divórcios motivados pelo vício em bets e jogo do tigrinho*. 2024. Acesso em: 23 set. 2024. Disponível em: <https://www.bbc.com/portuguese/articles/cy4l4p8dy3lo>. Citado na página 72.

BAUMER, B.; ZIMBALIST, A. *The Sabermetric Revolution: Assessing the Growth of Analytics in Baseball*. [S.l.]: University of Pennsylvania Press, 2014. Citado na página 17.

BUCHDAHL, J. *Restringir ou não restringir? Eis a questão*. 2020. Acesso em: 30 jan. 2025. Disponível em: <https://www.pinnacle.com/betting-resources/pt/educational/to-restrict-or-not-to-restrict-that-is-the-question/7h72dyzcczxjzdzs>. Citado na página 54.

BUCHDAHL, J. *Uma história em duas casas de apostas: modelos especializados x modelos recreativos*. 2021. Acesso em: 30 jan. 2025. Disponível em: <https://www.pinnacle.com/betting-resources/pt/educational/a-tale-of-two-bookmakers-sharp-vs-recreational-models/mc22s4yrepqa8uhh>. Citado na página 54.

BURKOV, A. *The Hundred-Page Machine Learning Book*. [S.l.]: Andriy Burkov, 2019. ISBN 9781999579517. Citado 6 vezes nas páginas 21, 22, 23, 24, 25 e 26.

CAO, C. *Sports Data Mining Technology Used in Basketball Outcome Prediction*. Dissertação (Mestrado) — Technological University Dublin, 2012. Citado 2 vezes nas páginas 17 e 49.

CLARKE, J.; JANDIK, T.; MANDELKER, G. The efficient markets hypothesis. *Expert financial planning: Advice from industry leaders*, v. 7, p. 126–141, 2001. Citado na página 37.

CORTIS, D. Expected values and variances in bookmaker payouts: A theoretical approach towards setting limits on odds. *Journal of Prediction Markets*, v. 9, p. 1, 07 2015. Citado 2 vezes nas páginas 32 e 34.

COWAN, D. *Poisson Distribution*. 2016. Acesso em: 2 fev. 2025. Disponível em: <https://www.ml-science.com/poisson-distribution>. Citado na página 53.

DOTAN, G.; LEW, V.; LIU, Z. *Beating the Book: A Machine Learning Approach to Identifying an Edge in NBA Betting Markets*. Dissertação (Mestrado) — University of California, Los Angeles, 2020. AAI28001181. Citado 2 vezes nas páginas 46 e 49.

ELMASRI, R.; NAVATHE, S. *Sistemas de Banco de Dados*. 4. ed. [S.l.]: Pearson, 2004. Citado na página 31.

GANDAR, J. et al. Informed traders and price variations in the betting market for professional basketball games. *The Journal of Finance*, American Finance Association, Wiley, v. 53, n. 1, p. 385–401, 1998. Citado na página 33.

GONZáLEZ, J. M. et al. The poisson model limits in nba basketball: Complexity in team sports. *Physica A: Statistical Mechanics and its Applications*, v. 464, p. 182–190, 2016. Citado na página 53.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. ed. [S.l.]: Springer, 2009. Citado 2 vezes nas páginas 21 e 22.

HAYKIN, S. S. *Neural networks and learning machines*. 3. ed. [S.l.]: Pearson Education, 2009. Citado 3 vezes nas páginas 23, 24 e 26.

HOUDT, G. V.; MOSQUERA, C.; NáPOLES, G. A review on the long short-term memory model. *Artificial Intelligence Review*, v. 53, 12 2020. Citado 2 vezes nas páginas 26 e 27.

HSIEH, C.-H.; BARMISH, B. R. On kelly betting: Some limitations. In: *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. [S.l.]: IEEE, 2015. p. 165–172. Citado na página 37.

HUBACEK, O.; SOUREK, G.; ZELEZNY, F. Exploiting sports-betting market using machine learning. *International Journal of Forecasting*, v. 35, n. 2, p. 783–796, 2019. Citado 7 vezes nas páginas 18, 35, 40, 44, 49, 55 e 56.

KARLIS, D.; NTZOUFRAS, I. Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, v. 52, n. 3, p. 381–393, 2003. Citado na página 53.

KELLY, J. L. A new interpretation of information rate. *Bell System Technical Journal*, v. 35, n. 4, p. 917–926, 1956. Citado 2 vezes nas páginas 35 e 36.

KHDER, M. Web scraping or web crawling: State of art, techniques, approaches and application. *International Journal of Advances in Soft Computing and its Applications*, 2021. Citado na página 31.

KUBATKO, J. et al. A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sports*, v. 3, 2007. Citado 5 vezes nas páginas 17, 18, 41, 47 e 56.

KUMAR, A.; LIANG, P. S.; MA, T. Verified uncertainty calibration. In: *Advances in Neural Information Processing Systems*. [S.l.]: Curran Associates, Inc., 2019. v. 32. Citado 2 vezes nas páginas 29 e 30.

LANGVILLE, A. N.; MEYER, C. D. *Who's 1?: The Science of Rating and Ranking*. [S.l.]: Princeton University Press, 2012. ISBN 9780691154220. Citado na página 58.

NAKAMURA, J. *Setor de apostas online cresceu 734% desde 2021, aponta pesquisa*. 2024. Acesso em: 7 ago. 2024. Disponível em: <https://www.cnnbrasil.com.br/economia/negocios/setor-de-apostas-online-cresceu-734-desde-2021-aponta-pesquisa/>. Citado na página 17.

NICULESCU-MIZIL, A.; CARUANA, R. Predicting good probabilities with supervised learning. *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*, p. 625–632, 01 2005. Citado na página 44.

POVOA, L. et al. O mercado de apostas esportivas on-line: impactos, desafios para a definição de regras de funcionamento e limites. *Núcleo de Estudos e Pesquisas/CONLEG/Senado*, 2023. Citado na página 17.

RAINIO, O.; TEUHO, J.; KLéN, R. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, v. 14, 2024. Citado 2 vezes nas páginas 28 e 29.

ROLLA, L. T.; LIMA, B. N. B. de. *Probabilidade*. [S.l.]: s.n., 2024. Citado 2 vezes nas páginas 34 e 53.

SARKER, I. Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, Springer, v. 2, n. 6, p. 420, 2021. Citado 4 vezes nas páginas 24, 25, 44 e 55.

SHLEIFER, A. *Inefficient Markets: An Introduction to Behavioral Finance*. Oxford University Press, 2000. ISBN 9780198292272. Disponível em: <https://doi.org/10.1093/0198292279.001.0001>. Citado na página 38.

SHVETS, A. *Chain of Responsibility*. 2021. Acesso em: 30 jan. 2025. Disponível em: <https://refactoring.guru/pt-br/design-patterns/chain-of-responsibility>. Citado na página 51.

STEKLER, H.; SENDOR, D.; VERLANDER, R. Issues in sports forecasting. *International Journal of Forecasting*, v. 26, n. 3, p. 606–621, 2010. Citado 3 vezes nas páginas 17, 33 e 40.

THORP, E. O. Chaper 9 - the kelly criterion in blackjack sports betting, and the stock market*. In: ZENIOS, S.; ZIEMBA, W. (Ed.). *Handbook of Asset and Liability Management*. San Diego: North-Holland, 2008. p. 385–428. ISBN 978-0-444-53248-0. Citado 2 vezes nas páginas 35 e 36.

TORRES, R. Prediction of nba games based on machine learning methods. In: *Universidade de Wisconsin*. [s.n.], 2013. Disponível em: <https://api.semanticscholar.org/CorpusID:46432106>. Citado na página 49.

VRAčAR, P.; ŠTRUMBELJ, E.; KONONENKO, I. Modeling basketball play-by-play data. *Expert Systems with Applications*, v. 44, p. 58–66, 2016. ISSN 0957-4174. Citado na página 55.

WALSH, C.; JOSHI, A. Machine learning for sports betting: Should model selection be based on accuracy or calibration? *Machine Learning with Applications*, v. 16, p. 100539, 2024. Citado 9 vezes nas páginas 19, 30, 35, 40, 44, 47, 49, 57 e 81.

WHEATCROFT, E. Profiting from overreaction in soccer betting odds. *Journal of Quantitative Analysis in Sports*, v. 16, n. 3, p. 193–209, 2020. Citado na página 33.

ŠTRUMBELJ, E.; VRAčAR, P. Simulating a basketball match with a homogeneous markov model and forecasting the outcome. *International Journal of Forecasting*, v. 28, n. 2, p. 532–542, 2012. ISSN 0169-2070. Citado na página 55.

Apêndices

APÊNDICE A – Tabela estatísticas NBA

Tabela 11 – Siglas das estatísticas e suas descrições

Sigla	**Descrição**
FG	Arremessos de campo: número total de cestas de 2 e 3 pontos convertidas.
FGA	Tentativas de arremessos de campo: número de arremessos tentados em jogo.
FG%	Percentual de acerto nos arremessos de campo: FG/FGA.
3P	Cestas de 3 pontos convertidas.
3PA	Tentativas de arremesso de 3 pontos.
3P%	Percentual de acerto em arremessos de 3 pontos: 3P/3PA.
FT	Lances livres convertidos.
FTA	Tentativas de lance livre.
FT%	Percentual de acerto nos lances livres: FT/FTA.
TRB	Rebotes totais: número de rebotes defensivos e ofensivos coletados.
ORB	Rebotes ofensivos coletados enquanto o time estava no ataque.
DRB	Rebotes defensivos coletados enquanto o time estava na defesa.
AST	Assistências: passes que resultam diretamente em uma cesta.
STL	Roubos de bola: número de vezes que um jogador desarmou um adversário.
BLK	Tocos: número de bloqueios realizados em arremessos adversários.
TOV	Erros ofensivos: número de perdas de posse de bola.
PF	Faltas pessoais cometidas.
TS%	Percentual de arremessos verdadeiros: mede a eficiência do arremesso considerando lances livres, cestas de 2 e 3 pontos.
ESTAT%	Percentual da estatística: estima a porcentagem da estatística em relação ao total disponível na partida.
ORtg	Rating ofensivo: estima a quantidade de pontos produzidos por um jogador ou time a cada 100 posses de bola.
DRtg	Rating defensivo: estima a quantidade de pontos cedidos por um jogador ou time a cada 100 posses de bola.

Fonte: Walsh e Joshi (2024), com modificações

APÊNDICE B – Mercados de apostas considerados no estudo

Tabela 12 – Mercados de apostas considerados no estudo

Mercado	**Número de apostas de valor**
Pontos do jogador	23647
Total de pontos na partida	17172
Total de pontos da equipe	9496
Total de pontos da equipe no 1º tempo	6494
Total de pontos no 1º tempo	6115
Pontos + rebotes do jogador	5540
Pontos + rebotes + assistências do jogador	5079
Pontos + assistências do jogador	4520
Total de pontos da equipe no 1º quarto	4095
Total de pontos no 1º quarto	3674
Rebotes do jogador	1994
Rebotes + assistências do jogador	1794
Assistências do jogador	1708
Total de pontos no 2º quarto	1630
Total de pontos da equipe no 2º quarto	1614
Cestas de três convertidas pelo jogador	690
Vitória da equipe	671
Tocos do jogador	424
Vitória no 1º tempo	385
Vitória no 1º quarto	328
Vitória no 2º quarto	237
Roubos de bola do jogador	75
Roubos + tocos do jogador	31
Turnovers do jogador	20
Primeiro cesto incluindo lances livres	6
Rebotes do jogador no 1º quarto	1
Pontos do jogador no 1º quarto	1
Vitória no 3º quarto	1
Total de pontos da equipe no 3º quarto	1
Total de pontos no 3º quarto	1

Fonte: Autor.