UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

LEONARDO ROCHA

ANÁLISE DA EFICIÊNCIA DOS ARREMESSOS DOS JOGADORES DA NBA EM MOMENTOS DE PRESSÃO

Revisado pelo Orientador

Assinatura do Orientador

Data ____/ ____ / 2025

PRESIDENTE PRUDENTE

2025

LEONARDO ROCHA

ANÁLISE DA EFICIÊNCIA DOS ARREMESSOS DOS JOGADORES DA NBA EM MOMENTOS DE PRESSÃO

Relatório Final para Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/Unesp para aproveitamento na disciplina TCC 1.
Orientador(a): Prof. Dr. Guilherme Aparecido Santos Aguilar.

PRESIDENTE PRUDENTE
2025

R672a Rocha, Leonardo
Análise da eficiência dos arremessos dos jogadores da NBA em momentos de pressão / Leonardo Rocha. -- Presidente Prudente, 2025
42 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências e Tecnologia, Presidente Prudente
Orientador: Guilherme Aparecido Santos Aguilar

1. Análise de dados. 2. Redes neurais (computação). 3. NBA. 4. Modelagem Preditiva. 5. Momentos de Pressão. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

TERMO DE APROVAÇÃO

Leonardo Rocha

ANÁLISE DA EFICIÊNCIA DOS ARREMESSOS DOS JOGADORES DA NBA EM MOMENTOS DE PRESSÃO

Relatório de Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

gov.br Documento assinado digitalmente
GUILHERME APARECIDO SANTOS AGUILAR
Data: 02/12/2025 09:16:37-0300
Verifique em https://validar.iti.gov.br

Orientador:

_____

Prof. Dr. Guilherme Aparecido Santos Aguilar
Departamento de Estatística

gov.br Documento assinado digitalmente
RICARDO PUZIOL DE OLIVEIRA
Data: 02/12/2025 09:25:31-0300
Verifique em https://validar.iti.gov.br

_____

Prof. Dr. Ricardo Puziol de Oliveira
Departamento de Estatística

Presidente Prudente, 02 de dezembro de 2025.

AGRADECIMENTOS

À minha família, agradeço pelo amor incondicional, pelo apoio e por cada gesto de incentivo ao longo desta jornada. Vocês foram meu alicerce nos momentos de desafio e minha maior inspiração para seguir em frente.

A Deus, expresso minha profunda gratidão pelo dom da vida, pela força concedida em cada etapa deste trabalho e pela sabedoria que me guiou desde o início até a conclusão deste TCC.

Ao meu professor orientador, Dr. Guilherme Aparecido Santos Aguilar, agradeço pela paciência, pelos ensinamentos e pelo olhar crítico que sempre soube orientar meus caminhos com confiança e entusiasmo. Seu comprometimento e dedicação foram fundamentais para o desenvolvimento deste estudo.

A todos que, direta ou indiretamente, colaboraram para a realização deste trabalho, meu muito obrigado.

RESUMO

Este trabalho enfatiza a relevância da coleta e análise de dados no esporte, ilustrada pelo caso de sucesso dos Oakland Athletics na MLB, que, mesmo com recursos limitados, adotaram a estatística para contratar jogadores compatíveis e alcançar alto nível competitivo. No basquete, a análise de dados também tem se difundido, conforme Schumaker et al. (2010), sendo fundamental selecionar informações relevantes para detectar padrões de desempenho. A aplicação de Inteligência Artificial (IA) no esporte, particularmente na NBA, tem permitido otimizar estratégias de jogo e métodos de treinamento, como exemplo o Advanced Scout (Bhandari, 1997). Além disso, algoritmos de IA têm sido utilizados para personalizar o treinamento de jogadores (Yoon et al., 2019). Com base em Metulini e Le Carre (2020), o estudo propõe modelar a eficiência dos arremessos de jogadores profissionais de basquete em situações de pressão, identificando as variáveis que mais influenciam o erro dos arremessos. Para isso, serão extraídos dados da temporada regular 2022/2023 da NBA, disponibilizados pela BigDataBall Company, e criadas oito variáveis indicativas de pressão (por exemplo, Shot_clock, Diff_p, Ratio, Miss.pl, Shot.type, Time, Quarter e Poss.type). A metodologia envolve análise descritiva e aplicação de redes neurais, com uso da função ReLU. Espera-se identificar como diferentes cenários de dificuldade afetam a assertividade dos arremessos, gerar previsões precisas do desempenho em situações de pressão com o objetivo de contribuir para o avanço das aplicações de técnicas estatísticas e de aprendizado de máquina no esporte e em outras áreas de decisão sob pressão.

Palavras-chave: NBA, Redes Neurais, Análise de Dados, Modelagem Preditiva e Momentos de Pressão.

ABSTRACT

This work emphasizes the relevance of data collection and analysis in sports, illustrated by the success of the Oakland Athletics in the MLB, who, despite limited resources, adopted statistical methods to recruit suitable players and achieve a high level of competitiveness. In basketball, data analysis has also become widespread, as noted by Schumaker et al. (2010), with the careful selection of relevant information being essential to detecting performance patterns. The application of artificial intelligence (AI) in sports—particularly in the NBA—has enabled the optimization of game strategies and training methods, exemplified by the Advanced Scout system (Bhandari, 1997). Furthermore, AI algorithms have been employed to personalize player training (Yoon et al., 2019). Based on Metulini and Le Carre (2020), this study proposes to model the efficiency of professional basketball players' shot-making under pressure by identifying the variables that most influence shot outcome. To this end, data from the 2022/2023 regular NBA season provided by BigDataBall Company will be extracted and eight pressure-indicating variables will be created (e.g., Shot_clock, Diff_p, Ratio, Miss.pl, Shot.type, Time, Quarter, and Poss.type). The methodology involves descriptive analysis and the application of neural networks using the ReLU activation function. It is expected that this research will reveal how different pressure scenarios affect shot accuracy and generate precise performance forecasts under pressure, thereby contributing to the advancement of statistical and machine-learning techniques in sports and other high-pressure decision-making contexts.

Keywords: NBA; Neural Networks; Data Analysis; Predictive Modeling; Pressure Situations.

Sumário

1 INTRODUÇÃO

Em Lewis (2004), é possível notar a importância do levantamento dos dados, uma vez que essa ideia fez com que um time de baseball (Oakland Athletics), no qual se encontrava em uma má fase chegasse a um grande nível na MLB (Major League Baseball). A equipe possuía recursos financeiros escassos e sofriam com a perda de seus principais atletas. Entre todas essas dificuldades, o técnico enxergou uma estratégia eficaz através da análise estatística e isso fez com que a tornasse um time competitivo entre a elite. Eles conseguiram realizar análises pontuais para as melhores contratações de novos jogadores que se enquadravam no padrão da equipe. Já no basquete, diversas técnicas são aplicadas com a missão de otimizar os dados.

De acordo com Schumaker et al. (2010), a análise de dados está cada vez mais comum no meio esportivo, milhares de dados são absorvidos e analisados, independentemente do esporte. Através da coleta, processamento e interpretação dos dados é possível obter grandes informações a respeito dos jogadores e das equipes, podendo alterar seu desempenho de forma eficaz. Entretanto, a dificuldade não se encontra na absorção dos dados, mas sim em quais dados coletar de forma que ele seja relevante para o estudo, por exemplo, para encontrar padrões em jogadores ou equipes. O uso de técnicas de análises estatísticas favorece nas respostas obtidas e aumenta a chance de uma tomada de decisão correta. A técnica de aprendizado de máquina e simulação vem sendo muito utilizada por técnicos com intuito de encontrar a melhor estratégia para próximos jogos de seus times. Nos últimos anos, a aplicação da IA (Inteligência Artificial) no esporte tem despertado um interesse crescente entre pesquisadores e profissionais. No basquete, especificamente, a utilização de técnicas de IA tem se mostrado promissora não apenas na análise do desempenho dos jogadores e equipes, mas também na otimização de estratégias de jogo e no desenvolvimento de sistemas de treinamento mais eficazes.

Na National Basketball Association (NBA) os treinadores e jogadores possuem acesso ao Advanced Scout. Ele foi criado em 1997 por um cientista norte-americano chamado Inderpal Bhandari e seus colaboradores, pertencentes ao IBM´s Watson Research Center. Este programa é utilizado pelas 20 equipes da NBA e se faz muito importante para todas, ele colabora de diferentes formas, como por exemplo detectando padrões de comportamento ou até mesmo respondendo questões pré-definidas. Em Perše et al. (2006), os autores apresentam um sistema de

reconhecimento automático de ações no basquete, utilizando técnicas avançadas de visão computacional e aprendizado de máquina. Esse sistema foi capaz de identificar e classificar diferentes tipos de jogadas e movimentos dos jogadores em tempo real, fornecendo uma análise detalhada do desempenho durante as partidas. Além disso, a IA tem sido aplicada no desenvolvimento de modelos preditivos para predizer o resultado de jogos de basquete com precisão cada vez maior. Em um estudo conduzido por Li et al. (2021), os pesquisadores propuseram um modelo de previsão de resultados de jogos de basquete baseado em redes neurais convolucionais, que levava em consideração uma gama de variáveis, incluindo estatísticas de jogadores, histórico de confrontos e condições de jogo.

Outro aspecto importante da aplicação da IA no basquete é a personalização de estratégias de treinamento e desenvolvimento de jogadores. Pesquisas como a realizada por Yoon et al. (2019) exploraram o uso de algoritmos de IA para analisar o estilo de jogo de jogadores individuais e identificar áreas específicas de melhoria. Essa abordagem permite que os treinadores adaptem seus métodos de ensino de acordo com as necessidades individuais de cada jogador, maximizando seu potencial de desenvolvimento.

Com base no artigo de Metulini e Le Carre (2020), o estudo proposto será realizado com o intuito de se obter um modelo que apresentará a eficiência dos arremessos de jogadores de basquete profissionais, levando em consideração a dificuldade apresentada para se realizar um arremesso e como o jogador reage em diferentes situações que o jogo lhe proporciona. “O objetivo é identificar as variáveis com maior influência no acerto ou erro de arremessos realizados pelos jogadores e por meio de previsões buscar quais jogadores possuem a maior probabilidade de acerto em cada tipo de arremesso (1 ponto, 2 pontos ou 3 pontos) em momentos de pressão.

As técnicas que utilizaremos serão regressão logística multinominal e redes neurais, porém vale ressaltar que o propósito é resolver o problema, então caso essas técnicas não sejam eficazes é possível a utilização de outras. Essas técnicas serão investigadas através do livro Izbicki e dos Santos (2020) e Fávero, Luiz Paulo, e Patrícia Belfiore (2017).

1.1 Objetivos

Este estudo tem como objetivo geral modelar o desempenho de arremessos de jogadores de basquete em situações de pressão, aplicando técnicas estatísticas e de aprendizado de máquina. Pretende-se compreender como diferentes cenários de dificuldade, presentes durante os jogos, influenciam a assertividade dos arremessos.

1.2 Objetivos específicos

O primeiro objetivo é identificar os fatores que aumentam a pressão sobre os jogadores durante os arremessos, o segundo é analisar o impacto de cada fator de dificuldade no desempenho individual dos jogadores e desenvolver um modelo estatístico que integre essas dificuldades como covariáveis para prever a eficácia dos arremessos. O terceiro consiste em avaliar a precisão do modelo para estimar o desempenho dos jogadores em situações de pressão e por fim o quarto objetivo é comparar a variação no desempenho entre os jogadores em diferentes cenários de dificuldade.

Esses objetivos estabelecem uma base para examinar como a pressão influencia o desempenho dos atletas e ajudar a desenvolver estratégias mais robustas para prever o sucesso dos arremessos.

1.3 Justificativa

A ideia de realizar esse projeto se dá pela crescente demanda de análises de desempenho esportivo baseado em dados. Durante os jogos, os jogadores enfrentam cenários complexos que desafiam sua habilidade de tomar decisões rápidas e realizar arremessos práticos. Compreender os fatores que mais impactam a precisão dos arremessos em situações de pressão pode fornecer insights valiosos para treinadores, atletas e equipes esportivas, permitindo o desenvolvimento de estratégias e treinamentos mais eficazes.

Esse trabalho pode contribuir para a sociedade ao ampliar o campo de aplicação de técnicas estatísticas e de aprendizado de máquina, demonstrando como essas ferramentas podem ser usadas para modelar e prever comportamentos humanos em contextos de alta pressão. Além disso, os conhecimentos gerados podem ser adaptados para outros cenários em que uma tomada de decisão sob pressão seja essencial, como áreas da saúde, gestão e educação.

2 REVISÃO BIBLIOGRÁFICA

A literatura sobre análise de desempenho no basquete combina estudos estatísticos, computacionais e comportamentais. Metulini e Le Carré (2019) destacam a influência de variáveis contextuais, como diferença de pontos e tempo restante, na probabilidade de sucesso dos arremessos. Trabalhos na área de ciência de dados aplicada ao esporte (Schumaker et al., 2010; Goldsberry, 2019) indicam que a modelagem preditiva pode auxiliar na compreensão de padrões de eficiência e tomada de decisão.

No campo do aprendizado de máquina, redes neurais têm sido utilizadas para reconhecer jogadas, prever resultados e analisar movimentações de jogadores (Yoon et al., 2019; Li et al., 2021). Conforme Haykin (2007), essas redes são especialmente úteis quando há relações não lineares entre as variáveis, característica presente em contextos esportivos.

Além dos aspectos técnicos, estudos de psicologia do esporte, como Baumeister (1984), investigam fenômenos ligados à pressão, sugerindo que o desempenho pode variar conforme fatores emocionais e situacionais.

3 METODOLOGIA

3.1 Introdução referente a redes neurais

O uso de redes neurais começou com a percepção de que o cérebro humano processa informações de maneira muito diferente dos computadores digitais tradicionais. O cérebro é um sistema complexo, não linear e altamente paralelo, formado por neurônios que trabalham juntos para realizar tarefas como reconhecimento de padrões e controle de movimentos com mais rapidez do que qualquer computador digital.

Um neurônio em desenvolvimento permite que o sistema nervoso se adapte ao ambiente. De forma geral, uma rede neural é um sistema criado para imitar como o cérebro realiza certas tarefas. Ela pode ser programada em um computador e é composta por várias “unidades de processamento”, chamadas de neurônios, que se conectam entre si para formar uma estrutura mais complexa.

Assim, uma rede neural é um processador distribuído e paralelo, formado por neurônios, com a capacidade de aprender com experiências e guardar esse conhecimento para uso futuro.

O aprendizado da rede acontece por meio de algoritmos de aprendizagem, que ajustam os pesos sinápticos (os valores que ligam os neurônios) de forma organizada, buscando alcançar um objetivo determinado.

As redes neurais se destacam por sua estrutura paralela e pela capacidade de aprender e generalizar, ou seja, conseguir responder corretamente mesmo a dados que não foram vistos durante o treinamento.

As principais propriedades das redes neurais incluem:

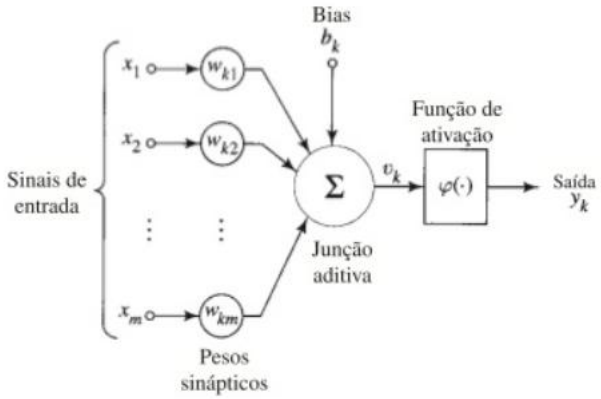
1. Não-linearidade: Quando a rede possui neurônios com comportamento não linear, ela também se torna não linear. Essa característica está presente em toda a estrutura da rede.
2. Mapeamento entre entrada e saída: A rede recebe exemplos e ajusta seus pesos para reduzir a diferença entre a saída desejada e a saída obtida.
3. Adaptabilidade: Redes neurais treinadas em certos ambientes podem ser facilmente ajustadas para lidar com mudanças no ambiente, mantendo sua estabilidade e bom desempenho.
4. Resposta baseada em evidências: Além de identificar padrões, a rede pode indicar o grau de confiança que tem em cada decisão tomada.
5. Uso de informação contextual: Cada neurônio é influenciado pela atividade dos outros neurônios da rede, o que permite maior integração da informação.
6. Tolerância a falhas: Mesmo que um neurônio pare de funcionar, a rede continua operando, embora com possível perda de qualidade na resposta.
7. Implementação em VLSI (Very Large Scale Integration): Por ter uma estrutura paralela, a rede pode ser muito rápida em certas tarefas. A tecnologia VLSI permite implementar redes com comportamentos complexos de maneira organizada e eficiente.

3.2 Modelos de um neurônio

Um neurônio é uma unidade de processamento de informação fundamental para o funcionamento de uma rede neural.

A seguir, temos o modelo de um neurônio:

Figura 1: *Funcionamento de uma rede neural.*



Os $x´s$ são os sinais de entradas, ou seja, os dados que alimentam o neurônio, podendo ser características do objeto de estudo. Cada entrada é multiplicada por um peso associado. Os pesos, também conhecido como pesos sináptico, determinam a importância de cada entrada.

"***Bias***", (vício, em inglês), é um valor adicional somado ao total, que ajuda o modelo a se ajustar melhor. Ele desloca a função de ativação para que o modelo consiga aprender padrões mais complexos.

O neurônio irá calcular a soma ponderada das entradas mais os **bias**, como exemplificado na fórmula a seguir:

$$u_k = \sum_{i=1}^{m} x_i \, w_{ki} + b_k$$

O valor de u_k é utilizado na próxima etapa. Nesta etapa entra a função não linear do neurônio.

A função de ativação, representada por $\varphi(u)$, define a saída de um neurônio em termos do campo local induzido u. A seguir, temos três funções de ativação:

1. Função Limiar. Para este tipo de função de ativação, temos:

$$\varphi(u) = \begin{cases} 1 \; se \; v \geq 0 \\ 0 \; se \; v < 0 \end{cases}$$

Tal neurônio é referido na literatura como o modelo de McCulloch-Pitts. Neste modelo, a saída de um neurônio assume o valor 1, se o campo local induzido daquele neurônio é não-negativo, e 0 caso contrário.

2. Função Linear por Partes. Para a função linear temos:

$$\varphi(u) = \begin{cases} 1, u \geq +\frac{1}{2} \\ u, +\frac{1}{2} > u > -\frac{1}{2} \\ 0, \quad u \leq -\frac{1}{2} \end{cases}$$

Onde assume-se que o fator de amplificação dentro da região linear de operação é a unidade.

3. Função Sigmóide. A função Sigmóide, cujo gráfico tem a forma de"S", é a forma mais comum de função de ativação utilizada na criação de redes neurais artificiais. Ela possui um balanceamento adequado entre comportamento linear e não linear. Um exemplo de função Sigmóide é a função logística, definida por:

$$\varphi(u) = \frac{1}{1 + \exp(-au)}$$

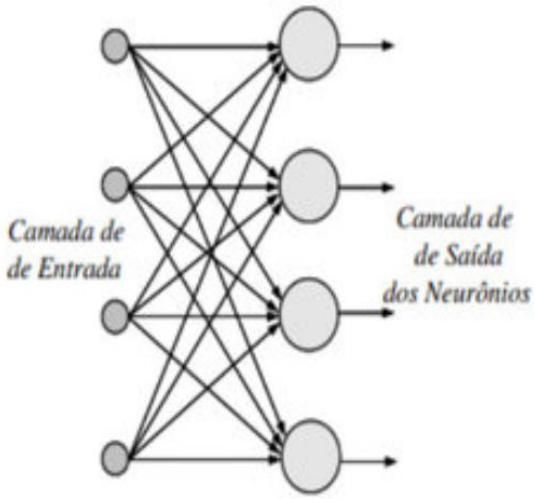
Onde "a" é o parâmetro de inclinação da função Sigmóide. A cada variação de "a" obtemos funções Sigmóides com diferentes inclinações.

3.3 Redes alimentadas adiante com camada única

Em redes neurais, os neurônios estão organizados em camadas. Na forma mais simples de uma rede em camadas, elas fluem apenas em uma direção: de uma camada de entrada para uma camada de saída (nós computacionais).

A figura abaixo apresenta exatamente o que acontece. Neste caso temos 4 nós, tanto na camada de entrada quanto de saída. Esta é uma rede conhecida como "rede de camada única", sendo que a designação "camada única" se refere à camada de saída de nós computacionais (neurônios). Os nós na camada de entrada são as variáveis explicativas e os nós na camada de saída a resposta. As setas ligam a variável a uma resposta e cada ligação tem um peso que indica o quanto aquela entrada influencia na resposta.

Figura 2 *- Rede alimentada adiante com camada única.*



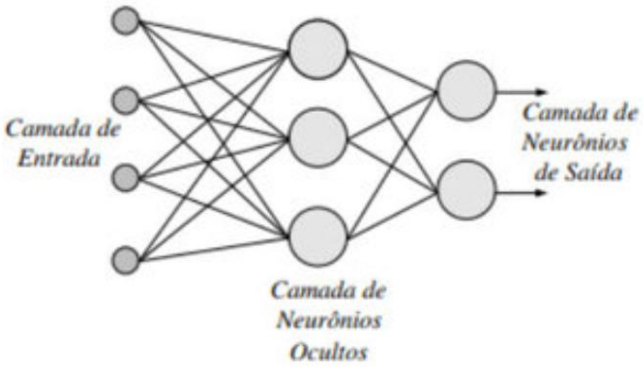
3.4 Redes alimentadas diretamente com múltiplas camadas.

Uma rede neural alimentada se distingue pela presença de uma ou mais camadas ocultas. A função dos neurônios ocultos é intervir entre a entrada externa e a saída da rede de uma maneira mais útil. Adicionando uma ou mais camadas ocultas, capacitamos nossa rede para extrair estatísticas de ordem elevada. A habilidade de os neurônios ocultos extraírem estatísticas de ordem elevada é particularmente valiosa quando o tamanho da camada de entrada é grande.

Os nós de fonte da camada de entrada da rede fornecem os respectivos elementos do padrão de ativação (vetor de entrada), que constituem os sinais de entrada aplicado aos neurônios (nós computacionais) na segunda camada. Os sinais de saída da segunda camada são utilizados como entradas para a terceira camada, e assim por diante para o resto da rede. O conjunto de sinais de saída dos neurônios da camada de saída (final) da rede constitui a resposta global da rede para padrão de ativação fornecido pelos nós de fonte da camada de entrada (primeira).

A imagem 3 apresenta uma rede alimentada adiante totalmente conectada com uma camada oculta e uma camada de saída.

Figura 3: *Rede alimentada adiante totalmente conectada*



As camadas estarem totalmente conectadas garante que todas as entradas influenciem em todas as saídas, passando por todos os caminhos possíveis, assim a rede possui mais flexibilidade para aprender padrões complexos.

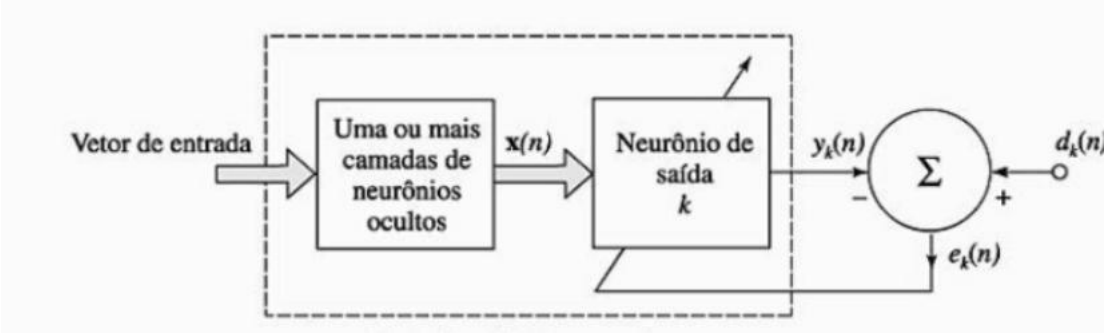
3.5 Processo de aprendizagem

Um dos principais objetivos de uma rede neural é a habilidade de aprender a partir de seu ambiente e melhorar seu desempenho através de aprendizagem. O processo de aprender de uma rede neural ocorre através de um processo interativo de ajustes aplicados aos pesos sinápticos e níveis de **bias**.

Um conjunto preestabelecido de regras bem-definidas para a solução de um problema de aprendizagem é denominado um algoritmo de aprendizagem. Temos um conjunto de ferramentas representado por uma variedade de algoritmos de aprendizagem, cada qual oferecendo vantagens específica

Imagine um neurônio k que constitui o único nó computacional da camada de saída de uma rede neural alimentada adiante, como a imagem a seguir:

Figura 4: *Redes de múltiplas camadas alimentadas adiante*



O neurônio k é acionado por um vetor de sinal x(n) produzido por uma ou mais camadas de neurônios ocultos, que são acionados por um vetor de entrada aplicado aos nós de fonte (camada de entrada) da rede neural. O argumento n representa o instante de tempo discreto. O sinal de saída do neurônio k é representado por $y_k(n)$. Este sinal de saída, representando uma única saída da rede neural, é comparado com uma resposta desejada, representada por $d_k(n)$. Consequentemente, é produzido um sinal de erro, representado por $e_k(n)$. O erro aciona um mecanismo que a ideia é aplicar uma sequencia de ajustes corretivos aos pesos sinápticos do neurônio k. Os ajustes corretivos são projetados para aproximas passo a passo o sinal de saída $y_k(n)$ da resposta desejada $d_k(n)$. Este é o objetivo é alcançado minimizando uma função de custo.

No presente trabalho, a função de custo empregada foi a Cross-Entropy, amplamente utilizada em problemas de classificação. A Cross-Entropy mede a discrepância entre as probabilidades previstas pelo modelo e os valores reais observados nas classes do conjunto de dados. Em tarefas de classificação binária, ela é definida como:

$$LogLoss = -[\, y \log(p) + (1 - y) \log(1 - \mathrm{p})]$$

onde y representa a classe verdadeira (0 ou 1) e p é a probabilidade estimada pela rede para a classe positiva. Essa função penaliza principalmente situações em que o modelo erra com alta confiança, produzindo gradientes maiores e, portanto, permitindo ajustes mais eficientes aos pesos sinápticos. Por esse motivo, a Cross-Entropy promove um aprendizado mais rápido, estável e estatisticamente apropriado para modelos que produzem saídas probabilísticas, como é o caso das redes neurais utilizadas neste estudo.

Assim, o processo de aprendizagem da rede neural consiste em ajustar seus parâmetros internos de maneira a minimizar a função de custo baseada em entropia cruzada, o que resulta em melhores predições e em maior capacidade de discriminar corretamente as classes diante das informações disponíveis.

3.6 Perceptrons de múltiplas camadas

A rede consiste de um conjunto de unidades sensoriais (nós de fonte) que constituem a camada de entrada, uma ou mais camadas ocultas de nós computacionais e uma camada de saída de nós computacionais. O sinal de entrada

se propaga para frente através da rede, camada por camada. Estas redes neurais são normalmente chamadas de perceptrons de múltiplas camadas.

Um perceptron de múltiplas camadas tem três características distintas:

1. O modelo de cada neurônio da rede inclui uma função de ativação não linear. Uma forma normalmente utilizada de não-linearidade que satisfaz esta exigência é uma não-linearidade sigmóide definida pela função logística:

$$y_j = \frac{1}{1 + \exp(-v_j)}$$

Onde v_j é o campo local induzido (valor calculado na entrada de um neurônio antes da aplicação da ativação) do neurônio j, e y_j é a saída do neurônio.

2. As redes neurais artificiais possuem camadas intermediárias chamadas de camadas ocultas, que ficam entre a entrada e a saída. Essas camadas são responsáveis por processar e transformar os dados de entrada em representações mais abstratas, facilitando a identificação de padrões complexos. Cada camada oculta extrai características específicas dos dados, permitindo que a rede compreenda e resolva tarefas sofisticadas.
3. A rede exibe um alto grau de conectividade, determinado pelas sinapses da rede. Uma modificação da rede requer uma mudança na população das conexões sinápticas ou de seus pesos.

Através dessas características e habilidade de aprender por treinamento que o perceptron de múltiplas camadas deriva seu poder computacional.

O desenvolvimento do algoritmo de retro propagação representa um marco nas redes neurais, pois fornece um método computacional eficiente para o treinamento de perceptrons de múltiplas camadas.

A retro propagação permite que a rede neural aprenda a partir dos dados de treinamento, ajustando seus parâmetros internos para melhorar sua capacidade de generalização e previsão.

3.7 Função de ativação

Para este trabalho, será utilizada a função de ativação sigmóide na camada de saída. A ideia de uma função de ativação é trazer a não linearidade no modelo, permitindo que a rede aprenda de forma mais eficaz.

A função sigmóide é definida da seguinte forma:

$$\varphi'\left(v_j(n)\right) = \frac{1}{1 + \exp\left(-a v_j(n)\right)}$$

Em que $a > 0$ e $-\infty < v_j(n) < \infty$. Do qual $v_j(n)$ é o campo local induzido do neurônio j. De acordo com a não – linearidade, a amplitude de saída se encontra dentro de um intervalo $0 \leq y_i \leq 1$.

Para as camadas ocultas, utilizou-se a função de ativação Leaky ReLU (Leaky Rectified Linear Unit).

A Leaky ReLU é uma variação da ReLU tradicional, amplamente adotada em redes neurais devido à sua simplicidade, eficiência computacional e capacidade de introduzir não linearidade ao modelo.

A função Leaky ReLU é definda como:

$$f(x) = \begin{cases} x, & x > 0 \\ \propto x, & x \leq 0 \end{cases}$$

onde $\propto$ é um pequeno valor positivo, tipicamente entre 0,01 e 0,1.

A principal diferença em relação à ReLU tradicional está no tratamento das entradas negativas. Enquanto a ReLU simplesmente zera todos os valores negativos, a Leaky ReLU permite uma pequena inclinação quando $x \leq 0$. Isso evita o problema conhecido como *"dead ReLU"*, no qual neurônios podem parar de aprender ao receber apenas gradientes nulos.

3.8 Modelo de treinamento

As redes neurais são treinadas para ajustar os parâmetros (pesos sinápticos) do modelo, com o objetivo de minimizar o erro entre as previsões e os valores reais.

O Gradiente Descendente é o algoritmo utilizado para calcular a derivada da função de erro, ajustando os pesos na direção que melhora a performance do modelo.

Considerando que o conjunto de dados utilizado neste trabalho é composto por todos os arremessos realizados na temporada regular da NBA de 2022/2023, o método de aprendizado escolhido é o Mini-Batch Learning (Aprendizado por Mini-Lotes).

O funcionamento dessa técnica consiste em dividir o conjunto de dados em pequenos lotes, chamados de mini-batches. A cada lote, os pesos do modelo são ajustados. Para aplicá-la, o dataset deve ser particionado em várias partes menores.

Cada mini-lote é então passado pelo modelo, em uma etapa chamada de Forward Pass, onde são feitas as previsões e calculada a função de custo (que mede o erro da previsão). Em seguida, calcula-se o gradiente da função de custo em relação aos pesos, processo conhecido como Backpropagation. Por fim, os pesos são atualizados utilizando um algoritmo de otimização, como o Gradiente Descendente, promovendo o aprendizado do modelo.

Para aprimorar o processo de convergência e evitar problemas como sobreajuste (overfitting) ou estagnação do aprendizado, foram utilizados dois callbacks essenciais durante o treinamento: EarlyStopping e ReduceLROnPlateau. Os callbacks são rotinas automáticas que monitoram o desempenho do modelo durante as épocas e realizam intervenções sempre que necessário.

O EarlyStopping foi configurado para acompanhar a métrica AUC no conjunto de validação. Quando essa métrica deixa de apresentar melhoria após um número definido de épocas, o treinamento é interrompido antecipadamente, garantindo que o modelo não continue ajustando seus pesos sem ganho real de desempenho. Além disso, a opção restore_best_weights assegura que os pesos correspondentes ao melhor resultado obtido durante o treinamento sejam recuperados ao final do processo.

Complementarmente, o ReduceLROnPlateau foi empregado para ajustar dinamicamente a taxa de aprendizagem (learning rate) do otimizador Adam. Sempre que a evolução da métrica de validação demonstra sinais de estagnação, o callback reduz automaticamente o valor da taxa de aprendizagem, permitindo que o modelo realize atualizações mais delicadas nos pesos e alcance convergência de forma mais precisa.

3.9 Avaliação do modelo

Assim que o modelo for testado é muito importante avaliarmos como foi sua eficiência. Contudo a matriz de confusão é uma parte muito importante nesse processo. Ela permite analisar de forma rápida o desempenho do modelo. Os valores que compõem a matriz são obtidos fornecendo os segmentos do conjunto de teste ao método de classificação e comparando sua predição com a classe correta de cada segmento.

Após a comparação os valores são classificados em quatro possíveis opções:

1. **Verdadeiro Positivo (VP):** segmentos que pertencem a classe positiva e foram classificados como positivos.
2. **Falso Positivo (FP):** segmentos que pertencem a classe negativa e foram classificados como positivos.
3. **Falso Negativo (FN):** segmentos que pertencem a classe positiva e foram classificados como negativos.
4. **Verdadeiro Negativo (VN):** segmentos que pertencem a classe negativa e foram corretamente classificados como negativos.

O foco é sempre acertar os VP e VN, sendo assim, nossa acurácia será boa.

A acurácia é uma métrica de desempenho de modelos que indica a proporção de previsões corretas em relação ao total de casos avaliados. Ela é definida por:

$$\text{Acurácia} = \frac{\text{VP} + \text{VN}}{VP + FP + FN + VN}$$

Relacionado a matriz de confusão, também conseguimos calcular o Recall, que é dado pela seguinte equação:

$$Recall = \frac{VP}{VP + FN}$$

Essa métrica busca mensurar a capacidade do modelo de identificar corretamente os verdadeiros positivos, ou seja, de “todos os exemplos que são realmente positivos, quantos meu modelo identificou corretamente?”.

Outra medida importante é a curva ROC (Receiver Operating Characteristc), gráfico que apresenta o desempenho do modelo de um classificador binário à medida que se varia o limiar da decisão. O eixo Y (vertical) é composto pela taxa de verdadeiros positivos e no eixo X (horizontal) a taxa de falsos positivos. A área da curva é conhecida como AUC (Area Under the Curve), quanto maior a área, ou seja, quanto mais próximo de 1 for o valor, melhor será o modelo.

4 RESULTADOS EXPLORATÓRIOS

4.1 Apresentação dos Dados

Os dados que estão sendo trabalhado são da temporada regular da NBA 2022/2023, sendo retirado do site BigDataBall Company (UK)

(https://www.bigdataball.com/). Os dados fornecem lance a lance de todos os jogos da temporada citada. Ao total está sendo utilizado 9 variáveis, conforme a tabele abaixo. Ao total a base é composto por 9 variáveis e 294.734 observações (linhas).

Tabela 1 – *Variáveis do Conjunto de Dados*

Variável	Tipo	Descrição
Shot	Categórica	Variável binária, que mostra se o jogador errou ou acertou o arremesso, 1 caso tenha acertado e 0 caso contrário
Shot_Clok	Numérica	Tempo restante (0–24 s) no cronômetro de posse de bola quando o arremesso foi feito (ex.: valor 12 indica que faltavam 12 s na posse).
Diff_p	Numérica	Diferença de pontos da equipe mandante em relação à visitante após o arremesso.
Ratio	Numérica	Razão de arremessos certos em relação ao total de arremessos da equipe até aquele momento (valor entre 0 e 1).
Miss_pl	categórica	Indica se o jogador acertou (1) ou errou (0) o arremesso anterior ao atual.
Shot_type	Categórica	Tipo de arremesso convertido: 1 ponto, 2 pontos ou 3 pontos.
Time	Numérica	Tempo (em segundos) até o final do quarto em que o arremesso foi tentado.
Quarter	Categórica	Período do jogo em que o arremesso foi realizado: 1°, 2°, 3° ou 4°.
Poss_type	Categórica	Tipo de posse quando o arremesso foi realizado, podendo ser original (arremesso dentro os 24s da posse) ou reset (arremesso após redefinição da posse da bola)

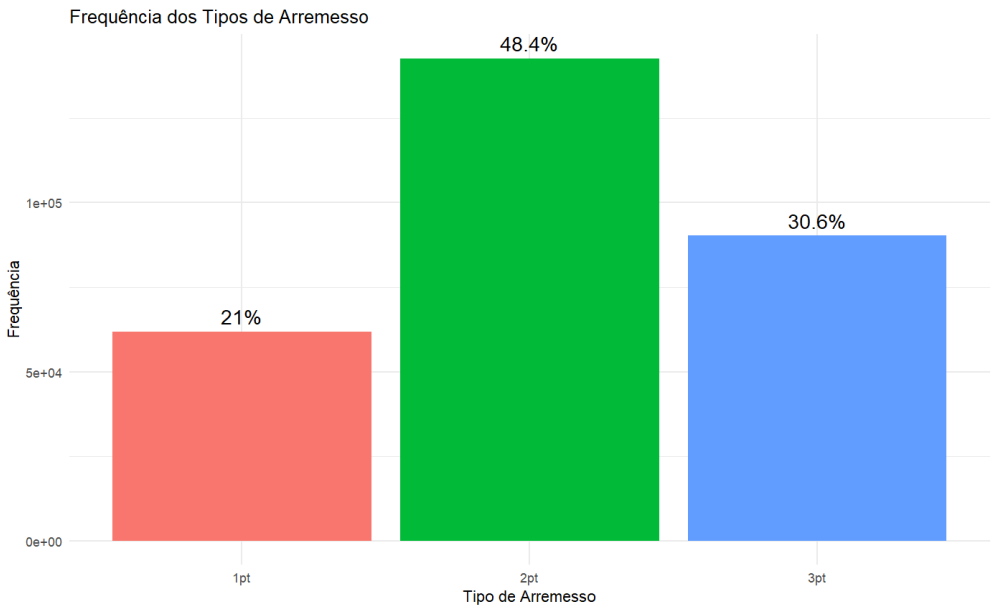
A base foi importada no software “R” e foi iniciado alguns tratamentos iniciais e análises exploratórias.

4.2 Análise Descritiva

O primeiro passo de análise é a descritiva, pois assim conseguimos entender como as variáveis se comportam e se há algo totalmente irreal na base. A primeira variável que trabalhei foi a “type_shot”.

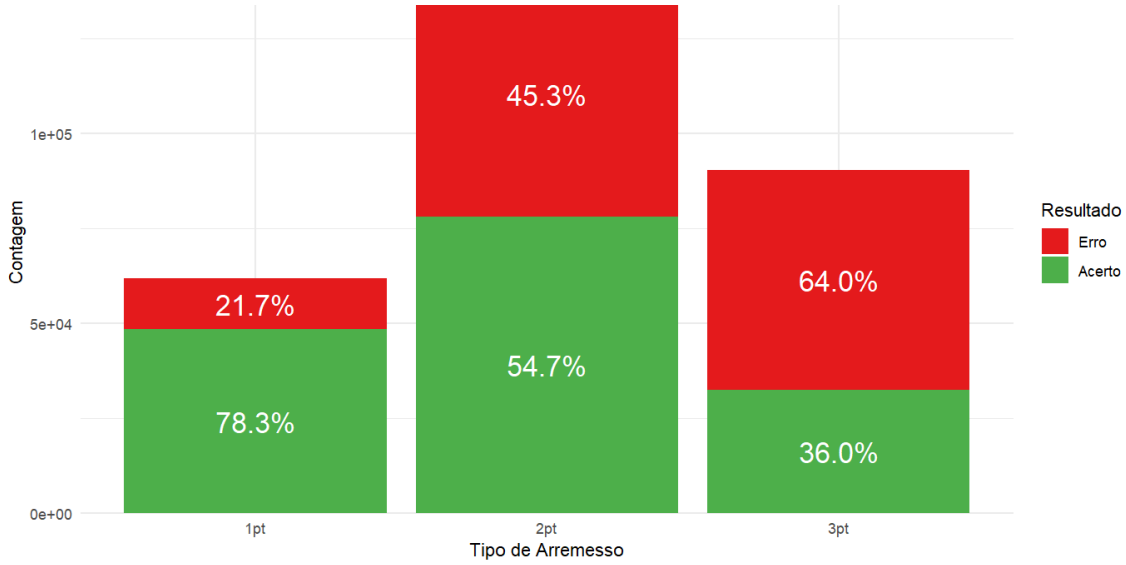
Realizando sua análise, é possível perceber que de modo geral, a maior parte dos arremessos são bolas de 2 pontos, representando 48,4% do total arremessado na temporada inteira. Em seguida tempos a tentativa de 3 ponto, representando 30,6% do total e por último arremessos de 1 ponto, que são os lances livros, representando 21% do total. Isso indica que a maioria dos arremessos realizados durante os jogos é de 2 pontos.

***Figura 5**: Frequência dos tipos de arremessos*



Dentre esses arremessos, conseguimos ver a relação do tipo de arremesso com a porcentagem de acerto com a imagem a seguir:

***Figura 6**: Erros vs Acertos por tipo de arremesso*

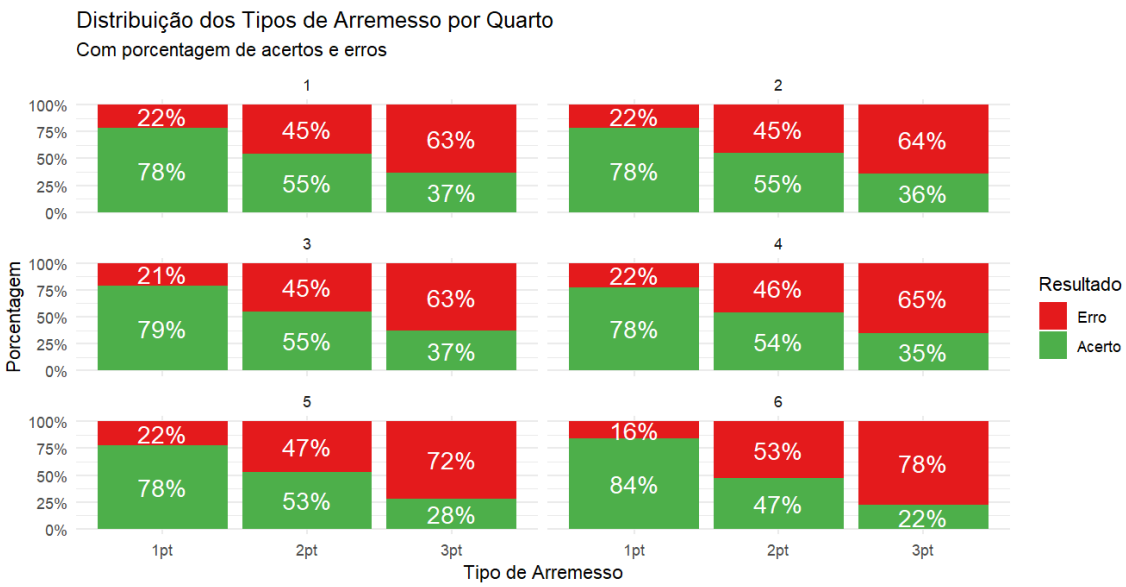


Do total geral de arremessos de 1 ponto, 78,3% são convertidos, podendo ser considerado um tipo de arremesso “fácil”, pela distância e por não ter algum adversário

marcando. Referente ao total geral dos arremessos de 2 pontos, 54,7% deles são acertados, mostrando que o aproveitamento é mais que a metade, porém um pouco equilibrado, isso pode nos mostrar que esse tipo de arremesso não é tão simples, então o fator pressão pode ter relação com a porcentagem de acerto. Por fim, falando dos arremessos de 3 pontos, o gráfico mostra que a taxa de erro é maior que a de acerto, 64% do total do total não é convertido em ponto, mostrando que esse tipo de arremesso é o mais difícil de se acertar, isso pode ser causado pela distância ou até mesmo pelos momentos de pressão, já que é o arremesso que mais vale ponto e pode ampliar ou diminuir a diferença de pontos das equipes. Portanto, a variável “type_shoot” pode ser bem relevante no estudo, uma vez que as taxas de acerto para cada tipo de chute são diferentes, principalmente para 1pt e 2pt, pois possuem uma “chance” de acerto maior.

Agora, trazendo essa análise de outra forma, será verificado como os tipos de arremessos se comportam ao longo dos períodos e a porcentagem de acerto em cada.

Figura 7: *Distribuição do tipo de arremesso por quarto*



Como pode-se ver, o volume de acertos de cada tipo de arremesso sofre poucas mudanças ao decorrer dos períodos. Os acertos de 1pt se mantém constante em todo período, apenas no último (6°) tem um aumento a porcentagem de acerto. O arremesso de 2 pontos também apresenta um padrão constante, em média a taxa de acerto em todos os períodos é de 53,1%, só não é maior pois no último há uma queda na porcentagem. Já o arremesso de 3pt é o que sofre mais mudanças, mas somente na prorrogação. Do primeiro ao quarto período (tempo normal) a média de acerto de

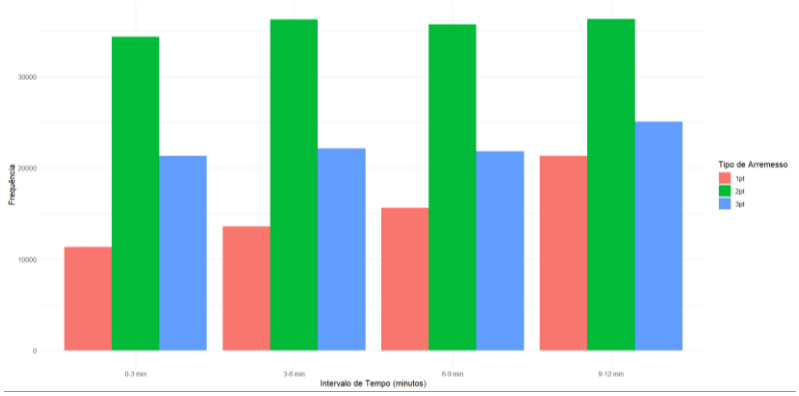
3 pontos é de 36,25%, já na prorrogação essa média diminui, indo para 25% de aproveitamento.

Analisando agora os arremessos por tempo, foi separado os 12 minutos de jogo em 4 partes, sendo:

- 0 – 3 minutos;
- 3 – 6 minutos;
- 6 – 9 minutos;
- 9 – 12 minutos;

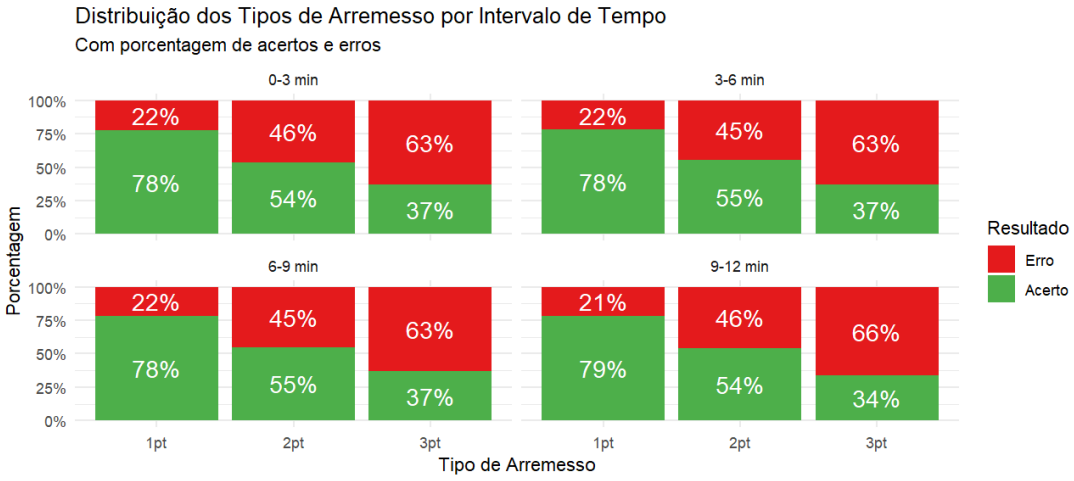
A frequência dos arremessos, de forma geral, é constante, havendo um leve aumento dos arremessos do intervalo de 9 a 12 minutos.

Figura 8 *– Frequência dos tipos de arremessos por intervalo de tempo*



Agora analisando os intervalos de tempo com a taxa de acerto por tipo de arremesso, tem-se o seguinte resultado:

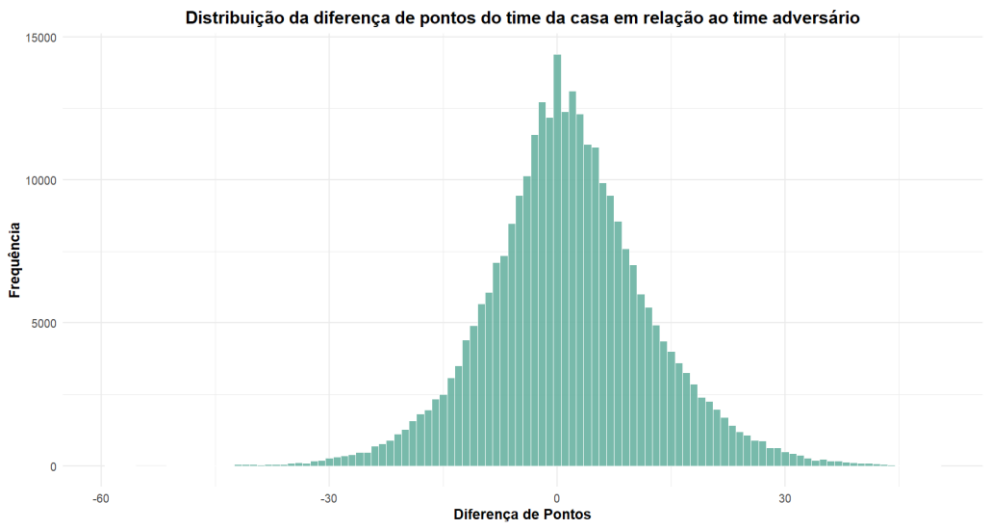
Figura 9*: Distribuição dos tipos de arremessos por intervalo de tempo*



É visto um equilíbrio na porcentagem de acertos para cada tipo de arremesso em cada intervalo de tempo, como mostrado na figura 9. Os jogadores costumam seguir com a mesma taxa de acertos em diferentes momentos do jogo. Isso mostra que o tempo é uma variável que pode não influenciar no acerto do arremesso do jogador.

A variável diferença de pontos é bem interessante. Trazendo um histograma, conseguimos analisar como ela se comporta.

***Figura 10**: Histograma diferença de pontos*



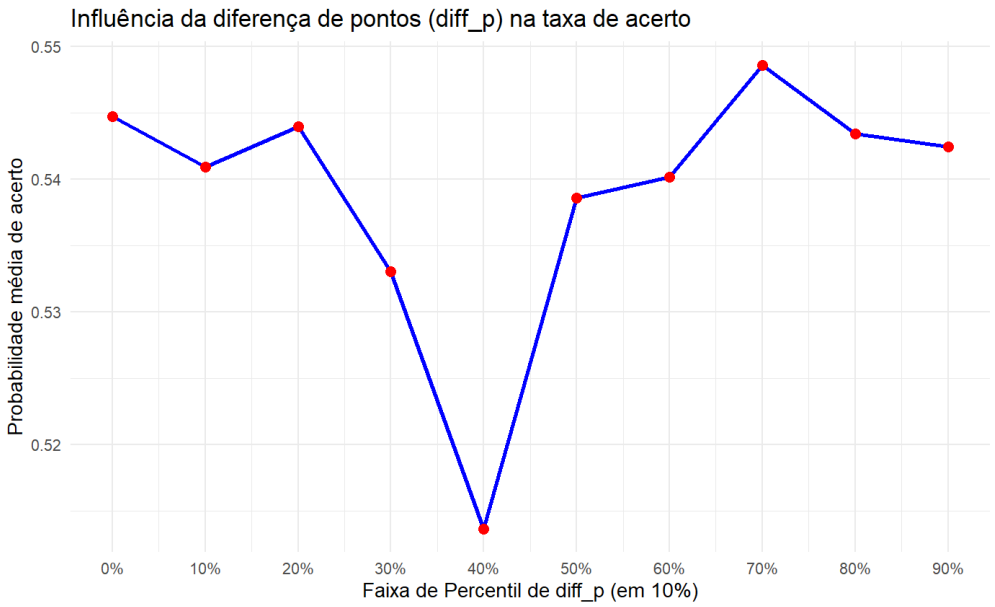
O histograma está praticamente centrado no 0, a média da diferença de pontos é 1,46, mostrando que as partidas de forma geral são bem competitivas, não havendo muita vantagem jogar fora ou dentro de casa.

Para investigar a existência de dependência entre o contexto da partida e o desempenho individual dos jogadores, foi analisada a relação entre a diferença de pontos do time da casa em relação ao visitante no momento do arremesso (diff_p) e a probabilidade de acerto (shot). A variável diff_p apresenta ampla variação, oscilando entre −59 e +50 pontos, refletindo cenários que vão desde grandes desvantagens até vantagens substanciais.

Com o intuito de identificar padrões de comportamento ao longo dessa distribuição, a amostra foi segmentada em dez grupos iguais (decis), definidos pelos percentis de 10% em 10%. Para cada grupo, calculou-se a probabilidade média de acerto. Os intervalos correspondentes aos percentis são: −59 a −11 (0–10%), −11 a −7 (10–20%), −7 a −4 (20–30%), −4 a −1 (30–40%), −1 a +1 (40–50%), +1 a +3 (50–

60%), +3 a +6 (60–70%), +6 a +10 (70–80%), +10 a +15 (80–90%) e +15 a +50 (90–100%). A Figura correspondente apresenta a evolução da média de acertos em cada intervalo.

Figura 11 *– Influência da Diferença de Pontos na Taxa de Acerto*



Os resultados indicam que a relação entre a diferença de pontos e o desempenho nos arremessos não é linear. Observa-se inicialmente que, em situações de grande desvantagem, a taxa média de acerto permanece relativamente elevada, sugerindo possível redução da pressão defensiva adversária ou maior liberdade para execução dos arremessos. Entretanto, à medida que a desvantagem diminui e o placar se aproxima do equilíbrio (intervalo entre −4 e +1 pontos, correspondente aos decis entre 30% e 50%), verifica-se a menor probabilidade de acerto dentre todas as faixas analisadas.

Esse comportamento forma um “vale” no centro da distribuição e coincide com momentos de maior intensidade competitiva, nos quais tanto a pressão psicológica quanto o rigor defensivo tendem a aumentar. Em jogos equilibrados, os atletas podem enfrentar decisões mais rápidas e arremessos menos confortáveis, o que explica a redução no desempenho ofensivo.

Por outro lado, quando o time da casa passa a liderar o placar, observa-se uma recuperação gradual da eficiência dos arremessos, alcançando o melhor desempenho no intervalo de +6 a +10 pontos (70–80%). Nessa condição, a equipe tende a atuar com maior confiança, encontrando arremessos mais bem selecionados e sob menor

pressão direta da defesa. Por fim, mesmo em situações de vantagem bastante elevada (acima de +10 pontos), a eficiência permanece acima da média geral, embora com leve estabilização, possivelmente em razão da entrada de reservas ou diminuição do ritmo competitivo.

De maneira geral, os achados demonstram que a diferença de pontos exerce influência significativa sobre o desempenho de arremessos, sendo capaz de explicar variações importantes na probabilidade de acerto. O resultado reforça que o contexto competitivo desempenha papel central na performance ofensiva no basquete, e que estados psicológicos associados ao equilíbrio ou desequilíbrio do placar podem afetar a tomada de decisão e a qualidade técnica dos arremessos. Esses padrões justificam a inclusão de diff_p como variável relevante nos modelos preditivos aplicados nas seções subsequentes deste trabalho.

5 Resultados

5.1 Modelo Redes Neurais

A partir da base original de dados, foi construída uma base reduzida contendo apenas as variáveis relevantes para a modelagem. A variável resposta selecionada foi shot, que indica o sucesso (1) ou fracasso (0) de um arremesso. As demais variáveis do conjunto de dados representam informações relacionadas ao contexto da jogada, incluindo características do lance e condições de pressão no momento do arremesso.

Inicialmente, procedeu-se ao pré-processamento dos dados. A variável dependente shot, originalmente categórica, foi convertida em valores numéricos por meio do método cat.codes, uma vez que redes neurais artificiais requerem que a variável resposta esteja representada de forma numérica para viabilizar o processo de otimização. Em seguida, foi construída a matriz de preditores (X) a partir da exclusão da coluna alvo, contendo apenas as variáveis explicativas.

As variáveis categóricas foram tratadas por meio da técnica de one-hot encoding, implementada via criação de variáveis dummies. Em cada variável categórica, uma das categorias é omitida (drop-first) para evitar problemas de multicolinearidade. Por exemplo, caso a variável type_shoot contenha três categorias distintas, são criadas duas variáveis indicadoras representando a presença/ausência

de cada categoria. Essa transformação é essencial, uma vez que redes neurais não conseguem interpretar diretamente variáveis categóricas sem codificação apropriada.

As variáveis de tempo foram transformadas em segundos para ser possível ser inserida no modelo.

Posteriormente, aplicou-se a padronização das variáveis numéricas utilizando o método StandardScaler, que transforma cada variável para média zero e desvio padrão igual a um. A padronização é fundamental em redes neurais, pois reduz diferenças de escala entre as variáveis e contribui para a estabilidade numérica durante o treinamento, permitindo que o algoritmo de otimização converja de forma mais eficiente.

Para divisão dos dados, utilizou-se a técnica de train-test split, separando 75% das observações para o conjunto de treinamento e os 25% restantes para o conjunto de teste. A divisão foi estratificada, garantindo que a proporção entre classes fosse preservada tanto no treinamento quanto no teste. Dentro do próprio conjunto de treinamento, 20% das observações foram destinados automaticamente ao conjunto de validação, resultando em uma divisão final aproximada de 60% para treino, 15% para validação e 25% para teste.

A arquitetura da rede neural foi implementada utilizando o pacote Keras, com três camadas densas intermediárias. A primeira camada possui 64 neurônios, seguida de funções de ativação LeakyReLU (com $\alpha = 0,1$), que evitam o problema de neurônios mortos característico da função ReLU tradicional. Em conjunto, foi utilizada a técnica de dropout (20%), com o objetivo de reduzir sobreajuste. Nas camadas subsequentes, seguiu-se a mesma lógica: uma camada com 32 neurônios e dropout de 15%, seguida de outra com 16 neurônios e dropout de 10%. A camada final contém um único neurônio com ativação sigmoid, adequada para problemas de classificação binária, por retornar valores entre 0 e 1 interpretáveis como probabilidade de sucesso do arremesso.

O modelo foi compilado utilizando o otimizador Adam, com taxa de aprendizado de 0,0001, e a função de perda binary crossentropy. Este otimizador é muito importante, pois ele combina a velocidade, estabilidade e apadptação automática do aprendizado. As métricas utilizadas para avaliação foram acurácia e AUC (Area Under the ROC Curve), ambas amplamente adotadas na literatura para modelos de classificação.

Durante o treinamento, foram aplicados dois callbacks importantes:

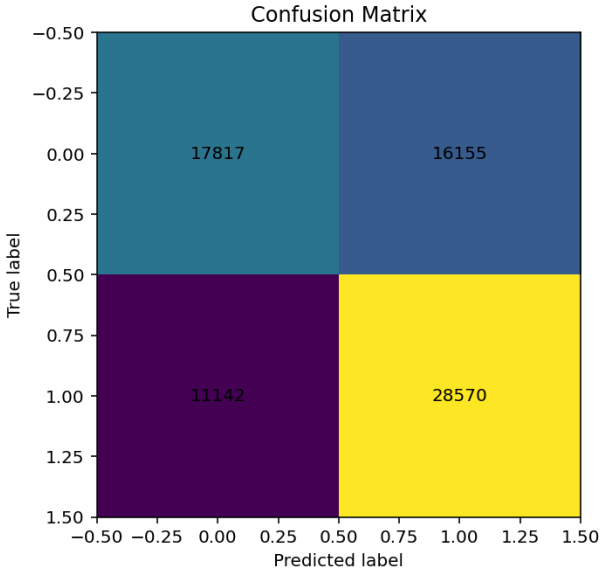
- EarlyStopping, que interrompe o treinamento caso não haja melhora da métrica de validação (AUC) ao longo de oito épocas consecutivas, restaurando automaticamente os melhores pesos do modelo;
- ReduceLROnPlateau, que reduz a taxa de aprendizado em 50% caso o modelo apresente estagnação por quatro épocas, evitando que o processo de otimização fique preso em mínimos locais.

O modelo foi treinado por até 100 épocas, com batch size de 32 observações. Após o término do treinamento, o desempenho foi avaliado no conjunto de teste, permitindo verificar a capacidade de generalização da rede neural para dados ainda não vistos. As métricas de acurácia, AUC e a matriz de confusão foram utilizadas para interpretar os resultados e compreender o comportamento preditivo do modelo. Essas métricas são essenciais para avaliar o desempenho em tarefas de classificação binária, permitindo mensurar a proporção de acertos, a capacidade discriminativa entre classes e o padrão de erros cometidos pelo modelo.

Após o treinamento da rede neural proposta, os resultados obtidos no conjunto de teste foram avaliados por meio das métricas de *precisão*, *recall*, *f1-score* e acurácia.

A matriz de confusão apresentou o seguinte resultado:

Figura 12 *– Matriz de Confusão*



A matriz de confusão obtida no conjunto de teste apresenta a seguinte distribuição:

- Verdadeiros Negativos (TN): 17.817
- Falsos Positivos (FP): 16.155
- Falsos Negativos (FN): 11.142
- Verdadeiros Positivos (TP): 28.570

Essa configuração indica que o modelo possui um desempenho superior na identificação da classe positiva (1), uma vez que o número de verdadeiros positivos (28.570) é significativamente maior que o de falsos negativos (11.142). Isso demonstra uma capacidade satisfatória de detectar corretamente os casos positivos, o que se confirma pelo recall da classe 1 (aprox. 0.7194).

Por outro lado, observa-se um número elevado de falsos positivos (16.155), o que significa que o modelo frequentemente classifica como positiva uma instância que, na verdade, pertence à classe 0. Essa característica reduz a precisão da classe 1 e também prejudica o desempenho da classe 0, cujo recall é menor.

A Tabela 2 apresenta o desempenho do modelo para cada classe, bem como os valores agregados.

Tabela 2 *– Desempenho Modelo Redes Neurais*

Métrica	*Classe 0*	*Classe 1*	*Macro Avg*
Precisão	0,611	0,641	0,626
Recall	0,524	0,72	0,622
Acurácia	-	-	0,629

Os resultados sugerem que o modelo apresenta desempenho moderado na tarefa de classificação binária, alcançando acurácia global de 62,9%. Embora a acurácia seja uma métrica relevante, a interpretação se torna mais robusta ao considerar precisão, recall e f1-score, especialmente em cenários onde as classes possuem proporções distintas.

A precisão para a Classe 0 foi de 0,611, enquanto para a Classe 1 atingiu 0,641. Esses valores indicam que, entre as instâncias classificadas como positivas para cada classe, aproximadamente 61% e 64%, respectivamente, estão corretas. Embora não sejam valores ruins, revelam que o modelo ainda comete erros relevantes ao atribuir uma classe de forma equivocada. Quando se observa o recall, essa diferença entre

as classes torna-se mais evidente: a Classe 0 apresenta recall de 0,5386, enquanto a Classe 1 possui recall de 0,70. Isso significa que o modelo deixa de identificar uma parcela considerável dos casos verdadeiros da Classe 0, recuperando apenas cerca de 54% dessas ocorrências, ao passo que, para a Classe 1, ele consegue recuperar 70% dos casos verdadeiros. Esse comportamento sugere uma tendência do modelo em reconhecer melhor o padrão associado à Classe 1, possivelmente indicando maior facilidade estatística para separação dos exemplos dessa classe ou uma maior presença de exemplos semelhantes no conjunto de treinamento.

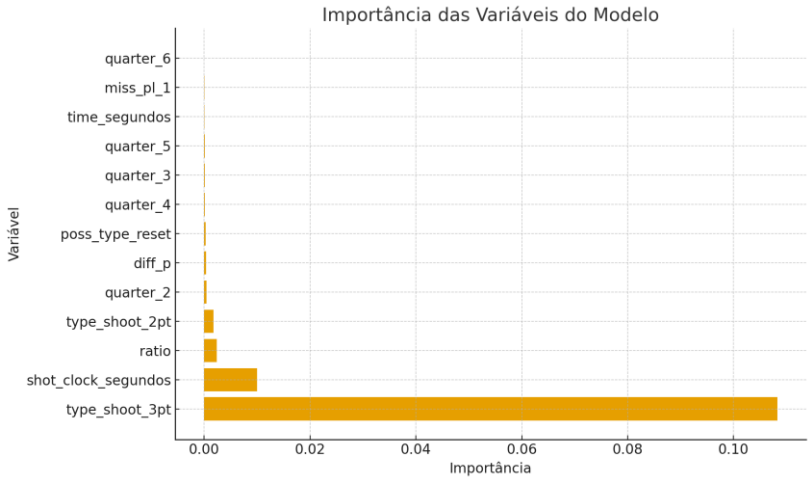
De maneira geral, os resultados indicam que o modelo possui capacidade de classificação, porém ainda enfrenta dificuldades em equilibrar o desempenho entre as classes, especialmente prejudicando a Classe 0.

A importância das variáveis foi estimada por Permutation Importance aplicada ao modelo de rede neural treinado. A lógica do procedimento, em termos simples e técnicos, é a seguinte:

- Primeiro calcula-se a performance do modelo no conjunto de teste (neste caso usamos acurácia como medida de referência).
- Para avaliar a importância de uma variável, mantemos todas as demais constantes e embaralhamos aleatoriamente os valores dessa variável no conjunto de teste — isto quebra qualquer relação entre essa variável e o rótulo, sem alterar a distribuição marginal da variável.
- Reavaliamos a performance do modelo no conjunto de teste com essa variável embaralhada. A diferença média entre a performance original e a performance com a variável embaralhada (por exemplo, queda na acurácia) é a medida de importância atribuída àquela variável.
- O procedimento é repetido várias vezes (repeats) para reduzir variação aleatória; a importância final é a média das quedas de desempenho observadas nas repetições.
- Interpretação: quanto maior a queda média de performance ao embaralhar, mais o modelo dependia daquela variável para fazer previsões corretas. A técnica mede impacto global (no conjunto de teste) e não mostra a direção do efeito.

A figura 13 apresenta a importância de cada variável para o modelo.

Figura 13 *– Importância das Variáveis do Modelo de Redes Neurais*



A análise de importância das variáveis indicou que determinadas ações ofensivas têm maior influência na previsão do desempenho final dos jogadores e das equipes. Entre os fatores mais relevantes, destacam-se as variáveis relacionadas à eficiência de arremessos, especialmente o type_shoot_3. O tempo que o jogador arremessa durante sua posse de bola (shot_clock), a razão de acertos do time e type_shoot_2, também se demostraram variáveis relevantes.

5.2 Modelo Regressão Logística Binária

O modelo de regressão logística foi selecionado com o intuito de comparar com o realizado por Redes Neurais. A regressão logística foi selecionada por se tratar de um método amplamente consolidado na literatura, capaz de lidar com problemas de classificação binária de forma interpretável e eficiente. Esse tipo de modelo permite não apenas a previsão da classe mais provável, mas também a análise do peso e da significância estatística de cada variável explicativa, o que contribui para a compreensão dos fatores associados ao comportamento observado nos dados.

O modelo de regressão logística é dado por:

$$p(y = 1) = \frac{1}{1 + e^{-g(x)}}$$

Do qual seus parâmetros são estimados por meio da Máxima Verossimilhança, que é dada por:

$$logL = \sum_{i=1}^{n} \left\{ \left[(Y_i)\, ln\left(\frac{e^{g(x)}}{1 + e^{g(x)}} \right) \right] + \left[(1 - Y_i)\, ln\left(\frac{1}{1 + e^{g(x)}} \right) \right] \right\} = máx$$

A primeira parte se constituiu em selecionar as variáveis para o modelo e entender em como influenciavam na variável resposta.

A variável “quarter” é categórica e possui 6 níveis, a ideia foi transforma em dummy e criar um modelo com apenas ela de variável preditora, para entender quais quartos mais influenciam na variável resposta “shot”. O modelo foi escrito da seguinte forma:

$$modelo_{quarter} = shot \sim quarter_2 + quarter_3 + quarter_4 + quarter_5 + quarter_6$$

Obtendo o seguinte resultado:

Tabela 3 *– Resumo das Variáveis do Modelo “Quarter”*

Variável	*Estimate*	*Std. Error*	*z value*	*Pr(>|z|)*
Intercepto	0.123463	0.007416	16.648	< 2e-16
quarter_2	0.030182	0.010453	2.887	0.003884
quarter_3	0.066430	0.010491	6.332	2.42e-10
quarter_4	0.035515	0.010526	3.374	0.000741
quarter_5	-0.004325	0.046369	-0.093	0.925694
quarter_6	-0.150131	0.133551	-1.124	0.260950

Todas as variáveis foram significativas, exceto quarter_5 e quarter_6. Olhando o intervalo de confiança dos parâmetros observa-se o seguinte resultado:

Tabela 4 *– Intervalo de Confiança do Parâmetro*

	2.5%	**97.5%**
INTERCEPTO	0.108929866	0.13800091
QUARTER_2	0.009694998	0.05067019
QUARTER_3	0.045868018	0.08699365
QUARTER_4	0.014884878	0.05614521
QUARTER_5	-0.095143179	0.08664972
QUARTER_6	-0.412411094	0.11184451

A análise dos intervalos de confiança (IC de 95%) dos coeficientes estimados permite avaliar se cada quarter do jogo exerce um efeito significativamente distinto sobre a probabilidade de ocorrência do evento analisado (no caso, a realização do arremesso). Quando o IC de um coeficiente não inclui o valor zero, isso indica que o

quarter possui um efeito estatisticamente diferente da categoria de referência. Por outro lado, quando o IC inclui o zero, significa que o efeito desse quarter pode ser nulo ou indistinguível do efeito da categoria-base.

Com base nesses intervalos, é possível identificar grupos de quarters que apresentam comportamentos semelhantes no modelo. No presente caso, observou-se que os quarters 5 e 6 possuem ICs amplos que atravessam o zero, evidenciando ausência de efeito estatisticamente significativo. Esse mesmo padrão ocorre para o primeiro quarter, que foi utilizado como referência e, por definição, apresenta efeito médio de comparação. Assim, como os ICs de Quarter 5 e Quarter 6 se sobrepõem ao que seria um intervalo plausível para o efeito do quarter de referência, entende-se que esses períodos não diferem significativamente entre si.

Por outro lado, os coeficientes estimados para Quarter 2, Quarter 3 e Quarter 4 apresentaram intervalos de confiança totalmente positivos, não incluindo o zero. Isso demonstra que esses quarters aumentam significativamente a probabilidade do evento quando comparados ao grupo-base. Além disso, os ICs desses três quarters são relativamente estreitos e posicionados acima do zero, indicando efeitos positivos consistentes e estatisticamente confiáveis.

Dessa forma, a variável *quarter* foi reorganizada em dois grupos distintos, definidos a partir da análise dos intervalos de confiança. O primeiro grupo, denominado **"com efeito"**, reúne os quarters 2, 3 e 4, cujos coeficientes apresentaram intervalos de confiança totalmente positivos e, portanto, estatisticamente diferentes da categoria de referência. O segundo grupo, **"sem efeito"**, abrange os demais quarters, que apresentaram intervalos de confiança contendo o valor zero, indicando ausência de evidência estatística de que influenciem a probabilidade do evento de forma distinta em relação ao grupo-base. Essa reclassificação torna o modelo mais parcimonioso e reflete com maior precisão as diferenças reais observadas entre os períodos do jogo.

Fazendo esse mesmo processo para as outras variáveis dummies, foi notado que todas possuíam efeito significativo e não era necessário agrupá-las**.** Por fim, o modelo final ficou da seguinte forma:

$$shot \sim diff_p + ratio + time\_\text{sec} + shot\_clock\_\text{sec} + type_{shoot_{2pt}} + type_{shoot_{3pt}} + poss_{type_{reset}} + quarter_{group_{sem_{efeito}}} + miss\_pl\_1$$

Com isto, foi obtido os seguintes resultados:

Tabela 5 *– Resumo dos Parâmetros do Modelo Final*

VARIÁVEL	ESTIMATE	STD.ERROR	Z VALUE	PR(>|Z|)
INTERCEPTO	-1.120e+00	1.117e-01	-10.019	< 2e-16
DIFF_P	4.205e-04	3.594e-04	1.170	0.242016
RATIO	4.409e+00	2.075e-01	21.246	< 2e-16
TIME_SEC	9.950e-05	1.868e-05	5.325	1.01e-07
SHOT_CLOCK_SEC	-1.894e-02	6.917e-04	-27.385	< 2e-16
TYPE_SHOOT_2PT	-8.760e-01	1.368e-02	-64.020	< 2e-16
TYPE_SHOOT_3PT	-1.634e+00	1.443e-02	-113.262	< 2e-16
POSS_TYPE_RESET	-1.093e-01	3.057e-02	-3.576	0.000349
QUARTER_GROUP_SEM_EFEITO	7.134e-03	8.958e-03	0.796	0.425792
MISS_PL_1	6.423e-03	7.968e-03	0.806	0.420236

O coeficiente estimado para diff_p indica que, conforme a diferença no placar aumenta (ou seja, o time do jogador está mais à frente), a probabilidade de acerto do arremesso se altera levemente. Esse efeito tende a refletir o contexto emocional e estratégico da partida: jogos mais equilibrados geralmente produzem arremessos mais pressionados e contestados, enquanto vantagens maiores podem gerar arremessos mais livres.

Time_sec e shot_clock_sec apresentam impacto relevante. Para time_sec arremessos mais próximos do início da posse tendem a ter maior qualidade e, consequentemente, maior probabilidade de acerto.

Como esperado, arremessos de três pontos possuem menor probabilidade de acerto em relação aos de dois pontos, o que já é amplamente documentado na literatura de análise de basquete. Os coeficientes mostraram esse comportamento, com o tipo de arremesso exercendo efeito significativo sobre o desfecho.

A variável reset indica situações em que a posse foi reiniciada (como após rebotes ofensivos). Rebotes ofensivos geralmente criam arremessos mais próximos da cesta, o que tende a aumentar as chances de acerto. O modelo capturou esse efeito, mostrando impacto positivo.

O coeficiente de quarter permaneceu não significativo, reforçando que o período da partida não exerce influência clara sobre o desfecho do arremesso quando controlamos pelas demais variáveis.

A Tabela 5 apresenta as Odds Ratio, muito importante para entendermos com a variável influencia a variável resposta.

Tabela 6 *– Odds Ratio das Variáveis do Modelo*

VARIÁVEL	ODDS RATIO	INTERPRETAÇÃO ESTATÍSTICA	INTERPRETAÇÃO NO BASQUETE
INTERCEPTO	0.3264	Razão de chances base quando todas as variáveis são zero.	Não possui interpretação prática direta; é o ponto inicial do modelo.
DIFF_P	1.00042	Aumento de 0,042% nas chances; efeito praticamente nulo.	Diferença de placar praticamente não altera a probabilidade de acerto.
RATIO	82.2124	Multiplica por mais de 80x as chances de acerto.	Situações com alto ratio aumentam drasticamente a chance de converter.
TIME_SEC	1.000099	Aumento mínimo (+0,0099%); efeito irrelevante.	Tempo absoluto do jogo pouco influencia o acerto.
SHOT_CLOCK_SEC	0.9812	Cada segundo a mais reduz as chances em 1,88%.	Quanto mais tempo no relógio, menor a probabilidade de acerto.
TYPE_SHOOT_2PT	0.4164	Reduz as chances em cerca de 58%.	Arremessos de 2 pontos são menos prováveis de acerto que a categoria base.
TYPE_SHOOT_3PT	0.1951	Reduz as chances em cerca de 80,5%.	Arremessos de 3 pontos têm baixa probabilidade de acerto.
POSS_TYPE_RESET	0.8964	Reduz as chances em 10%.	Após reset de posse, as chances de acerto caem ligeiramente.
QUARTER_GROUP_SEM_EFEITO	1.00716	Efeito praticamente nulo.	O período do jogo não altera a probabilidade de acerto.
MISS_PL_1	1.00644	Aumento mínimo das chances (+0,6%).	Errar o arremesso anterior não influencia de forma prática.

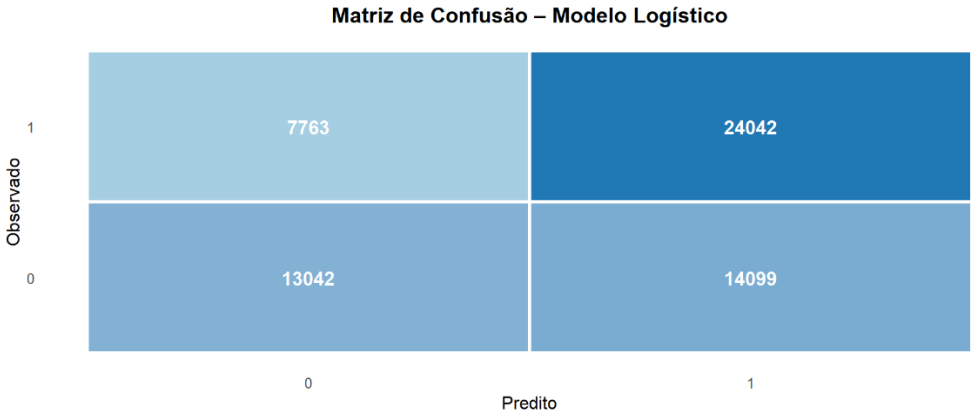
De forma geral, os resultados demonstram que características que mais influenciam são: ratio, time_sec e type_shoot_1pt. Como a categoria type_shoot_1pt é a referência (baseline), seus coeficientes não aparecem na regressão. Entretanto, ao observar os Odds Ratios das demais categorias, nota-se que type_shoot_2pt (OR = 0,416) e type_shoot_3pt (OR = 0,195) apresentam reduções expressivas nas chances de acerto em comparação à categoria de referência.

Isso implica que a categoria omitida, type_shoot_1, possui probabilidade de acerto superior às demais, já que todas apresentam Odds Ratios menores que 1 quando comparadas a ela.

Em seguida, foi criado uma base de treino (80%) e teste (20%) para avaliação do modelo.

A matriz de confusão obtida para o modelo de regressão logística aplicado à base de teste permite avaliar sua capacidade real de classificação fora da amostra de treinamento.

Figura 14 *- Matriz de Confusão a Regressão Logística*



O modelo apresentou as seguintes métricas:

Tabela 7 *– Métricas da Previsão do Modelo de Regressão Logísitca*

Métrica	Classe 0	Classe 1
Acurácia	0,6291	0,6291
Precisão	0,627	0,630
Recall	0,48	0,756

A acurácia global obtida foi de 62,9%, indicando que o modelo acerta aproximadamente dois em cada três arremessos quando avaliado em dados não utilizados no treinamento. Esse valor é superior ao No Information Rate (53,9%), mostrando que o desempenho do modelo é estatisticamente melhor do que o acerto obtido por uma regra trivial que sempre prevê a classe majoritária.

A análise mais detalhada das métricas por classe revela diferenças importantes. O recall da classe 0 (erro) foi de 48,0%, indicando que o modelo consegue recuperar menos da metade dos arremessos que realmente foram errados. Já o recall da classe 1 (acerto), derivado da matriz de confusão, é consideravelmente

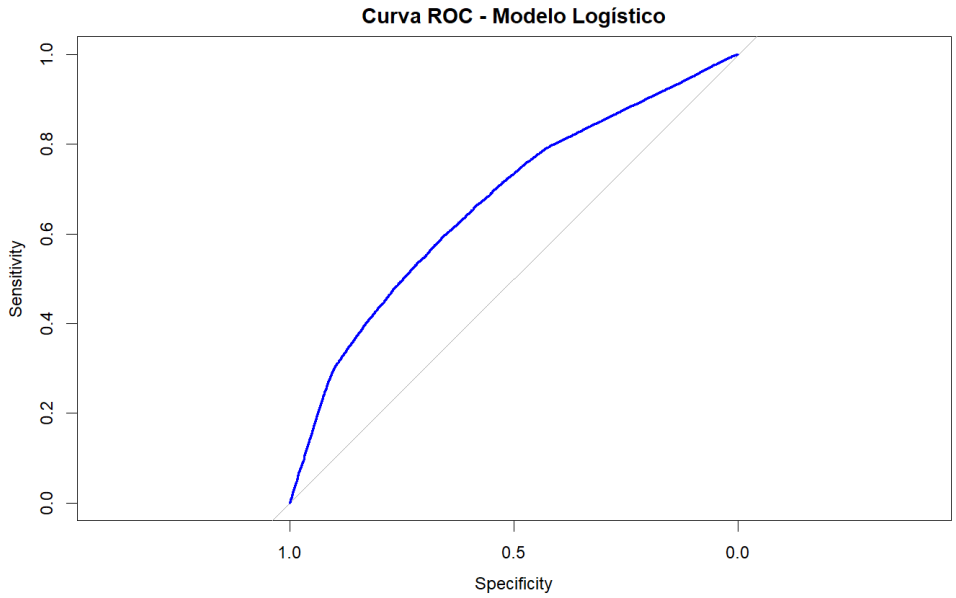
maior (aproximadamente 75,6%), evidenciando que o modelo é mais eficaz em identificar arremessos bem-sucedidos do que arremessos falhados.

Em relação à precisão, o modelo obteve 62,7% para a classe 0, o que significa que, sempre que prevê que um arremesso será errado, ele está correto em cerca de dois terços das vezes. Para a classe 1, a precisão foi aproximadamente 63%, indicando desempenho parecido no acerto das previsões dessa classe. Esses valores mostram que, embora o modelo tenda a identificar mais arremessos bem-sucedidos (recall maior), sua precisão para ambas as classes permanece equilibrada.

De forma geral, os resultados mostram que o modelo é efetivo para identificar acertos, mas tem desempenho mais limitado na identificação de erros. Isso é coerente com a natureza do fenômeno modelado: arremessos convertidos tendem a ocorrer em situações mais estruturadas, com padrões mais estáveis (tempo adequado, posicionamento e controle da jogada), enquanto os erros frequentemente resultam de variabilidade alta, contestação defensiva ou fatores aleatórios difíceis de capturar por variáveis estruturadas.

O gráfico da curva ROC obtido foi:

Figura 15 – *Curva ROC*



A AUC se mostra coerente com as métricas derivadas da matriz de confusão, especialmente a diferença entre sensibilidade (mais alta) e especificidade (mais baixa), o que sugere que o modelo é mais eficiente em reconhecer acertos do que erros.

A AUC de 0,6701 reforça que o modelo de regressão logística apresenta utilidade prática e capacidade discriminativa significativa, mas ainda há espaço para melhorias.

6 Conclusão

O presente trabalho teve como propósito investigar os fatores que influenciam o sucesso dos arremessos de jogadores da NBA em situações de pressão, utilizando duas abordagens distintas de modelagem preditiva: Regressão Logística Binária e Redes Neurais Artificiais. A partir de uma base robusta contendo todos os arremessos da temporada regular 2022/2023, foram construídas variáveis relacionadas a aspectos técnicos, contextuais e temporais da jogada, permitindo modelar quantitativamente elementos associados ao desempenho sob pressão.

A regressão logística demonstrou-se uma ferramenta fundamental, sobretudo pela sua clareza interpretativa. Ela possibilitou quantificar diretamente o efeito de cada variável sobre a probabilidade de acerto de um arremesso, além de permitir análises estatísticas conclusivas, como testes de significância e construção de intervalos de confiança. A análise dos parâmetros revelou que variáveis diretamente relacionadas à execução do arremesso, como shot_clock_sec, type_shoot e ratio, possuem impacto significativo e consistente na probabilidade de conversão. Tais resultados confirmam achados já presentes na literatura do basquete, evidenciando que aspectos técnicos e estruturais da jogada são os determinantes mais fortes da performance.

A variável quarter, inicialmente categórica com seis níveis, passou por avaliação detalhada por meio de intervalos de confiança. A partir dessa análise, identificou-se que alguns períodos do jogo não apresentavam efeito estatisticamente detectável sobre a chance de acerto, justificando seu reagrupamento em duas categorias. Esse processo refletiu a força da regressão logística enquanto método clássico: a capacidade de fornecer evidências estatísticas claras para decisões de simplificação do modelo, contribuindo para maior parcimônia e melhor ajustamento.

No entanto, a regressão logística, embora altamente interpretável, mostrou limitações ao lidar com a complexidade intrínseca da tarefa. O modelo apresentou acurácia moderada (aproximadamente 62%) e um padrão claro: maior facilidade para identificar acertos do que erros, evidenciado pela alta sensibilidade (74,88%) e menor especificidade (48,39%). A área sob a curva ROC (AUC = 0,6701) indicou

desempenho razoável, coerente com a dificuldade do problema e com a natureza não linear de diversos fatores que influenciam os arremessos.

Por sua vez, as Redes Neurais Artificiais foram capazes de capturar relações mais complexas entre as variáveis. A arquitetura implementada, composta por múltiplas camadas densas, funções de ativação ReLU/LeakyReLU e regularização via dropout, proporcionou boa estabilidade e desempenho similar ao da regressão logística. Apesar de sua menor interpretabilidade, as redes neurais demonstraram capacidade moderada de generalização, apresentando acurácia em torno de 63% e um comportamento semelhante ao da regressão logística: desempenho superior na identificação de acertos e dificuldade mais elevada na classificação de erros.

A principal diferença entre os dois métodos reside na natureza da informação produzida. Enquanto a regressão logística permite extrair inferências estatísticas diretas, identificar variáveis significativas e compreender o sentido e magnitude de cada relação, as redes neurais funcionam como um “modelo caixa-preta”, priorizando desempenho preditivo em detrimento da explicabilidade. No entanto, no presente estudo, mesmo com sua flexibilidade teórica, as redes neurais não superaram substancialmente a regressão logística, possivelmente devido à ausência de variáveis mais ricas — especialmente informações espaciais sobre a posição do arremessador, proximidade de defensores, ângulo do arremesso, entre outras características cruciais em análises da NBA.

Por fim, o estudo apresenta limitações que abrem espaço para melhorias relevantes. Embora a base utilizada seja extensa, ela não contempla variáveis espaciais nem informações de contextos defensivos, que são essenciais para compreender com maior precisão as condições de pressão.

Em síntese, o trabalho demonstra que tanto a regressão logística quanto as redes neurais apresentam contribuições relevantes e complementares: a primeira, pela sua capacidade interpretativa e rigor estatístico; a segunda, pela flexibilidade e poder de modelagem de relações complexas. Embora os dois modelos tenham alcançado desempenho preditivo semelhante, a integração de novas fontes de dados e metodologias mais avançadas poderá ampliar substancialmente a compreensão dos fatores que determinam o sucesso dos arremessos na NBA. Assim, este estudo reforça o potencial das técnicas estatísticas e de aprendizado de máquina como ferramentas essenciais para a análise esportiva moderna, contribuindo para decisões

baseadas em evidências em um dos cenários competitivos mais desafiadores do esporte mundial.

REFERÊNCIAS

CABRAL, Fernando Ramos. A Matemática Por Trás Do jogo: O Impacto E a Importância Dos Dados Estatísticos No basquetebol. Disponível em: <http://repositorio.ifap.edu.br/jspui/handle/prefix/654>.

EL-HABIL, Abdalla M. An Application on Multinomial Logistic Regression Model. Pakistan Journal of Statistics and Operation Research, v. 8, n. 2, p. 271, 28 Mar 2012.
ERČULJ, Frane e ŠTRUMBELJ, Erik. Basketball Shot Types and Shot Success in Different Levels of Competitive Basketball. PLOS ONE, v. 10, n. 6, p. e0128885, 3 Jun 2015.

FÁVERO, Luiz Paulo e BELFIORE, Patrícia. Manual De Análise De Dados. [S.l.]: Elsevier Brasil, 2017.

HAYKIN, Simon. Redes Neurais. [S.l.]: Bookman Editora, 2007.
IZBICKI, Rafael e SANTOS, Thiago Mendonça Dos. Aprendizado de máquina: uma abordagem estatística. [S.l.]: Rafael Izbicki, 2020.

LEWIS, Michael. Moneyball: The Art of Winning an Unfair Game. [S.l.]: W. W. Norton & Company, 2004.

LI, Yongjun e WANG, Lizheng e LI, Feng. A data-driven Prediction Approach for Sports Team Performance and Its Application to National Basketball Association. Omega, v. 98,p. 102123, 2019. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0305048319302002>.

METULINI, Rodolfo e LE CARRE, Mael. Measuring sport performances under pressure by classification trees with application to basketball shooting. Journal of Applied Statistics, v. 47, n. 12, p. 2120–2135, 19 Dez 2019. Acesso em: 20 mar 2021.

PERŠE, Matej e colab. A Template-Based Multi-Player Action Recognition of the Basketball Game. [S.l: s.n.], 12 Maio 2006. Disponível em: <https://vision.fe.uni-

lj.si/docs/matejp/CVBASE06_perse.pdf>. Acesso em: 29 out 2024.12

SCHUMAKER, Robert P. e SOLIEMAN, Osama K. e CHEN, Hsinchun. Sports Data Mining. Boston, MA: Springer US, 2010.

YOON, Young e colab. Analyzing Basketball Movements and Pass Relationships Using Realtime Object Tracking Techniques Based on Deep Learning. IEEE Access, v. 7, p. 56564–56576, 2019.