Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Departamento de Estatística

Predição do desempenho de jogadores da *National Basketball Association*

Alberto Torres Bueno Júnior

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos

Centro de Ciências Exatas e de Tecnologia

Departamento de Estatística

Predição do desempenho de jogadores da *National Basketball Association*

Alberto Torres Bueno Júnior

Orientador: Prof. Dr. Márcio Alves Diniz

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística da Universidade Federal de São Carlos -DEs - UFSCar, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos, SP

Setembro de 2024

Federal University of São Carlos

Exact and Technology Sciences Center

Department of Statistics

Predicting *National Basketball Association* players performance

Alberto Torres Bueno Júnior

Advisor: Prof. Dr. Marcio Alves Diniz

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos

November 2024

Alberto Torres Bueno Júnior

Predição do desempenho de jogadores da *National Basketball Association*

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Alberto Torres Bueno Júnior e aprovado pela banca examinadora.

Aprovado em 06 de setembro de 2024

Banca Examinadora:

- Prof. Dr. Márcio Alves Diniz (Orientador)
- Prof. Dra. Maria Silvia de Assis Moura
- Prof. Dr. Anderson Luiz Ara Souza

A todos que me apoiaram no caminho.

Resumo

O trabalho tem como objetivo buscar uma alternativa para predição do desempenho técnico e tático em quadra de atletas da *National Basketball Association*, (NBA), a liga profissional de basquete masculino do Estados Unidos. O foco será no desempenho individual dos jogadores, apresentando uma visão alternativa às predições mais comuns, que possuem perspectiva concentrada no desempenho dos times.

Além disso, serão apresentados e analisados índices de desempenho individual já existentes e popularizados, que sejam comparáveis com o Modelo Linear Generalizado e a Árvore de Regressão ajustados com o objetivo de predição.

Após o ajuste dos modelos propostos, a performance dos mesmos será comparada para períodos distintos, possibilitando conclusões mais variadas.

Palavras-chave: *predição, esportiva, basquete, jogadores, NBA*.

Abstract

The objective of this study is to explore an alternative for predicting the technical and tactical performance of athletes in the National Basketball Association (NBA). The focus will be on the individual performance of players, providing an alternative perspective to more common predictions, which are centered on team's performance.

Furthermore, existing and popularized individual performance indices will be presented and analyzed, allowing comparison with the Generalized Linear Model and Regression Tree models, which are adjusted with the objective of predicting the performances of players in new games.

After the proposed models are fitted, each respective performance will be compared over different seasons, enabling more varied insights.

Keywords: *prediction, sports, basketball, players, NBA*.

Lista de Figuras

1.1 Quadra de basquete com medidas. Fonte: (Alves, 2011) 22

1.2 Posições dos jogadores. Fonte: (Brasil, 2016). 23

2.1 Exemplo de uma Árvore de Regressão simples. Fonte: (IZBICKI e SANTOS, 2020). 50

2.2 Exemplo de uma Árvore de Regressão sobre preços de seguro de saúde nos EUA. Fonte: Choi (2017) 51

2.3 Comparação entre os índices de Informação e Gini. (THERNEAU e ATKINSON, 2019) 54

3.1 Histogramas das principais métricas avançadas. 63

3.2 Densidades das métricas de conversão de arremessos. 64

3.3 Distribuição do BPM dos jogadores em relação ao mando. 66

3.4 Distribuição do `Game Score` dos jogadores em relação ao mando. 67

3.5 Distribuição do `PER` dos jogadores em relação ao mando. 67

3.6 Distribuição do Índice de Desempenho Proposto. 69

3.7 Boxplot do Índice de Desempenho Proposto. 70

3.8 Dispersão dos resíduos do modelo. 74

3.9 Histograma e QQ-Plot dos resíduos do MLG. 75

3.10 Nós e folhas do modelo de árvore de decisão. 77

Lista de Tabelas

2.1 Descrição de todas as variáveis presentes no conjunto de dados. 38

3.1 Estatísticas resumo de métricas avançadas. 61

3.2 Estatísticas resumo de métricas ofensivas. 62

3.3 Estatísticas sumárias de métricas defensivas. 63

3.4 Médias das métricas avançadas, por posição. 65

3.5 Tabela de contagem para as posições. 66

3.6 Tabela de contagem para o mando de jogo. 66

3.7 Tabela de médias de algumas métricas, por temporada. 68

3.8 Medidas resumo da variável resposta. 68

3.9 Estimativas do modelo GLM por validação cruzada. 72

3.10 Métricas de desempenho do modelo. 73

3.11 Métricas de desempenho das predições do MLG na temporada 2022-2023. . 76

3.12 Métricas de desempenho das predições do MLG, distribuição Gama. 76

3.13 Métricas de desempenho das predições do MLG, distribuição Poisson. . . . 76

3.14 Métricas de desempenho das predições da Árvore de Regressão, conjunto de treinamento. 78

3.15 Métricas de desempenho das predições da Árvore de Regressão, para a temporada de 2022-2023. 79

Sumário

1 Introdução e Objetivo **19**

1.1 Conceitos e história do basquete 19

1.2 Contexto do estudo 25

2 Material e Métodos **35**

2.1 Conjunto de dados 35

2.2 Metodologias 43

2.2.1 Modelos Lineares Generalizados 43

2.2.2 Árvores de Particionamento Recursivo 49

2.2.3 Criação da Árvore Complexa 53

3 Resultados **61**

3.1 Análise Descritiva e Exploratória 61

3.2 Modelagens e Predições 70

4 Comentários Finais **81**

Referências Bibliográficas **83**

A Códigos **87**

Capítulo 1

Introdução e Objetivo

1.1 Conceitos e história do basquete

Com a modernização do esporte, consequência da vasta ampliação do uso de tecnologias para a medição de desempenho dos atletas, muitos novos fatores começaram a influenciar o dia-a-dia das equipes. Isso permitiu o registro quantitativo de diversos atributos e, graças ao uso da modelagem estatística, essas informações coletadas previamente permitem fazer análises mais abrangentes e completas sobre o desempenho dos jogadores em uma série de jogos futuros. Dentre todos os esportes que adotaram essa nova dinâmica analítica, o basquete figura entre os mais notórios.

Neste contexto, a *National Basketball Association* (NBA), atualmente considerada a principal liga masculina de basquete de todo o mundo, vem ampliando seus esforços para aumentar sua qualidade e competitividade e o uso da análise estatística sobre suas partidas é parte desse esforço. Com isso, o interesse de fãs, profissionais e pesquisadores vem proporcionalmente aumentado para desenvolver novas técnicas ou análises que possam causar um resultado impactante.

A NBA, seu nome atual, foi fundada em 6 de junho de 1946 sobre a alcunha de

Basketball Association of America (BAA). Já em 1949 ocorreu a fusão com outra liga de basquete, a *National Basketball League* (NBL) que culminou na NBA. No seu período inicial, existiam apenas 17 equipes e a primeira partida, ainda pela BAA ocorreu em 1º de novembro de 1946 entre o *Toronto Huskies* e o *New York Knickerbockers*, na cidade de Toronto-CAN e terminou com vitória da equipe de New York (Reference, 2024b). Curiosamente, ambas equipes ainda existem nos dias atuais, com a equipe de Toronto sendo rebatizada como *Raptors* e adotando toda uma nova identidade visual, enquanto que os *Knicks* passaram por apenas remodelações gráficas no seu logo, mantendo o nome e identidade por todos esses anos.

Ao longo das décadas o esporte, sendo dinâmico e mutável, passou por algumas "eras" dentro da liga. Assim são conhecidas as dinastias: grandes equipes, jogadores e técnicos que foram responsáveis por enormes alterações na dinâmica do jogo. Sendo assim, faz-se essencial contextualizar brevemente todo esse período histórico que têm influência inegável nas partidas jogadas atualmente na NBA e no restante do mundo.

A primeira grande era marcante foi protagonizada pelos *Boston Celtics*, que nas décadas de 50 e 60 dominaram a liga e venceram 11 títulos em um período de 13 anos. Entre 1959 e 1966 todos os oito títulos em disputa foram vencidos pela equipe, um recorde histórico e que possivelmente nunca será quebrado (WCBV5, 2024). O lendário treinador Red Auerbach comandava grandes estrelas do esporte, com os principais sendo Bill Russel, Bob Cousy e John Havlicek, todos imortalizados no *Hall of Fame* da NBA (Memorial, 2024).

Depois dessa dinastia dos Celtics de Russell e cia, foi a vez de Larry Bird, outro *Hall of Famer* e um dos maiores jogadores da história, trazer uma nova era de domínio para a cidade de Boston. Porém, na gênese da maior rivalidade entre equipes da NBA, o *Los Angeles Lakers* liderado por Magic Johnson buscou rivalizar pela disputas dos títulos. A

liga se beneficiou incrivelmente por essa nova rivalidade entre as superequipes, de grandes cidades (além de tudo), com os Celtics vencendo em 1975-1976, 1980-1981, 1983-1984 (final contra os Lakers) e 1985-1986. Já os Lakers, venceram em 1979-1980, 1981-1982, 1984-1985 (finais contra os Celtics), 1986-1987 (finais contra os Celtics) e 1987-1988.

Entrando agora na época mais icônica, por conta da presença ilustre do maior jogador da história, a dinastia passa para Chicago, a comando do *Bulls* de **Michael Jordan**. A equipe venceu seis campeonatos na década de 90, nunca tendo vencido antes e também não conquistou nenhum título após esse domínio. Foram dois *three peats*, que é o nome para tricampeonatos que ocorrem de forma consecutiva, de 1991 até 1993 e de 1996 até 1998. Além de toda a mudança que Jordan causou dentro da quadra, a popularidade que ele trouxe para a liga e a inspiração que foi e ainda é para jovens atletas são contribuições imensuráveis.

A partir dos anos 2000, a liga se aproveitou de toda essa renovada popularidade para expandir sua influência e atrair talentos internacionais para atuar por suas equipes. Alguns desses talentos internacionais tiveram carreiras com marcas expressivas, como Dirk Nowitzki, Manu Ginóbili, Tony Parker, Yao Ming e Pau Gasol. Destes, apenas Ming não ganhou um título, embora tenha tido uma carreira longeva e com grande importância para a expansão no mercado chinês.

Mantendo essa tendência, atualmente temos os últimos cinco títulos de *Most Valuable Player* (MVP) sendo entregues a jogadores nascidos fora dos EUA, uma sequência inédita e que pode se estender ainda mais nos próximos anos. Como contexto, o MVP é um título dado ao jogador que foi mais "valioso", ou seja, que teve o melhor desempenho entre todos durante a temporada regular, o principal craque.

Devido ao aumento da diversidade de nacionalidades nos principais atletas da NBA, os últimos anos tiveram uma mudança gradual no estilo de jogo, mas que veio para ser a

tendência por muito tempo: o aumento dos arremessos de três pontos.

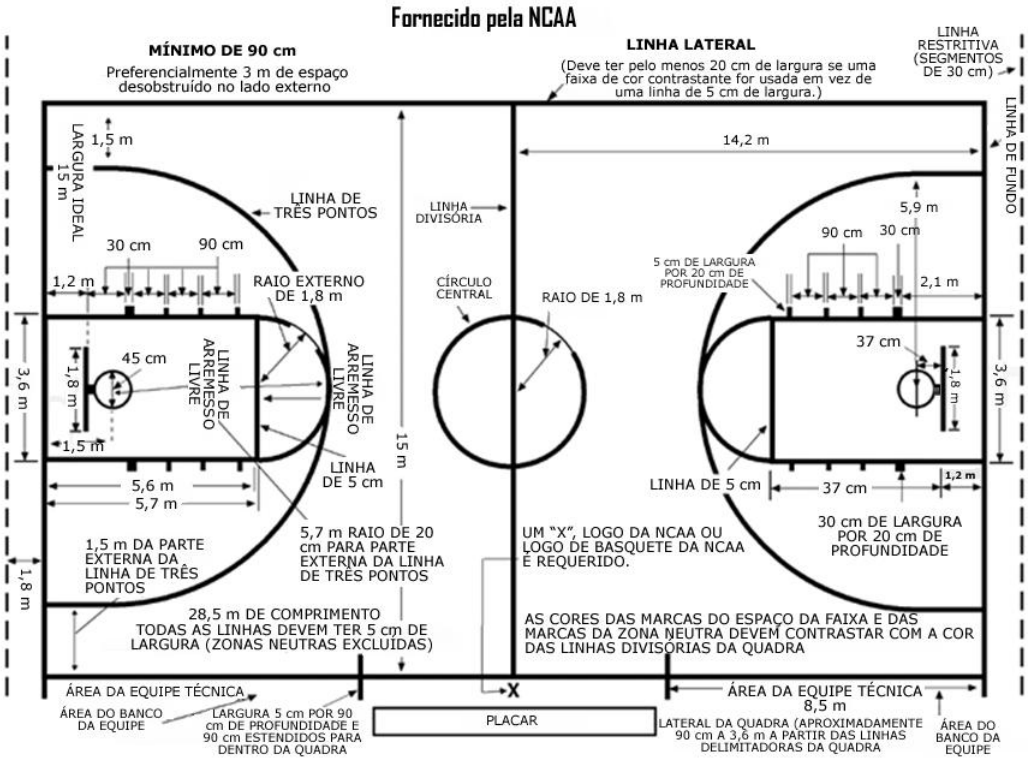

Figura 1.1: Quadra de basquete com medidas. Fonte: (Alves, 2011)

.

A regra que alterou a pontuação permitindo os arremessos de três pontos foi implementada em 1979. Como visto na Figura 1.1 a famosa "linha de 3" é um semicírculo que envolve o garrafão e a cesta/tabela em ambos os lados da quadra. Sendo assim, o time com posse de bola que acertar um arremesso onde o jogador esteja com ambos os pés para trás da linha do semicírculo, marcará três pontos como consequência. Já arremessos onde o jogador esteja dentro do semicírculo (sua linha perimetral inclusa), a equipe marcará dois pontos. Em situações de faltas de jogo ou faltas técnicas, um jogador arremessará da linha de arremesso livre, quando cada arremesso destes convertidos resulta em um ponto. São essas todas as possíveis pontuações em um jogo de basquete.

Mesmo com a mudança regulamentar ocorrendo a partir da temporada de 1979, a verdadeira revolução no uso frequente do arremesso de três só veio com Steph Curry e o *Golden State Warriors*. Curry, o inconteste melhor arremessador da história (NBA,

2022), criou uma carreira baseada na bola tripla e acumula sucessos, recordes e títulos, forçando com que as demais equipes ajustassem seu estilo de jogo.

Outra importante mudança recente é em relação a posição e estilo de jogo dos atletas.

Figura 1.2: Posições dos jogadores. Fonte: (Brasil, 2016).

Embora, devido ao aumento dos arremessos do perímetro (as bolas de três), o jogo tenha se tornado menos posicional, ainda existem algumas delimitações que diferenciam os estilos de jogadores que uma equipe precisa se reunir para ter um desempenho mínimo em diversas áreas. Sendo assim, uma divisão das posições dos cinco jogadores em quadra é:

- **1-Armador (*Point Guard*):** atleta que "arma" o time, com a principal função de ser o maestro do ataque, com boas habilidades de drible, controle de bola, infiltração, passe e arremesso. Normalmente, são os atletas que mais ficam com a bola em mãos e a posição favorita dos atletas mais baixos. Exemplos: Steph Curry e Steve Nash;

- **2-Ala-armador (*Shooting Guard*):** características semelhantes ao do armador, porém com mais enfoque na parte de arremesso. Na maioria dos times, é a posição com o melhor arremessador do elenco, principalmente relacionado as bolas de três, tendo também um importante papel defensivo, visto que precisa marcar o principal

arremessador do adversário. Exemplos: Klay Thompson e Michael Jordan;

- **3-Ala (*Small Forward*):** em geral, o principal pontuador de uma equipe. Posição de grande atleticismo, com jogadores versáteis de grande agilidade, altura e controle de bola, além de terem um jogo bem completo, dividindo as ações entre arremessos a distância e bandejas. Exemplos: LeBron James e Julius Erving;

- **4-Ala-pivô (*Power Forward*):** jogadores de grande poderio físico e atlético, com grande enfoque em atuar numa posição próxima a cesta, porém com a habilidade de "esticar" a quadra. Defensivamente, são sempre requisitados para defenderem atletas de características diferentes, defendendo o garrafão e a meia distância. Exemplos: Julius Randle e Tim Duncan;

- **5-Pivô (*Center*):** normalmente o jogador mais alto do time, com a principal função de marcar pontos próximo da cesta, além de pegar rebotes defensivos e ofensivos. Na defesa, são os responsáveis por proteger a área pintada, o garrafão de defesa. Exemplos: Nicola Jokic e Shaquille O'Neal.

Com a grande variação de táticas, enfoque no arremesso de três e a melhora de condicionamento e atleticismo dos atletas, tais posições não são regras bem definidas para o estilo de jogo de uma equipe. Muitos jogadores possuem características híbridas, jogando em posições *combo*, onde basicamente atuam de maneira diferente de acordo com o quinteto em quadra, o momento do jogo, dentre outros fatores. Ademais, alguns focam em se especializar, como os alas-armadores *"3 and D"*, especialistas em arremesso do perímetro e defesa.

Além disso, a NBA possui um conjunto de regras específico, que difere das utilizadas em competições internacionais, que seguem o regramento da *International Basketball Federation* (FIBA). O jogo é cronometrado, com quatro períodos de 12 minutos cada, sendo

cada posse de bola começando com um máximo de 24 segundos, sendo que subsequentes posses de bola no mesmo ataque retomam com 14 segundos. Por exemplo, se um arremesso não é convertido e um rebote ofensivo acontece, a posse de bola continua com o time que errou o arremesso, agora com 14 segundos para concluir a jogada.

Levando em consideração a mudança no estilo de jogo e procurando ampliar seus mercados, a NBA continua nos seus esforços de expansão, aproveitando a revigorada popularidade mundial dos seus *superstars*, com a possibilidade de expansão do total de equipes e jogos em território fora dos EUA e Canadá.

1.2 Contexto do estudo

Analisando os acontecimentos que impactam cada jogo, diversos fatores podem influenciar direta e indiretamente em seu resultado, como: lesões, onde o jogo é disputado, quantos jogos a equipe já realizou naquela temporada, qual foi o período de descanso antes do jogo, entre outros. Conseguir mensurar a relevância destes e de todos os outros fatores pode fazer uma grande diferença no resultado final da partida ou até mesmo alterar todo o rumo da temporada de uma equipe. Sendo assim, temos um ambiente muito dinâmico, complexo e valioso, combinação perfeita para despertar o interesse de inovação de técnicas analíticas.

Para entender melhor como vamos propor nossa análise, é extremamente valioso compreender o peculiar funcionamento da NBA, que difere muito de outras competições, também famosas mundialmente, como ligas de futebol. Fundamentalmente, seu ano-temporada (início em outubro de um ano e final em junho do ano seguinte, por exemplo: início em outubro de 2022 e final em junho de 2023, também chamada de temporada 22/23) é dividido em duas partes: a temporada regular e a pós-temporada, conhecida

como fase dos *playoffs*, ou eliminatória.

Antes do início de uma nova temporada, as franquias passam por um período de pré-temporada. Veja, aqui diz-se "franquia" por conta da liga funcionar exatamente dessa maneira. A liga é uma entidade esportiva que se divide em 32 partes, que são as franquias/times, cada uma controlada por um indivíduo ou grupo de indivíduos. Sendo assim, cada franquia é dona de 1/32 da liga, tendo voto em momentos cruciais de discussão sobre mudanças futuras.

Desse modo, cada franquia/equipe adota uma identidade e escolhe um mercado onde vai se estabelecer. Porém, a cidade de escolha não é condição permanente, visto que algumas equipes já se mudaram de cidades (e até de identidade, como o antigo *Seattle Supersonics*) de acordo com interesses financeiros. Embora dono de uma mesma fatia da liga, equipes que se estabelecem em cidades menores acabam sofrendo financeiramente devido à capacidade de outras equipes conseguirem captar mais recursos e atrair o desejo de jogadores, como em Nova York e Los Angeles, por exemplo.

Assim, alguns mecanismos foram implementados com o intuito de nivelar a competitividade e oferecer oportunidades menos desiguais para todas equipes. O principal destes é o *salary cap*, um teto/limite sobre o total da folha salarial paga a jogadores, impedindo a formação de super equipes em cidades maiores, mais atrativas. Embora interessante conceitualmente, esse limite não funciona perfeitamente e pode ser contornado mediante algumas manobras.

Durante a pré-temporada as equipes analisam o desempenho do ano-temporada anterior com o objetivo de buscar melhorias. Para isso, existem algumas principais opções.

- ***Draft*:** A liga realiza anualmente o *Draft*, um evento muito peculiar aos esportes praticados nos EUA. Nele, todas as equipes tem a oportunidade de selecionar jogadores que vem principalmente do esporte universitário, o principal caminho para

que jovens atletas alcancem o nível profissional na NBA. Com o intuito de balancear a competitividade e o nível dos elencos, a ordem das escolhas é inversamente proporcional ao desempenho do ano anterior, ou seja, as equipes com os piores desempenhos possuem a oportunidade de escolher primeiro, tendo a vantagem de escolher os jogadores que são consensualmente melhores promessas. Mais recentemente, alguns jogadores que atuam em outros continentes entraram para serem escolhidos no *Draft*, ampliando o horizonte e diversidade da liga. Para que esse mecanismo não causasse um desbalance claro durante a temporada, o famoso *tanking*, a liga adotou um sorteio desbalanceado para decidir a ordem de escolhas. Ou seja, o time com pior desempenho recebe uma probabilidade mais alta de ter a primeira escolha, com as probabilidades atribuídas diminuindo progressivamente para os demais times. Deste modo, não há um claro incentivo para que um time perca todos os jogos do ano, somente para garantir a escolha mais alta da próxima temporada;

- ***Free Agency*:** momento de contratações ou trocas, onde as equipes procuram jogadores que não possuem mais nenhum vínculo contratual com qualquer equipe ou oferecem ativos (jogadores, escolhas, dinheiro) para trocar por jogadores de outras equipes, que ainda tenham um contrato ativo;

- ***Young Players*:** com treinamentos, jogos amistosos, a *Summer League* e a *G-League* (uma liga secundária) as equipes também podem buscar desenvolver mais seus jovens talentos, que ainda não tiveram grande contribuição com o elenco principal, porém possuem grande potencial nos olhos da comissão técnica.

Além disso, neste período que antecede o início oficial da temporada as equipes fazem toda sua preparação para o início da nova competição, realizando treinamentos e ajustando aspectos que influenciam o desempenho dos atletas.

Já com o início de um ano-temporada, temos a primeira fase, conhecida como temporada regular. Esta é a principal e mais longa etapa da competição, onde cada equipe disputa 82 jogos com o objetivo de ter a melhor campanha possível, ou seja, proporção de vitórias e derrotas, para garantir uma classificação para a fase final.

Desses 82 jogos, 41 são disputados como mandantes e 41 como visitantes. Isto é muito importante, visto que o fator "mando de campo" é sempre muito valorizado, em alguns casos até em excesso, pelas equipes. Por conta da popularidade e senso comum ao redor de jogar no seu próprio ginásio e com vantagem de maior torcida, este fator sempre levanta questionamentos interessantes e demanda análises mais aprofundadas para chegar-se a uma conclusão baseada no histórico de cada equipe.

Nos dias de hoje, as 32 equipes que disputam a liga são divididas em duas conferências com 16 equipes cada uma: as conferência Leste e Oeste. Além disso, cada conferência é dividida em subconjuntos chamados de *divisões*, sendo:

- **Conferência Leste - Divisão do Atlântico:** Philadelphia 76ers, Boston Celtics, New York Knicks, Brooklyn Nets e Toronto Raptors;
- **Conferência Oeste - Divisão do Noroeste:** Denver Nuggets, Minnesota Timberwolves, Oklahoma City Thunder, Portland Trail Blazers e Utah Jazz;
- **Conferência Leste - Divisão do Central:** Chicago Bulls, Cleveland Cavaliers, Detroit Pistons, Indiana Pacers e Milwaukee Bucks;
- **Conferência Oeste - Divisão do Pacífico:** Golden State Warriors, Los Angeles Clippers, Los Angeles Lakers, Phoenix Suns e Sacramento Kings;
- **Conferência Leste - Divisão do Sudeste:** Charlotte Hornets, Atlanta Hawks, Miami Heat, Orlando Magic e Washington Wizards;

- **Conferência Oeste - Divisão do Sudoeste:** Dallas Mavericks, Houston Rockets, New Orleans Pelicans, Memphis Grizzlies e San Antonio Spurs.

Os nomes das divisões se referem à sua posição geográfica no país, com o intuito de diminuir o deslocamento dos times, visto que os "rivais" de divisão se enfrentam mais vezes entre si durante a temporada regular. Além disso, equipes de conferência diferentes possuem poucos duelos entre si, também aliviando a necessidade de deslocamentos muito grandes, o que vem por causar maior vantagem para a equipe mandante, visto que não despendeu energia em viagem para o referido jogo.

Embora as divisões separem os times em blocos dentro de cada conferência, estas não servem nenhum propósito classificatório para a pós-temporada, como acontece em outras ligas de esportes dos EUA, como a *National Football League* e a *Major League Baseball*, por exemplo. Atualmente, após a criação do torneio *play-in* na temporada 2019-2020 devido a pandemia de COVID-19, cada conferência classifica 10 times para os *playoffs*.

Destes, seis tem classificação direta, enquanto que os quatro demais participam do *play-in* em busca das duas vagas restantes. Essas vagas são definidas de maneira simples e objetiva, com um *ranking* das equipes baseados na sua proporção de vitórias/derrotas. Ou seja, o time com mais vitórias de cada conferência se classifica em 1º, aquele com o segundo maior saldo de vitórias em 2º e assim sucessivamente.

Ainda no decorrer da temporada regular, existe uma pausa no calendário, com o objetivo de oferecer um descanso aos atletas e também de dar espaço a um evento de grandes proporções midiáticas, o *All Star Game* ou *All Star Weekend* em tempos mais recentes. Nele, há uma mistura entre voto popular e de jornalistas que cobrem a liga para uma escolha dos jogadores com o melhor desempenho até aquele momento da temporada, para que duas equipes se formem e se enfrentem no jogo das estrelas.

Embora a ideia seja vistosa no papel e nas telas, o fim de semana deste evento re-

presenta simplesmente uma grande oportunidade para lucrar com propaganda, visto que os atletas escolhidos para a grande partida jogam de maneira despretensiosa, tentando poupar seus corpos de desgastes e possíveis lesões que possam impactar no restante da temporada. Sendo assim, este evento serve o maior propósito de criar discussões e debates entre aficionados, quadros de polêmicas em programas de televisão e alguns vídeos de jogadas plásticas em um *top 10*. Para quem gosta somente de esporte em alto nível e não só publicidade, vale mais focar em acompanhar outras partidas ao longo dos seis meses de temporada regular.

Com o fim das 82 rodadas que consistem na temporada regular, passamos para os *playoffs*. A partir da temporada 2019-2020 o sistema de torneio *play-in* passou a preceder os jogos dos *playoffs*. Neste mini-torneio, as equipes de cada conferência que ficaram entre as posições sete e dez da classificação se enfrentarão, visto que só oito equipes disputarão a primeira rodada dos *playoffs* (os seis melhores colocados se classificam de maneira automática).

Tendo os oito definidos de cada conferência, os *playoffs* são divididos em rodadas, no formato de confrontos eliminatórios: a primeira rodada com 8 equipes (quartas de final), semifinais de conferência com quatro equipes, finais de conferência e as finais da NBA, entre os dois campeões de suas respectivas conferências.

Em todas as rodadas, cada confronto ocorre em uma série de melhor de sete partidas, ou seja, a equipe que vencer quatro jogos ganha a rodada e avança. Aqui entra em cena mais um detalhe importante advindo do desempenho na temporada regular. As equipes carregam sua classificação geral, sendo isso um fator determinante na quantidade de partidas que serão disputados como mandante.

Por exemplo, quando o terceiro colocado geral enfrenta o sexto colocado geral, aquele com melhor classificação terá o direito de fazer a maioria das partidas da rodada como

mandante, sendo quatro das sete em total. O esquema genérico se dá por:

- **Jogo 1:** Mando do time com melhor classificação;
- **Jogo 2:** Mando do time com melhor classificação;
- **Jogo 3:** Mando do time com pior classificação;
- **Jogo 4:** Mando do time com pior classificação;
- **Jogo 5:** Mando do time com melhor classificação;
- **Jogo 6:** Mando do time com pior classificação;
- **Jogo 7:** Mando do time com melhor classificação;

Novamente, aqui temos o fator "mando de jogo" causando, em teoria, uma grande vantagem, visto que o time com melhor classificação joga mais vezes em casa, além de iniciar a série em seus mandos e ter em suas mãos a vantagem de abrir um placar geral de 2-0, caso vença os dois primeiros jogos. Como curiosidade, equipes que iniciaram uma série com duas vitórias, possuem um histórico de 425 séries vencidas, com apenas 34 viradas (of Baskteball, 2024).

Para as finais da NBA, o esquema de mando mantém-se o mesmo. Porém, em situações de confronto com equipes da mesma classificação numérica, 1º do Oeste contra 1º do Leste, o desempate se dará fazendo a proporção de vitória/derrotas da temporada regular. As finais são, sem dúvida, o momento mais importante da temporada. É produzido um grande conteúdo midiático em torno de todos os jogos, com a liga utilizando sua influência global para ampliar seus mercados e oferecer experiências únicas para fãs e novos telespectadores no grande embate entre as duas melhores equipes do mundo.

Considerando a grande importância e popularidade da liga, um tópico em alta é a predição esportiva, que utiliza uma combinação de métodos estatísticos, aprendizado de

máquina e análise de dados. Devido ao grande número de jogos e à vasta variedade de métricas que são mensuradas por partida, os modelos de predição têm acesso a uma quantidade significativa de informações, o que possibilita uma precisão mais precisa. Afinal, é sempre um tópico de grande interesse e debate quem se sagrará campeão daquele ano, ou até mesmo qual atleta levará o título de *MVP*.

Além das medidas tradicionais de cada jogo, como pontos, rebotes, assistência, roubos de bola, entre outros, a partir da temporada 2013-2014 temos o advento do *player tracking*. Com isto, a quantidade de informações disponíveis cresceu consideravelmente.

Outros fatores que podem ser subestimados por um olhar mais superficial são os extra quadra. Muito pode acontecer para alterar o destino de uma equipe numa temporada, como mudanças na diretoria, comissão técnica, punições, lesões extracampo e problemas de comportamento, fatores que também devem ser considerados, quando possível, na construção de modelos preditivos.

Entender todos esses conceitos e utilizar técnicas refinadas aumenta a possibilidade de identificar quais as principais variáveis que impactam no desempenho dos jogadores, que é definido por uma trama complexa dos fatores já elencados, assim como diversos outros. Por exemplo, podemos observar que jogadores com um alto índice de acerto nos arremessos de três pontos têm um impacto significativo no resultado final das partidas.

Entretanto, todo tipo de previsão possui suas limitações inerentes, além de muitas outras encontradas ao longo de sua aplicação. A situação não é diferente para a NBA, visto que o basquete é um esporte com um dinamismo único, que permite que momentos brilhantes e jogadas completamente inesperadas aconteçam, o que impacta consideravelmente no desempenho dos atletas. Fora isso, muitos momentos ao longo do dia podem afetar o desenrolar da partida, adicionando mais um quesito de imprevisibilidade.

Embora a missão não seja trivial, a vantagem e os *insights* que uma boa previsão

trazem supera todas as dificuldades do percurso. Uma boa previsão do desempenho dos atletas em quadra, nos contextos técnicos e táticos, pode mudar totalmente o rumo de uma equipe, visto que embora a temporada seja longa, algumas partidas chave funcionam como pêndulo em momentos cruciais: confrontos contra rivais na classificação, sequências de jogos fora de casa, partidas em dias consecutivos, etc. Além do aspecto mais pontual, o maior interesse na análise de dados permite uma melhora contínua do esporte, identificando pontos fortes e fracos que os atletas precisam focar.

Um dos principais objetivos deste trabalho é identificar os principais fatores que influenciam o desempenho dos jogadores nas partidas, o que envolve analisar as estatísticas individuais dos jogadores, além de considerar a interação entre os jogadores, táticas de jogo, estilo de jogo e fatores externos. Para essa análise, serão utilizadas técnicas de aprendizado de máquina, assim como outros tipos de modelagem mais usuais, como algoritmos de regressão. Além disso, análises exploratórias e descritivas que podem trazer já informações iniciais de grande valor para serem testadas no momento da modelagem e comparação com outros índices já utilizados comumente.

Um outro objetivo é entender como as conclusões obtidas com as predições podem impactar o futuro dos atletas e das equipes, auxiliando na criação de táticas personalizadas para cada time, focos de melhoria de alguns conceitos para os jogadores, assim como uma gestão mais assertiva de todos os lados em momentos de renovação contratual, trocas e as demais estratégias de gestão de elenco.

Embora seja um campo já explorado há algum tempo, existe uma grande limitação no mundo de predições esportivas, visto que a imprevisibilidade é algo inerente ao esporte, principalmente em esportes coletivos com um dinamismo tão exacerbado quanto o basquete. Sendo assim, abordaremos questões referentes ao que costumeiramente é atribuído como um “erro” da predição, visando um enfoque em tirar conclusões assertivas

ao analisar o resultado dos métodos utilizados, sem criar uma expectativa ao redor de uma infalibilidade.

Portanto, o objetivo específico deste trabalho é realizar uma investigação abrangente sobre a predição esportiva na NBA, utilizando técnicas estatísticas e de modelagem com o intuito de propor um ou mais modelos/índices capazes de trazer uma predição acurada sobre o desempenho dos jogadores. Serão utilizadas uma grande gama de variáveis e estatísticas, agregadas de fontes diversas, a fim de buscar desenvolver modelos preditivos precisos e informados, levando em consideração a forma de como as informações foram agregadas.

Capítulo 2

Material e Métodos

Para este capítulo, serão abordados os dados que foram utilizados, levando em consideração os objetivos já definidos. Além disso, é também fundamental apresentar a metodologia que será utilizada para a obtenção dos resultados, contextualizando assim as análises e conclusões.

2.1 Conjunto de dados

Devido ao aspecto dinâmico de um jogo de basquete, as estatísticas são compiladas de vários métodos diferentes, com o intuito de apresentar análises mais elaboradas e amplas para jogadores e equipes que atuam de maneiras únicas.

Sendo assim, durante o estudo serão utilizados dados da NBA com o escopo de *game logs*. Os *game logs* consistem nas estatísticas individuais de cada jogador em todos os jogos das respectivas temporadas. Com isso, é possível fazer um tipo de análise não tão comum: analisar focadamente no desempenho do jogador em cada jogo, buscando prever o desempenho, considerando como desempenho a performance técnica e tática em quadra, em todos os contextos possíveis de serem mensurados e analisados, nas próximas partidas.

Existem algumas outras possibilidades em agregar informações dos jogadores, como *totals* que representa uma soma de todas as estatísticas do atleta na temporada, *per 36 minutes* que representa uma extrapolação para simular uma atuação de 36 minutos do jogador e *per 100 possessions* que aglutina as estatísticas do jogador a cada 100 posses de bola em que ele atuou, todas estas versões possuem especificidades para fazer análises diferenciadas.

Os dados foram obtidos do *site basketball reference* (Reference, 2024a), que é uma das vertentes do *site sports reference* que é responsável por compilar informações das principais ligas esportivas dos EUA. Especialmente, as informações dos *game logs* foram obtidas do *site stathead* (Head, 2024), uma versão paga que disponibiliza uma plataforma muito mais flexível para que o usuário obtenha os dados com possibilidade de aplicar diversos filtros, ou, como no caso desse estudo, obter os dados brutos de todos os jogos sem qualquer tipo de filtro.

O período temporal inicialmente escolhido foi bastante amplo, com dados desde a temporada 1999/2000 até a temporada 2022/2023. Porém, para reduzir um pouco o volume massivo de informações, que resultam em processos computacionais demasiadamente demorados e, neste caso, atingindo o limite do poder computacional disponível, o conjunto foi reduzido para englobar um período de 15 temporadas, começando em 2008/2009. Embora um horizonte menor que o idealizado, a amplitude temporal é mais do que o suficiente para comportar informações de diferentes momentos e jogadores que passaram pela liga.

O conjunto de dados inicial consistia de 594.368 obervações, sendo que cada uma consiste no *log* com as estatísticas referentes a atuação na partida de um jogador. Ou seja, em uma partida em que 10 jogadores atuaram, o conjunto de dados possuirá 10 linhas referentes aquele jogo, uma para cada atleta que lá atuou. Para que a análise possa ser mais assertiva e traga conclusões mais conectadas com a dinâmica do esporte, foram

filtrados da base inicial as observações referentes a jogos onde o atleta atuou por menos de oito minutos na partida.

Embora, em alguns casos, alguém possa ter desempenhado um papel importantíssimo em uma partida mesmo com uma minutagem diminuta, a grande parcela desses casos são de jogadores que entraram em quadra em momentos muito específicos da partida, costumeiramente nos minutos finais, quando a diferença no placar já é irreversível e ambos os treinadores retiram seus principais jogadores de quadra, com o intuito de preservá-los fisicamente e evitar algum tipo de lesão. Sendo assim, após a aplicação do referido filtro, o conjunto a ser analisado consiste em 528.192 observações.

Adicionalmente, foi feita uma amostragem por jogador, buscando ainda reduzir o volume massivo de informações, porém ainda mantendo informações valiosas. Para cada equipe e temporada, foram selecionados os dois jogadores com a maior quantidade de minutos jogados naquele ano. Para homogeneizar a amostra, buscando evitar um viés de apenas os "melhores" jogadores, foram selecionados, aleatoriamente, outros jogadores da base, excluindo aqueles com a maior minutagem, já selecionados. Todo esse processo de amostragem, filtros, entre outros, pode ser visto no código no Apêndice A. Após isso, o conjunto de dados que será utilizado possui 273998 observações, referente a 661 jogadores.

Um grande número de variáveis foi escolhido, ampliando as possibilidades de análise e construção de modelo. Sua descrição detalhada pode ser vista na Tabela 2.1.

Tabela 2.1: Descrição de todas as variáveis presentes no conjunto de dados.

Variável	**Descrição**
Jogador	Nome do jogador
Pontos	Quantidade de pontos marcados na partida
Idade	Idade, em anos e dias, do jogador no data da partida
Time	Time do jogador
Oponente	Time oponente do jogador na partida
Titular	*Flag* que registra se o jogador foi titular na partida
Minutos	Quantidade de minutos jogados na partida
`FG`	Quantidade de arremessos de quadra convertidos
`FGA`	Quantidade de arremessos de quadra tentados
`FG%`	Porcentagem de arremessos de quadra convertidos
`2P`	Quantidade de arremessos de 2 convertidos
`2PA`	Quantidade de arremessos de 2 tentados
`2P%`	Porcentagem de arremessos de 2 convertidos
`3P`	Quantidade de arremessos de 3 convertidos
`3PA`	Quantidade de arremessos de 3 tentados
`3P%`	Porcentagem de arremessos de 3 convertidos
`FT`	Quantidade de lances livres convertidos
`FTA`	Quantidade de lances livres tentados
`FT%`	Porcentagem de lances livres convertidos
`TS%`	Eficiência geral de arremessos (em porcentagem)
`ORB`	Quantidade de rebotes ofensivos
`DRB`	Quantidade de rebotes defensivos
`TRB`	Quantidade de rebotes totais (ofensivos + defensivos)
`AST`	Quantidade de assistências
`STL`	Quantidade de roubos de bola
`BLK`	Quantidade de tocos/bloqueios
`TOV`	Quantidade de *turnovers* (perdas da posse)
`PF`	Faltas pessoais/individuais cometidas (*Personal Fouls*)
`Game Score`	Métrica avançada `Game Score`
`BPM`	Métrica avançada `Box Plus Minus`
`Plus Minus`	Métrica avançada `Plus Minus`
Posição	Posição do jogador
Temporada	Derivado da data, temporada referente da partida
Idade (anos)	Idade, somente em anos, do jogador
Mando	*Flag* que registra se a equipe do jogador foi mandante
Resultado	*Flag* que registra se a equipe do jogador venceu
Margem	Diferença de pontos no placar final da partida
`OT`	*Flag* que registra se a partida foi para a prorrogação
`PER`	Métrica avançada `PER`
`eFG`	Métrica avançada `eFG`
`TSA`	Métrica avançada `TSA`
`FTR`	Métrica avançada `FTR`
`TOV%`	Métrica avançada `TOV%`

Do conjunto obtido do site *stathead*, algumas das variáveis foram divididas e modifi-

cadas, com o objetivo de aproveitar ao máximo suas informações. A variável "Resultado Final" apresentava o placar final da partida, assim como uma marcação se o time do jogador em questão havia vencido o jogo, até com uma marcação especial em casos de prorrogação **OT** (*Over Time*). Com isso, foi possível criar as variáveis: "Resultado" que aponta vitória ou derrota, "Margem" que apresenta a diferença final entre a pontuação dos times no placar e "OT" que aponta se a partida acabou somente na prorrogação (*Over Time*), embora a ocorrência desse evento seja baixa na base. De forma similar, as variáveis "Temporada" e "Idade (anos)" foram construídas a partir de outras variáveis presentes no conjunto, trazendo uma informação mais sintetizada.

Algumas informações foram mantidas e construídas no formato categórico, seja por sua natureza ou com o intuito de utilização para agrupamento e/ou análises detalhadas para cada uma dessas categorias. Das colunas categóricas ou que foram categorizadas, temos

- Nome do jogador;
- Time do jogador;
- Time oponente;
- Titularidade;
- Posição;
- Temporada;
- Mando;
- Resultado.

De maneira adicional, “Mando” e “Resultado” também foram utilizadas no formato numérico, como *dummies* binárias, com o valor 1 indicando jogo em casa e vitória, respectivamente.

Além destas, a maior parte das variáveis incluídas no estudo são as mais comumente associadas a uma partida de basquete, como Pontos, `3P`, Assistências, entre outros. Com o grande aumento das análises esportivas e coleta de informações mais precisas, distintas e abrangentes, novas métricas e estatísticas, que ficaram conhecidas como avançadas, surgiram para enriquecer todos os tipos de estudos esportivos.

Levando isso em consideração, pode ser bem vantajoso mais dessas métricas e estatísticas para um estudo preditivo como este. Porém, ancorado no princípio da parcimônia, a escolha foi feita com base nas análises mais comuns sobre o esporte, também levando em consideração as estatísticas que possuem maior praticidade para serem calculadas no âmbito jogo-a-jogo. Como o conjunto de possíveis variáveis preditoras é demasiadamente vasto, o processo de adicionar a maior quantidade de variáveis possíveis torna-se desnecessariamente oneroso, justamente por não trazer nenhuma garantia de trazer mais qualidade para a predição.

Sendo assim, para melhor contextualização ao longo do estudo, será feita uma explicação levemente elaborada especificamente das variáveis utilizadas que não são de conhecimento mais amplo do jogo de basquete.

A métrica `Game Score` tem como objetivo avaliar o desempenho em quadra de um jogador, criada por John Hollinger em 2007 (Midfield, 2024), que utiliza uma forma para ponderar estatísticas de *box score*, como pontos, arremessos de quadra convertidos, rebotes ofensivos, rebotes defensivos, roubos de bola, assistências, entre outros. É, resumidamente, uma versão simplificada de outra métrica criada por *Hollinger*, o `PER`, *Player Efficiency Rating* (HOLLINGER, 2004). Diferente do `Game Score`, o `PER` também leva em

consideração o desempenho da equipe na partida, para bonificar ou penalizar a atuação do jogador. Por conta da natureza de sua formulação, o `PER` tem uma média de 15 para cada temporada da liga, sendo que, por exemplo, um jogador com `PER` igual a dez pode ser considerado como bem abaixo da média, considerando esta métrica especificamente.

Já o `Plus Minus` tem o foco de analisar o impacto que o devido atleta tem no placar, considerando as mudanças na pontuação que ocorrem quando ele está ou não em quadra. Sua fórmula é "Pontos anotados com o jogador em quadra" - "Pontos cedidos com o jogador em quadra", sendo que valores positivos indicam que o atleta tem um impacto positivo em aumentar a vantagem do seu time no placar. Com isso, é possível que cada time analise seus jogadores em grupo, com o intuito de encontrar trios, quartetos ou até quintetos ideais, que juntos possuam o maior impacto em ampliar sua vantagem no jogo. Uma versão derivada dessa métrica é o `BPM`, conhecido como *Box Plus Minus*, criada por Daniel Myers, que, assim como o `PER`, compara o desempenho em relação a média geral da liga (nesse caso, com a média em zero). Ou seja, quanto mais acima de zero seja o BPM de um jogador, melhor foi seu desempenho naquela partida ou temporada, dependendo de qual foi o escopo do cálculo.

Para a métrica `eFG`, *Effective Field Goal Percentage* (Haefner, 2021), sua contribuição principal é de levar em consideração que um arremesso de fora do perímetro, costumeiramente conhecido como um "arremesso de três", vale mais pontos do que um arremesso de dentro do perímetro, exatamente um ponto a mais. Sendo assim, o `eFG` é calculado por (`FG` + 0.5 * `3P`) / `FGA`.

O `TSA` (Reference, 2024c), *True Shooting Attempts* busca levar em consideração outro tipo de pontuação em um jogo de basquete, um lance livre. Traz uma visão que "complementa" a análise conjuntamente com o `eFG`, sendo calculado por `FGA` + 0.44 * `FTA`. Ainda sobre lances livres, o `FTR`, *Free Throw Rate*, mensura a habilidade que uma equipe

e/ou jogador possuem de conseguirem ir para a linha de lance livre, forçando faltas do adversário, sua fórmula é dada por `FTA/FGA`. Finalmente, o `TOV%`, *Turnover Percentage*, serve como uma estimativa da quantidade de *turnovers*, ou seja, perdas de posse, a cada 100 jogadas. Sua fórmula é dada por 100 * `TOV` / (`FGA` + 0.44 * `FTA` + `TOV`). Por exemplo, um jogador com `TOV%` de 5% comete cinco *turnovers* a cada 100 posses de bola com ele.

A variável de interesse foi construída como um índice de desempenho, com o objetivo de mensurar o desempenho do atleta em conjunto com o resultado da partida. O principal componente é a métrica `BPM` (acrônimo de *Box Plus Minus*), que é uma versão aprimorada do `Plus Minus`, buscando mensurar o impacto do atleta no resultado enquanto ele está jogando. Além de analisar a diferença no placar durante a presença do jogador em quadra, são utilizadas as métricas simples de *box score*, como: pontos, assistências, entre outros.

O `BPM` puramente é calculado de modo a manter a média de todos os jogadores da liga em zero. Sendo assim, atletas com desempenhos maiores que zero no `BPM` trazem uma contribuição acima da média para seu time, enquanto que atletas com `BPM` abaixo de zero são costumeiramente os últimos reservas da equipe.

Para compor a variável resposta junto com o `BPM` foram levados em considerações fatores em relação ao resultado do jogo, como a margem da vitória ou derrota, assim como uma penalização ou acréscimo no valor final de acordo com o resultado puro: vitória ou derrota.

A fórmula foi definida levando em consideração conceitos esportivos, com o intuito de agregar na métrica uma ponderação referente ao resultado final da partida. A construção é definida por:

$$Y = BPM + \frac{M}{100} * FR,$$

sendo que `M` representa a margem do placar, ou seja, a subtração dos pontos do vencedor e do perdedor. `FR` é a componente que considera se o time do jogador venceu a partida e se venceu no período regular. As ponderações foram decididas com o intuito de levar em consideração a importância das vitórias, além de ponderar também o fato do jogo ir para prorrogação, visto que jogar um período a mais, no mínimo, é custoso para todos jogadores.

- Se venceu e foi na prorrogação, o fator `FR` é $1,1$;
- se venceu no período regular, o fator `FR` é $1,15$;
- se perdeu e foi na prorrogação, o fator `FR` é $0,9$;
- se perdeu no período regular, o fator `FR` é $= 0,85$.

2.2 Metodologias

2.2.1 Modelos Lineares Generalizados

Os Modelos Lineares Generalizados foram apresentados inicialmente por NELDER e WEDDERBURN (1972) e popularizados por McCULLAGH e NELDER (1989) como uma proposta de melhoria para os utilizados na época. Tratavam-se de uma compilação de técnicas estatísticas que trouxeram uma grande robustez para os métodos paramétricos, com aplicações em diversas áreas, principalmente na área financeira (DINIZ e LOUZADA NETO, 2012).

Considerando as características de modelos lineares, sendo: as observações da amostra são i.i.d. (independentes entre si e identicamente distribuídas), as covariáveis são representadas por uma matriz $\mathbf{X}$ (matriz de delineamento, matriz de planejamento) e um vetor amostral das observações referentes a variável resposta de interesse. McCULLAGH

e NELDER (1989) define os três componentes presentes na classe dos MLG's. Em geral, a utilização de um MLG é mais adequada para problemas onde uma abordagem paramétrica (família de distribuições com número finito de parâmetros) é a ideal.

Porém, modelos paramétricos apresentam algumas desvantagens (SEGDWICK, 2015). Uma delas é lidar com conjuntos de dados no qual o número de covariáveis é superior ao de observações, assim como é necessário que os dados satisfaçam algumas suposições. Além disso, para alguns problemas o objetivo definido inicialmente não é referente a analisar a relação covariável(is)-resposta, mas sim estimar uma função que traga bons resultados preditivos.

Um MLG é composto por três componentes principais: a função de ligação, a distribuição da família exponencial e o modelo linear preditor. A função de ligação estabelece uma relação entre a média da variável resposta μ e a combinação linear dos preditores η, de forma que $g(\mu) = \eta$. Isso permite que η assuma qualquer valor real, enquanto μ está restrita ao intervalo da função de ligação. A variável resposta Y é assumida como pertencente a uma família exponencial de distribuições, como: Normal, Binomial, Poisson, Exponencial, Geométrica etc. Por fim, o preditor linear η é uma combinação linear dos preditores $X_1, X_2, \ldots, X_p$, onde $\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$, sendo β_i os coeficientes a serem estimados.

A variável resposta Y segue uma distribuição de probabilidade que é membro da família exponencial. A família exponencial uniparamétrica comporta distribuições uniparamétricas que suas funções de densidade podem ser escritas na forma:

$$f(y_i;\theta) = h(y_i)\, \exp\left\{A(\theta_i)t(y_i) - b(\theta_i)\right\},$$

em que θ é o parâmetro de interesse e $h(), A(), t()$ e $b()$ são funções reais. São definidas

por, como visto em Casella e Berger (2002):

- $h()$: o fator de escala, uma função que ajusta a densidade de probabilidade para garantir que seja válida;
- $A()$: função de ligação canônica;
- $t()$: função de estatística suficiente, depende da distribuição escolhida.
- $b()$: função de log partição.

Especialmente para os Modelos Lineares Generalizados, considerando a forma geral definida em (2.2.1) na forma canônica, com $A(\theta) = 1$ e $t(y_i) = 1$, adicionando um parâmetro $\phi > 0$:

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{1}{a_i(\phi)}[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\right\},$$

em que θ_i é dito parâmetro canônico, ϕ o parâmetro de dispersão e $a_i(\phi) = \frac{\phi}{w_i}$ com w_i sendo pesos definidos *a priori* para cada observação. Ademais, a esperança e variância da variável resposta pode ser definida como

$$E(Y_i) = \mu_i = b'(\theta_i) \text{ e } Var(Y) = a(\phi_i)b''(\theta_i) = a_i(\phi)V_i,$$

com V_i conhecida como a função de variância, dada por $V_i = \frac{d\mu_i}{d\theta_i}$. As covariáveis do modelo, inclusas na matriz $X_{i,j}$, são relacionadas aditivamente no preditor linear, com $i, j = 1, ..., n$, dado por

$$\eta_i = x_i^T\beta,$$

com β sendo o vetor de parâmetros, assim como em outros modelos lineares.

Função de Ligação

Um dos primeiros passos para modelar é a escolha da função de ligação. Essa escolha depende da natureza da variável de interesse. A função identidade, dada como $g(\mu) = \mu$ é a escolha padrão quando a variável resposta é numérica e pode assumir qualquer valor real, seja ele positivo ou negativo. Neste caso, a função de ligação é simplesmente a própria média da variável resposta, ou seja, não há transformação aplicada. Esta escolha é comum quando a distribuição normal (gaussiana) é utilizada, especialmente em contextos onde se assume que a variável resposta tem uma distribuição simétrica em torno da média e variância constante.

Já a função de ligação logarítmica, dada como $g(\mu) = \log(\mu)$ é usada quando a variável resposta é positiva e a relação entre os preditores e a resposta é multiplicativa. Isso significa que as mudanças nos preditores resultam em mudanças relativas (percentuais) na média da variável resposta. Esta função é comum em modelos com distribuição gama ou Poisson, que são adequadas para variáveis que assumem valores estritamente positivos e que frequentemente exibem assimetria.

Outra possibilidade é usar a função de ligação Logit, dada como $g(\mu) = \log(1 - \mu)$, que é amplamente utilizada em modelos logísticos, onde a variável resposta é binária, assumindo valores 0 ou 1. Com esta, é possível modelar a probabilidade de ocorrência de um evento como uma função linear dos preditores. Semelhante a Logit, temos a função de ligação Probit, que é construída pela inversa da função de distribuição acumulada da Normal $g(\mu) = \Phi^{-1}(\mu)$. A Probit é também usada em modelos onde a variável resposta é binária, servindo como uma alternativa a Logit (JESUS, 2015).

Uma outra alternativa é a função de Ligação Inversa, dada como $g(\mu) = \frac{1}{\mu}$ e que é utilizada em modelos onde a variável resposta seja estritamente positiva e a associação entre os preditores e a resposta segue uma estrutura inversa. Esta função é apropriada

em modelos com distribuição inversa gaussiana, comum em situações onde a resposta é o tempo até a ocorrência de um evento.

Uma das mais utilizadas técnicas avaliar a qualidade do modelo e a escolha da função de ligação é a validação cruzada.

Validação Cruzada

A validação cruzada é uma técnica utilizada para avaliar a capacidade preditiva de um modelo. O procedimento inicia pela divisão do conjunto de dados em teste e treinamento (ou desenvolvimento e validação), procedimento este que é bastante comum em modelagens. A seguir, o ajuste é feito na partição de treinamento e analisado na partição de teste, correndo várias iterações deste procedimento, testando diversas partições diferentes para esse cruzamento até encontrar a partição que apresenta os melhores resultados. Existem diferentes maneiras de fazer essa validação, sendo a *k-fold* e *leave-one-out* as mais comuns (Learn, 2024).

A validação cruzada *k-fold* consiste dividir os dados em k partes (ou folds). Assim, o modelo é treinado em $k-1$ partes e testado na parte restante. Esse processo é repetido k vezes, de forma que cada parte seja utilizada uma vez como conjunto de teste/validação. Já o procedimento *leave-one-out* é um pouco diferente, consistindo em selecionar apenas uma observação para cada iteração que será utilizada como conjunto de teste, com o modelo sendo treinado nas demais. O processo será repetido até que todas observações sejam utilizadas ao menos uma vez como conjunto de teste/validação (T Hastie, 2001).

Ao utilizar validação cruzada, são calculadas métricas de performance preditiva para os modelos ajustados em cada iteração, para um conjunto de funções de ligação que são de interesse para serem testadas. Sendo assim, o modelo com a função de ligação que apresentar as melhores métricas em mais iterações servirá como parâmetro para definição

da função de ligação.

Família de Distribuições e Modelo Preditor

Os parâmetros β de um MLG são usualmente estimados pelo método de máxima verossimilhança (MLE, *Maximum Likelihood Estimator*). A função de verossimilhança é construída a partir da distribuição da família exponencial e é maximizada em relação aos parâmetros β. Em alguns casos, a estimação pode ser feita usando algoritmos iterativos, como o método iterativo de reponderação dos mínimos quadrados (IRLS), que é uma generalização do método dos mínimos quadrados para MLG's (Dobson e Barnett, 2008). A avaliação do modelo é feita principalmente pela análise dos resíduos, que são as diferenças entre os valores observados e os valores ajustados pelo modelo, permitindo identificar possíveis problemas de ajuste. Além disso, a adequação do modelo pode ser verificada através de medidas como *deviance*, que compara a qualidade do ajuste de diferentes modelos. A análise de outliers e a verificação da suposição de independência dos erros também são etapas cruciais na validação do modelo (Agresti, 2015).

NELDER e WEDDERBURN (1972) também propuseram uma medida da qualidade do ajuste do modelo, conhecida como *deviance*. A *deviance* total é definida por

$$D = \sum_{i=1}^{n} d_i^2,$$

com d_i^2 sendo a componente de *deviance* para cada uma das observações e obtida por

$$d_i = -2 \int_{y_i}^{\mu_i} \frac{y_i - u}{V(u)} du. \tag{2.1}$$

Tomando $V(u_i)$ e μ_i para distribuições pertencentes a família exponencial canônica com forma geral vista em (2.2.1), tem-se que para a Distribuição Poisson $V(u_i) = 1$ e

$\mu_i = \exp(\eta_i)$. Com isso, é possível reescrever (2.1) como

$$d_i^2 = w_i \left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i) \right].$$

Uma das maneiras para medir a qualidade do ajuste do modelo aos dados é analisar seus resíduos, que são a diferença entre os valores observados e seus preditos para a variável resposta de interesse. O resíduo *deviance*, que é um dos utilizados para analisar qualidade de ajuste de modelos lineares (resíduo ordinário, resíduo de Pearson, resíduo de Pearson padronizado), é calculado por $r_i^D = \text{sinal}(y_i - \hat{\mu}_i\sqrt{d_i})$, em que a função sinal() é igual a -1 quando $x < 0$ e igual a 1 quando $x > 0$.

2.2.2 Árvores de Particionamento Recursivo

As Árvores de Decisão são uma das técnicas de particionamento recursivo, sendo divididas em Árvores de Classificação e Árvores de Regressão. Para esse estudo, serão utilizadas apenas Árvores de Regressão, sendo assim o desenvolvimento teórico será sobre estas.

Por ser um método não paramétrico, o particionamento recursivo apresenta algumas vantagens em relação a métodos mais tradicionais, como regressão linear, logística, entre outros. Uma das vantagens é a possibilidade de associações não lineares entre as covariáveis, que não precisam ser explicitamente especificadas antes do ajuste da árvore. Outra vantagem é em relação a utilização de interações de grau maior que dois (STROBL, MALLEY e TUTZ, 2009).

O algoritmo CART® proposto por BREIMAN *et al.* (1984) e os algoritmos ID3 e C4.5 (sucessor do ID3) propostos por QUINLAN (1986) são os mais tradicionalmente usados em estudos de Árvores de Decisão.

Para definição dos pontos de corte, para covariáveis categóricas ou numéricas, os algoritmos utilizam de medidas de impureza com o Índice de Gini sendo o mais comum.

Os particionamentos realizados nas árvores são chamados de nós, enquanto que um nó terminal, ou seja, aquele que apresenta um resultado de certo particionamento, é chamado de folha, já o nó inicial pode ser conhecido como raiz. A ligação entre nós é denominado como ramo, sendo que o tamanho dos ramos é variável. Todos os processos de criação de uma árvore são graficamente simples de entender, o que facilita a compreensão para públicos não especialistas, assim como a disseminação da metodologia para diferentes áreas do conhecimento.

O algoritmo de partição começa pelo primeiro nó e verifica sua condição (a covariável pode ser numérica ou categórica). Como os algoritmos implementados no pacote *rpart* dividem cada nó em dois, será considerado que se as condições dos nós forem satisfeitas a árvore seguirá a esquerda, e a direita caso contrário.

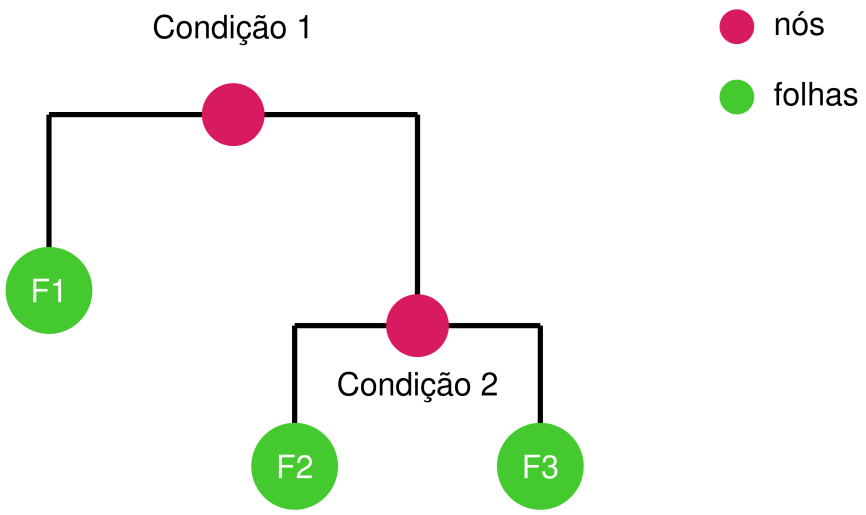

Figura 2.1: Exemplo de uma Árvore de Regressão simples. Fonte: (IZBICKI e SANTOS, 2020).

Na Figura 2.1 vê-se um exemplo simples do procedimento de uma Árvore de Parti-

cionamento Recursivo. Tudo inicia-se no primeiro nó, de cor vermelha, quando testa-se cada observação em relação a Condição 1 (primeira covariável da árvore). Caso satisfeita, segue-se a esquerda e a direita caso contrário, como dito anteriormente.

Caso a Condição 1 seja satisfeita, o procedimento chega ao fim e a observação pertencerá a Folha 1, de cor verde. Caso não seja, o procedimento segue para o teste da Condição 2, em vermelho. Após isso, depende do resultado, a observação pertencerá a Folha 2 ou a Folha 3, concluindo o algoritmo.

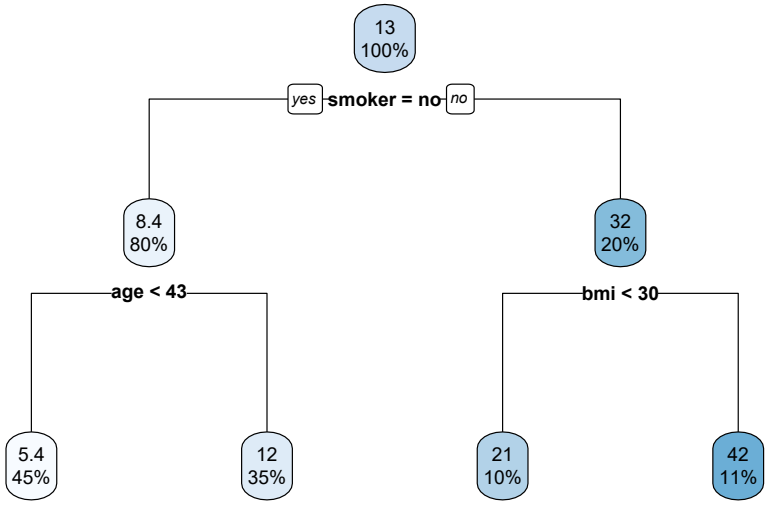

Figura 2.2: Exemplo de uma Árvore de Regressão sobre preços de seguro de saúde nos EUA. Fonte: Choi (2017)

Na Figura 2.2 vemos outro exemplo de construção de uma Árvore de Regressão, usando dados de para determinar o custo de seguros médicos (*insurance*) para pacientes internados em hospitais no EUA. Vê-se que nos nós os particionamentos podem ser feitos em covariáveis quantitativas e qualitativas apresentando mais uma vantagem dessa técnica.

A Árvore de Regressão utilizou tanto uma covariável qualitativa `Smoker`, particionando-a em "Yes" e "No", tanto covariáveis quantitativas como `Age` e `BMI`. Além disso, o processo automaticamente não utiliza as covariáveis no conjunto que não sejam relevantes para explicar a variável resposta, fazendo uma seleção de variáveis automaticamente (o que não

acontece em muitos modelos paramétricos, como os Modelos Lineares Generalizados, onde o ajuste do modelo é feito após a aplicação de um método de seleção de variáveis).

A forma geral do modelo, como definida por CLARKE, FOKOUÉ e ZHANG (2009) é:

$$Y = \sum_{i=1}^{j} \beta_j I(x \in R_i) + \epsilon.$$

As partições, $R_1, ..., R_j$, que são criadas nas covariáveis em estudo são distintas e disjuntas (IZBICKI e SANTOS, 2020), pois cada uma cria caminhos diferentes e sem intersecção entre si. Já os coeficientes da regressão β_j são estimados a partir dos dados; porém esses coeficientes β_j podem ser substituídos por outras funções de regressão para cada região R_i, oferecendo alternativas ao modelo aditivo.

Com isso, podemos utilizar a função

$$g(x) = \frac{1}{|R_i|} \sum_{i:x_i \in R_i} y_i, \tag{2.2}$$

vista em (2.2) para prever os valores da variável resposta. A função observa a região R_i de cada observação e calcula uma média da variável resposta y_i da região R_i.

Porém, a criação de uma árvore não é um procedimento simples e linear. Inicialmente, o algoritmo utilizado cria a árvore mais complexa possível, utilizando de todos os particionamentos recursivos a sua disposição. Após isso, referenciando novamente a situações que ocorrem com as árvores seres vivos, ocorre o processo de poda, com o intuito de ajustar o modelo de maneira mais adequada, para que ele utilize os apenas os recursos necessários para resolver o problema, evitando o *overfitting*.

2.2.3 Criação da Árvore Complexa

Inicialmente busca-se construir a melhor árvore/modelo possível, com o objetivo de ter as partições nós e folhas com maior homogeneidade. Para medir a razoabilidade de criar a árvore mais homogênea são propostos alguns indicadores, com um deles sendo o Erro Quadrático Médio (EQM):

$$EQM = \sum_{R} \sum_{i:x_i \in R} \frac{(y_i - \hat{y_R})^2}{n},$$

em que $\hat{y_R}$ representa o predito de cada observação pertencente a uma região R de interesse, para uma amostra de tamanho n. Porém, ainda é demasiado oneroso computar o EQM mínimo, o que leva a alterações no algoritmo com o intuito de diminuir os gastos computacionais e temporais.

O pacote *rpart* (THERNEAU e ATKINSON, 2019) do *software* de licença livre R (R Core Team, 2019), que será utilizado para a implementação desse estudo, apresenta algumas medidas para medir a redução do EQM nos nós. Tais medidas são utilizadas para auxiliar na decisão de escolha das covariáveis e partições que serão feitas em cada nó. A medida de Informação é definida por:

$$I(N) = \sum_{i=1}^{C} f(p_{iN}). \tag{2.3}$$

Com p_{iN} sendo a proporção de observações que estão no nó N e pertencerão a classe i em novas amostras, f sendo uma função de impureza e C a quantidade de classes. Então, temos que em (2.3) são somadas a função de impureza para um determinado nó em todas as classes.

Idealmente deseja-se que $I(N) = 0$, indicando que o nó N separa as observações perfeitamente, com a proporção de observações futuras que vão para a classe i após

passarem por N igual a um.

O pacote implementa duas alternativas para a função genérica f de impureza. O Índice de Informação toma $f(p) = -p\log(p)$ e o Índice de Gini toma $f(p) = p(1-p)$.

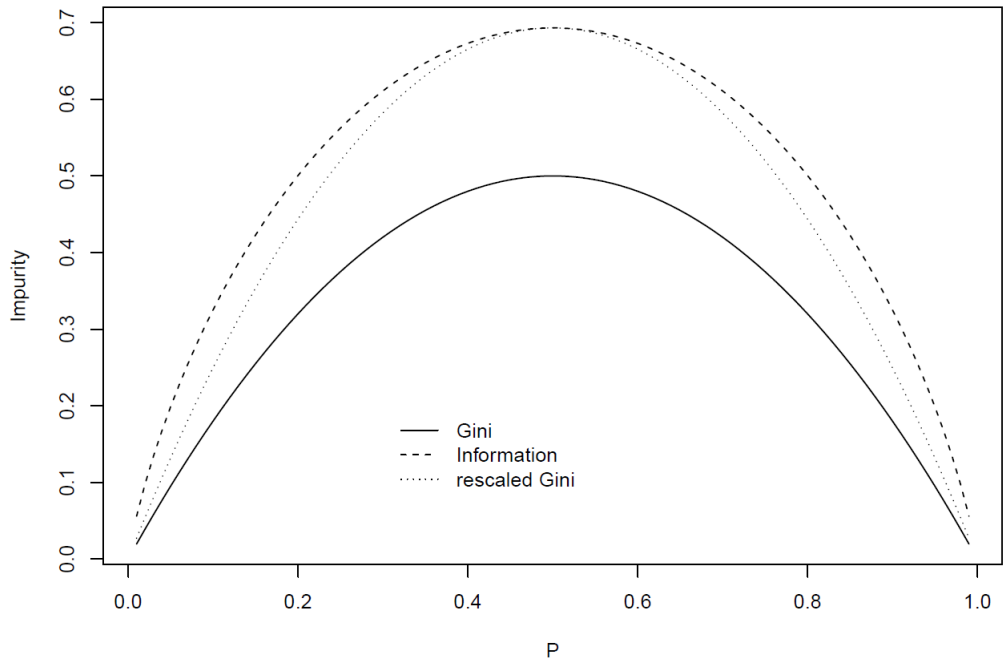

Figura 2.3: Comparação entre os índices de Informação e Gini. (THERNEAU e ATKINSON, 2019)

Na Figura 2.3 são apresentados os Índices de Informação, Índice de Gini e uma versão reescalada do Índice de Gini descritos na vinheta do pacote *rpart*, em uma situação de partição de um nó em duas classes/regiões. Percebe-se que para essa situação, o Índice de Informação e o Índice de Gini reescalado são curvas bem próximas e resultam na mesma partição.

Uma outra implementação, optando pelo método *anova* no pacote *rpart* THERNEAU e ATKINSON (2019), de particionamento é

$$SQ_N - (SQ_E - SQ_D), \tag{2.4}$$

onde $SQ_N = \sum(y_i - \bar{y})^2$ é a soma de quadrados do nó N, enquanto que SQ_E e SQ_D

são as somas de quadrados para a partição esquerda e direita, respectivamente, no nó N. Aqui, busca-se a partição que maximiza a soma de quadrados intragrupos.

Uma das opções é iniciar particionando todas as covariáveis em duas regiões distintas e disjuntas, regiões estas que possuam o menor EQM possível (2.2.3). IZBICKI e SANTOS (2020) apresenta essas regiões iniciais como:

$$R_1 = \{\mathbf{x} : x_i < t_1\} \text{ e } R_2 = \{\mathbf{x} : x_i \geq t_1\}, \tag{2.5}$$

onde x_i representa uma covariável do conjunto e t_1 representa esse primeiro corte que separa as regiões R_1 e R_2. Ou seja, a região R_1 comportará as observações que a covariável x_i é menor ou segue um determinado critério de corte t_1, enquanto que a região R_2 comportará as demais.

Com a definição do nó inicial, o passo seguinte é dividir as regiões R_1 e R_2, particionando-as de maneira análoga, e assim seguir até o final da árvore. A maior diferença é que agora uma das regiões destas precisa ser escolhida para sofrer o particionamento em um conjunto de covariáveis x_k e levando em consideração um corte t_2.

Os particionamentos das regiões apresentarão forma parecida com (2.5). Caso a decisão do algoritmo seja de particionar R_2:

$$R_1 = \{\mathbf{x} : x_i < t_1\} \text{ , } R_{2,1} = \{\mathbf{x} : x_i \geq t_1, x_k < t_2\} \text{ e } R_{2,2} = \{\mathbf{x} : x_i \geq t_1, x_k \geq t_2\}.$$

Após isso o particionamento recursivo será realizado para as demais regiões, R_1, $R_{2,1}$, $R_{2,2}$ e assim sucessivamente, até que atinja um critério de parada (definido manualmente) e o algoritmo chegue ao fim. Tal critério pode variar, porém uma das alternativas é definir o número de observações esperadas em cada folha, de acordo com o tamanho amostral, onde o processo termina quando as regiões das folhas atingem esse critério.

Finalizado o ajuste da árvore mais complexa é necessário fazer análises sobre o ajuste desse modelo, que como possui apenas um critério de parada, pode levar a um super-ajuste. Um modelo com super-ajuste, também referenciado pela alcunha em inglês *over-fitting*, funciona muito bem para uma amostra específica (a usada para treiná-lo). Entretanto, devido a essa especificidade ele sofre com a generalização, ou seja, seu poder de predição para novas amostras é pequeno.

Para evitar isso, existe o processo de poda da árvore, que tem o objetivo de podar o modelo para aumentar sua generalização e poder preditivo.

Uma outra opção, que é pouco utilizada, seria de usar procedimentos de parada durante o ajuste da árvore. Porém, esses procedimentos que focam em entender se a próxima partição de um nó traria alguma melhoria para o modelo levam a *underfitting*.

Poda da Árvore

Para melhorar o modelo inicial, a poda é feita para diminuir sua complexidade e melhorar seu poder para predizer novas observações.

O pacote *rpart* (THERNEAU e ATKINSON, 2019), no software R, implementa uma determinada maneira para o processo da poda, que será aqui descrito. Sejam $f_1, ..., f_c$ as folhas de uma árvore A. Define-se como $|A|$ o número de folhas da árvore e

$$R(A) = \sum_{i-1}^{c} P(f_i)R(f_i)$$

como o risco inerente a árvore $|A|$. Em relação a modelos de regressão, $|A|$ é análogo aos graus de liberdade, assim como $R(A)$ é a Soma de Quadrados dos Resíduos.

Considerando α um número real positivo que será utilizado para medir o "custo" de particionar uma nova covariável durante o ajuste do modelo, ou seja, α servirá como um critério para medir a complexidade do modelo. Seja $R(A_0)$ o risco de uma árvore que

ainda não sofreu nenhuma partição, tem-se:

$$R_\alpha(A) = R(A) + \alpha|A|,$$

como o custo de uma determinada árvore A, com α sendo um peso, utilizado para balancear minimizar o EQM e a complexidade do modelo. Sendo assim, A_α é definida como uma sub-árvore da árvore mais complexa e que possui custo mínimo dentre todas.

Desse modo, o interesse principal é encontrar uma única sub-árvore A_α em que o $R_\alpha(A)$ seja mínimo. Para encontrar A_α, alguns resultados são importantes:

1. Se A_1 e A_2 são sub-árvores da árvore complexa A com $R_\alpha(A_1) = R_\alpha(A_2)$, então A_1 é uma sub-árvore de A_2 ou A_2 é uma sub-árvore de A_1; então $|A_1| < |A_2|$ ou $|A_2| < |A_1|$;

2. Se $\alpha > \beta$ então $T_\alpha = T_\beta$ ou T_α é uma sub-árvore de T_β;

3. Sejam $\alpha_1, \alpha_2, \ldots, \alpha_m$ números reais; então $R_{\alpha_1}, R_{\alpha_2}, \ldots, R_{\alpha_m}$ e $T_{\alpha_1}, T_{\alpha_2}, \ldots, T_{\alpha_m}$ podem ser computados de maneira eficiente.

Considerando qualquer conjunto de sub-árvores de A, seu número máximo de folhas será $|A|$. Utilizando o resultado 2, tem-se que todos os possíveis valores de α podem ser agrupados em m intervalos, sendo $m \geq |A|$.

$$I_1 = [0, \alpha_1] \ ; \ I_2 = (\alpha_1, \alpha_2] \ ; \ldots ; \ I_m = (\alpha_{m-1}, \infty],$$

em que todos os $\alpha \in I_i$ são da mesma sub-árvore que minimiza (2.2.3). A maneira de encontrar o melhor valor possível para α é por validação cruzada. Os seguintes passos são implementados no R e descritos em THERNEAU e ATKINSON (2019):

1. Ajustar o modelo da árvore complexa e depois computar I_n descrito em (2.2.3). Para cada I seu "valor típico" é descrito por

$$\beta_1 = 0;$$

$$\beta_2 = \sqrt{\alpha_1 \alpha_2};$$

$$\vdots$$

$$\beta_{m-1} = \sqrt{\alpha_{m-2} \alpha_{m-1}};$$

$$\beta_m = \infty.$$

2. O conjunto de dados será dividido em g grupos $G_1; G_2; \ldots; G_g$ com tamanho homogêneo g/n. Para cada um dos grupos:

   - Será ajustada uma árvore complexa, excluindo o grupo G_i, e serão computados $A_{\beta_1}; A_{\beta_2}; \ldots; A_{\beta_m}$;
   - Serão calculados os valores de preditos, as folhas, das observações do grupo G_i para todas as árvores A_{β_j} com $1 \geq j \geq m$;
   - Computa-se o risco (2.2.3).

3. Somam-se os grupos G_i para obter uma estimativa do risco para cada β_j. Para o β com o menor risco, ajusta-se a árvore A_β para o conjunto completo, de treinamento, e esta é escolhida como a melhor árvore pós-poda.

Uma outra possibilidade, que praticamente pode trazer vantagens em determinados casos é uma análise gráfica de β *vs* risco. Em maioria, o comportamento dessa visualização é de uma grande queda inicial, seguida por um platô e um aumento progressivo. Uma maneira é escolher um dos modelos que está no platô, escolhendo aquele que apresenta menor risco e erro padrão desse risco.

CLARKE *et al.* (2009) descreve um processo de poda com mínimo custo e complexidade, que compartilha muitas semelhanças com o processo descrito em THERNEAU e ATKINSON (2019).

Inicialmente, a partir da árvore complexa criada, escolhe-se um nó arbitrário e retira-se todos os nós das partições subsequentes a este. O critério utilizado é definido como:

$$C(A;\alpha) = EQM + \alpha|A|. \tag{2.6}$$

Tem-se que $|A|$ define o número de folhas (nós terminais) de uma árvore A, enquanto que o EQM visto em (2.2.3) mantém-se como um indicador de qualidade das partições. O termo α é um peso, utilizado com o objetivo de determinar a importância do equilíbrio entre minimizar o EQM e a complexidade do modelo.

O critério formalmente descrito em (2.6) busca encontrar algum nó que esteja mais próximo a raiz (nó inicial) buscando um equilíbrio entre menor erro e complexidade do modelo. Caso haja algum tipo de empate na escolha do nó, leva-se em consideração aquele que possui mais nós em partições seguintes.

Busca-se árvores com os menores valores possíveis de $C(A;\alpha)$, visto que na construção leva-se em consideração minimizar o Erro Quadrático Médio e pesos (α) muito grandes acabam por penalizar árvores muito grandes, pois possuem maior complexidade e um maior número de folhas, e acabam por indicar pela escolha de árvores menores. Entretanto, α muito pequenos podem levar a escolha de árvores com *overfitting*.

Sendo assim, avaliando os extremos de α têm-se: com α muito grande a melhor escolha é de uma árvore com somente um nó (pouco complexa), com α muito pequeno a melhor escolha seria de uma árvore grande (vários nós) e muito complexa. Ou seja, é necessário escolher um valor ótimo de α para balancear tais extremos.

O processo avalia todos os nós da árvore, individualmente, retirando-os do modelo e observando a variação que isso causa no EQM. Com isso, é possível criar uma espécie de *ranking* dos nós em relação a sua influência no modelo. Sendo assim, é possível decidir analiticamente quais destes serão retirados da árvore, para que seja feito um novo ajuste. Basicamente, o processo de poda acaba por juntar algumas das regiões R_i, que tinham sido particionadas anteriormente, em um nó comum.

Capítulo 3

Resultados

Definidas todas as nuances que serão analisadas em quadra, o contexto geral dos dados utilizados e as metodologias que serão empregadas para a predição, é de suma importância uma análise descritiva e exploratória que possa apresentar características importantes do conjunto de dados, assim como indicar elementos de atenção que possam ser fatores de muita importância no momento de predição.

3.1 Análise Descritiva e Exploratória

Para entender o comportamento e a vantagem da utilização de métricas avançadas para predição, serão analisadas algumas medidas descritivas para cada uma destas.

Tabela 3.1: Estatísticas resumo de métricas avançadas.

Estatística	*Game Score*	BPM	*Plus Minus*	PER	eFG	TSA
Média	9.42	-0.23	0.42	15.93	0.51	10.89
Mediana	8.30	-0.20	0.00	15.58	0.50	9.88
Mínimo	-9.80	-45.10	-53.00	-41.11	0.00	0.00
Máximo	60.80	61.10	57.00	99.89	1.50	55.28
Desvio	7.43	8.28	11.63	11.70	0.23	6.28
Coeficiente Variação	0.788	36	27.69	0.734	0.45	0.576

Pela Tabela 3.1 podemos notar algumas características de cada uma das métricas

avançadas que foram destacadas. O `BPM` é uma métrica que apresenta muita amplitude, com mínimo e máximo de $-45,1$ até $61,1$, valores bem mais distantes do padrão de avaliação proposto por Myers, que é entre -4 e 12. Também no `Plus Minus`, uma versão mais simples do `BPM`, notamos os *outliers*, porém em ambos a média e mediana estão próximos de zero, o que indica que em média os jogadores possuem participação neutra nas partidas analisadas.

As demais métricas também apresentam grande amplitude. O *Game Score* e `PER` são métricas construídas de maneira similar, mas apresentam comportamento diferente, justificando a utilização de ambas para o início de modelagem. O `eFG` está entre 0 e $1,5$, com valores bons sendo a partir de $0,5$. O `TSA` apresenta média $10,89$, representando a quantidade média de arremessos que os jogadores tentam por partida.

Agora, considerando mais as métricas costumeiramente analisadas em jogos de basquete, é também valioso entender o seu comportamento por meio do cálculo de medidas resumo. Para facilitar a análise, serão divididas em dois blocos: ofensivas e defensivas. O primeiro pode ser visto na Tabela 3.2.

Tabela 3.2: Estatísticas resumo de métricas ofensivas.

Variável	`Pontos`	`FG%`	`2P%`	`3P%`	`FT%`	`TS%`	`ORB`	`AST`
Média	12.10	0.45	0.50	0.33	0.76	0.54	1.15	2.66
Mediana	11.00	0.46	0.50	0.33	0.83	0.54	1.00	2.00
Mínimo	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Máximo	71.00	1.00	1.00	1.00	1.00	1.50	18.00	25.00
Desvio	8.22	0.20	0.25	0.29	0.28	0.22	1.47	2.70
Coeficiente Variação	0.679	0.444	0.5	0.878	0.368	0.407	1.278	1.015

Pela Tabela 3.2 notamos uma média mais modesta de pontos, de $12,1$, carregadas pelos jogadores mais regulares. Além dos valores *outliers* que aparecem em todas as métricas, é possível notar que uma análise geral não traz informações muito valiosas, visto que as características diferentes dos jogadores se perdem nas médias.

Agora, para o bloco de informações que analisam o trabalho defensivo dos jogadores,

temos a Tabela 3.3.

Tabela 3.3: Estatísticas sumárias de métricas defensivas.

Variável	STL	BLK	TOV	TOV%	PF
Média	0.85	0.53	1.54	12.80	2.15
Mediana	1.00	0.00	1.00	11.11	2.00
Mínimo	0.00	0.00	0.00	0.00	0.00
Máximo	10.00	12.00	12.00	100.00	6.00
Desvio	1.02	0.90	1.46	12.46	1.44
Coeficiente Variação	1.2	1.698	0.948	0.973	0.669

Pela Tabela 3.3 notamos um comportamento similar nas métricas, grande amplitude e desvio padrão considerável. Vemos *Outliers* nos valores máximos e mínimos, assim como médias e medianas mais baixas, devido ao fator de analisar em conjunto um grupo muito heterogêneo.

Para complementar a análise tabular, serão apresentados gráficos referentes as métricas, agrupados com o intuito de trazer mais fluidez para a análise, além de buscar algumas características específicas para ajudar na predição.

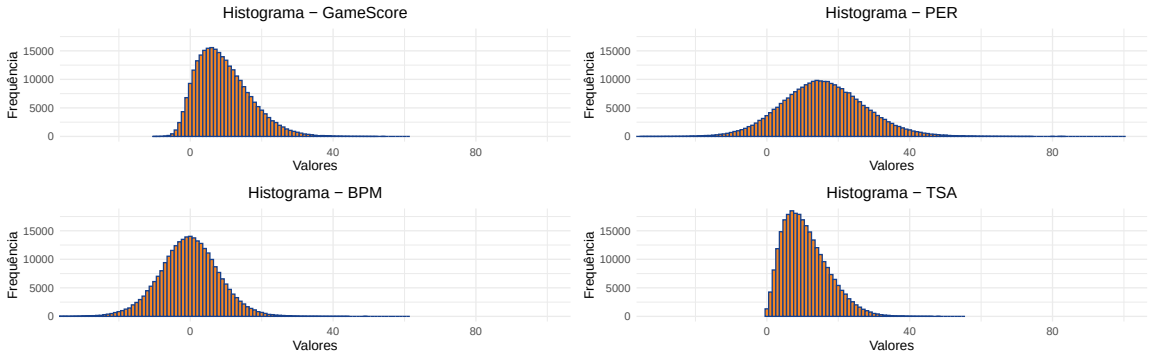

Figura 3.1: Histogramas das principais métricas avançadas.

Pela Figura 3.1 podemos observar a distribuição de algumas métricas, com destaque especial para `GameScore` e `TSA` que apresentam assimetria, com maior concentração na cauda inferior. Já para `BPM` e `PER` a distribuição é mais simétrica, o que é esperado devido a maneira de como tais métricas foram construídas.

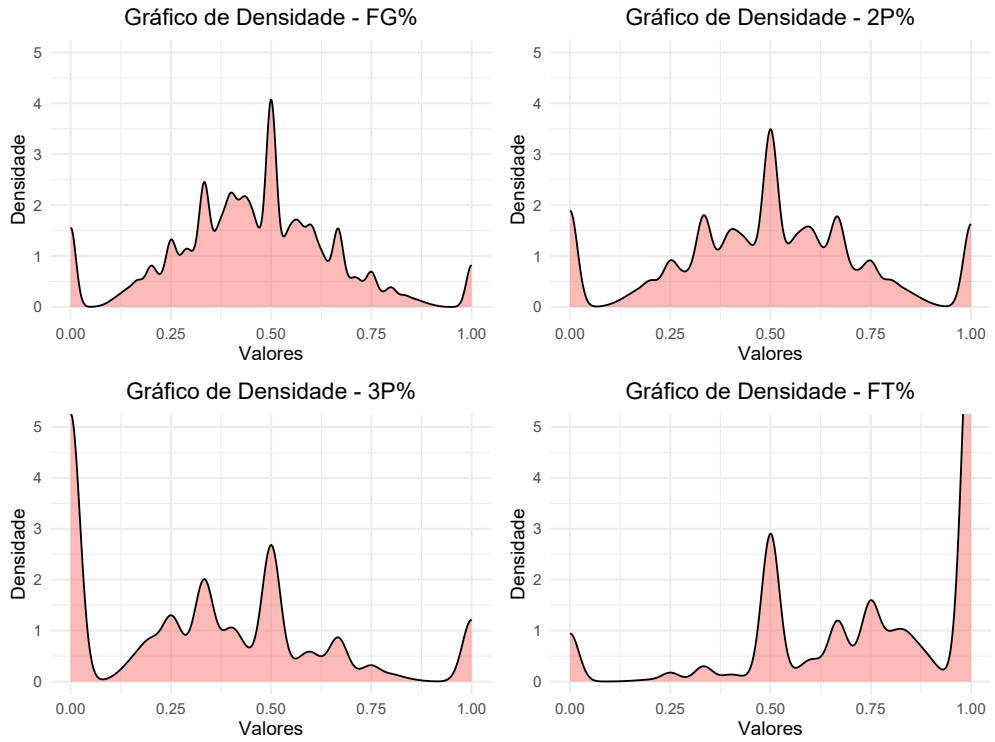

Figura 3.2: Densidades das métricas de conversão de arremessos.

Uma outra maneira de analisar as distribuições é pela Figura 3.2. Pelas estatísticas analisadas, vemos um comportamento até que esperado. As porcentagens de acertos de arremessos (em geral) e de arremessos de dois pontos (os mais comuns) apresentam uma distribuição sem concentrações, com medianas próximas de 0.5. Já para os lances livres e arremessos de três pontos, que são situações que acontecem com menos frequência (e são mais difíceis), temos vários picos na distribuição, evidenciando as concentrações.

Algumas das variáveis categóricas presentes no conjunto de dados possuem níveis que apresentam características bem distintas e interessantes do jogo de basquete. Por exemplo, a separação entre posições de atuação dos jogadores.

Cada equipe atua com cinco jogadores em quadra e, embora as principais características das posições tenha mudado ao longo do tempo, os jogadores são escolhidos com o intuito de montar um quinteto completo, que abranja todas as habilidades físicas, técnicas e táticas ou que tenha o maior enfoque possível em um ponto específico, como

montar um time com ótimos arremessadores. Sendo assim, serão analisadas algumas das métricas já mencionadas, buscando entender seu comportamento em cada um dos níveis da variável Posição.

Tabela 3.4: Médias das métricas avançadas, por posição.

Posição	*Game Score*	BPM	*Plus Minus*	PER	eFG	TSA
C	7.98	-1.12	0.33	15.75	0.51	8.04
C-F	9.50	0.04	-0.04	17.30	0.51	9.75
F	7.77	-1.07	-0.04	13.97	0.49	9.49
F-C	8.96	-0.40	0.53	15.97	0.50	9.91
F-G	10.10	0.71	1.06	15.53	0.50	11.85
G	8.51	-0.75	0.06	14.16	0.47	10.59
G-F	8.34	-0.76	0.28	13.57	0.49	10.98

A Tabela 3.4 apresenta as médias de métricas avançadas categorizadas por posição. A posição C-F (Pivô/Ala) obteve o maior Game Score (9.50) e PER (17.30), indicando um alto nível de eficiência e contribuição geral. Em contraste, a posição F (Ala) apresentou o menor Game Score (7.77) e PER (13.97), sugerindo menor impacto no desempenho individual. A posição F-G (Ala/Armador) se destacou com o maior BPM (0.71) e Plus Minus (1.06), demonstrando um impacto positivo significativo na performance coletiva, enquanto a posição C registrou um dos piores desempenhos em BPM (-1.12), evidenciando menor contribuição.

As posições de armadores (G e G-F) exibiram os maiores volumes de tentativas de arremesso (TSA), mas não foram as mais eficientes, com PER abaixo da média. De maneira geral, os jogadores nas posições híbridas (C-F, F-C e F-G) apresentaram melhor desempenho em métricas de eficiência e impacto, como Game Score, PER e Plus Minus, sugerindo maior versatilidade e contribuição positiva para seus times.

Tabela 3.5: Tabela de contagem para as posições.

Posição	Contagem
C	52364
C-F	25650
F	138271
F-C	51638
F-G	26293
G	185569
G-F	48407

Para entender a distribuição das posições, a Tabela 3.5 apresenta as contagens para cada uma das posições consideradas. As maiores prevalências estão nas posições F e G (*Forward* e *Guard*, respectivamente) que são mais genéricas. As posições de jogadores *combo*, ou seja, que desempenham mais de uma função em quadra podem ser diferenciais interessantes na predição do desempenho.

Tabela 3.6: Tabela de contagem para o mando de jogo.

Mando	Contagem
Casa	263407
Fora	264785

A Tabela 3.6 apresenta a contagem de jogos com base no mando. O número de jogos realizados fora de casa (264.785) é ligeiramente superior ao número de jogos em casa (263.407), com uma diferença de 1.378 jogos. Esse balanço sugere uma distribuição quase equitativa entre os jogos realizados em ambas as condições, indicando que não há um desequilíbrio significativo no número de jogos em função do mando.

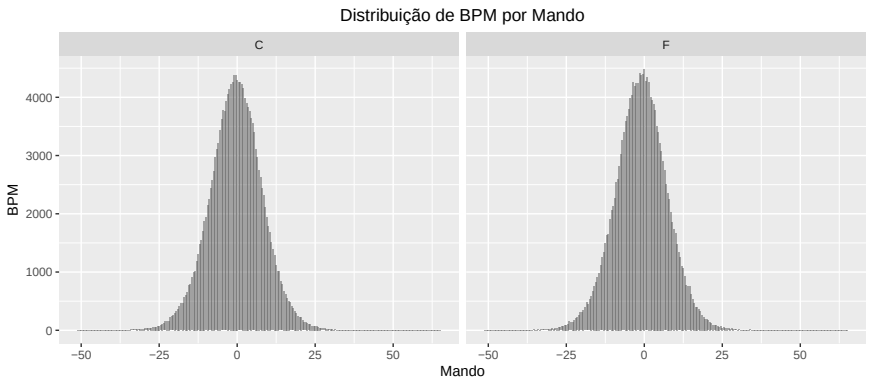

Figura 3.3: Distribuição do BPM dos jogadores em relação ao mando.

O gráfico apresentado na Figura 3.3 exibe duas distribuições de `BPM` (*Box Plus Minus*) separadas pelo mando, comparando jogos em casa (C) e fora (F). Ambas as distribuições têm um formato simétrico e seguem uma curva aproximadamente normal, com os valores de BPM concentrados em torno de zero, indicando que a maioria dos jogadores contribui de forma equilibrada entre os jogos em casa e fora. Esta observação é interessante, visto que há uma impressão geral de que os jogadores (e seu time) têm costumeiramente um desempenho melhor jogando em casa.

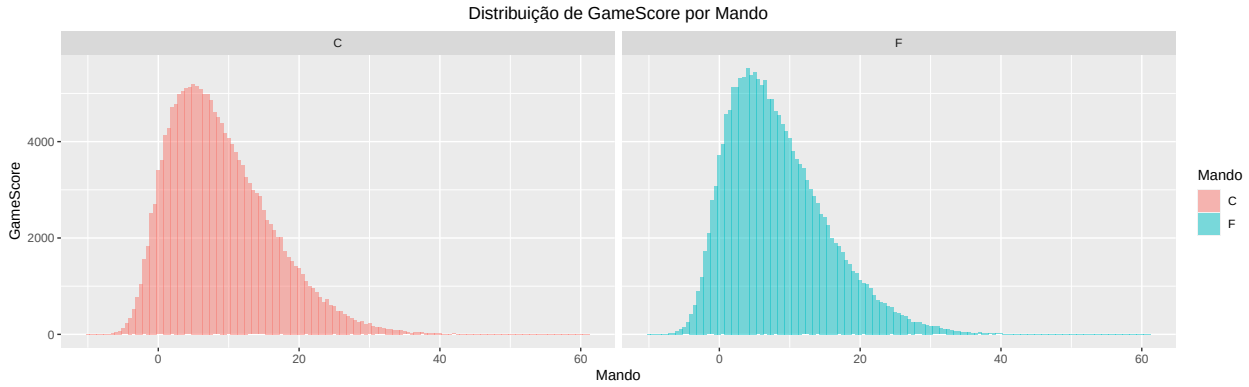

Figura 3.4: Distribuição do `Game Score` dos jogadores em relação ao mando.

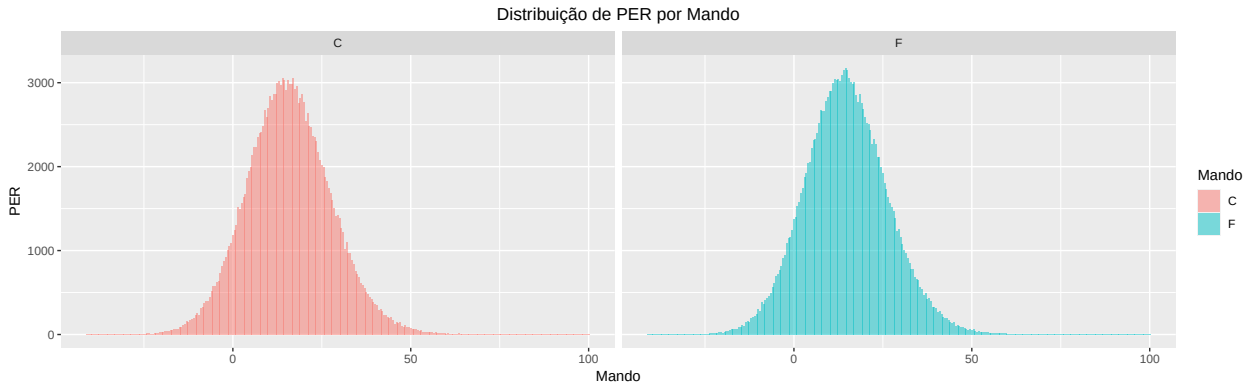

Figura 3.5: Distribuição do `PER` dos jogadores em relação ao mando.

Complementando o observado pelo `BPM`, nas Figura 3.4 e Figura 3.5 as métricas `Game Score` e `PER` não apresenta considerável diferença entre o comportamento dos jogadores considerando o mando do jogo.

Tabela 3.7: Tabela de médias de algumas métricas, por temporada.

Temporada	Média `Pontos`	Média `BPM`	Média `PER`
2008_2009	11.05	-0.79	14.20
2009_2010	11.17	-0.70	14.30
2010_2011	10.94	-0.78	14.23
2011_2012	10.74	-0.74	14.18
2012_2013	10.20	-0.76	13.65
2013_2014	10.57	-0.69	14.21
2014_2015	10.71	-0.68	14.47
2015_2016	10.45	-0.64	14.38
2016_2017	10.79	-0.67	15.17
2017_2018	10.99	-0.63	15.51
2018_2019	11.24	-0.54	16.13
2019_2020	11.53	-0.64	16.56
2020_2021	11.75	-0.61	16.88
2021_2022	11.62	-0.60	16.81
2022_2023	12.06	-0.64	17.37

Pela Tabela 3.7 podemos observar o comportamento de algumas variáveis ao longo das temporadas em estudo, analisando suas médias. Há um leve aumento, nas médias de pontos, `BPM` e `PER` com o decorrer dos anos. Porém, o aumento é gradual, indicando que não existe qualquer recorte de tempo que apresente algum evento único e que possa afetar a análise.

Além das possíveis variáveis preditoras, é de sumo interesse analisar o comportamento e distribuição do índice de desempenho proposto.

Tabela 3.8: Medidas resumo da variável resposta.

Mínimo	**1º Quartil**	**Mediana**	**Média**	**3º Quartil**	**Máximo**
-45.0195	-5.5340	-0.2120	-0.2309	5.0920	61.0915

Analisando as medidas resumo, apresentadas na Tabela 3.8, notamos uma grande amplitude, com valor mínimo de -45.0195 e máximo de 61.0915. Porém, a amplitude interquartil é bem menor, de aproximadamente 10 pontos, o que indica a possível presença de valores outliers nas caudas. Como a resposta é o `BPM` multiplicado por um fator, percebemos que essa alteração e a amostragem de um grupo específico de jogadores não causou

uma drástica modificação na média, permitindo que as interpretações dos resultados se mantenham similares.

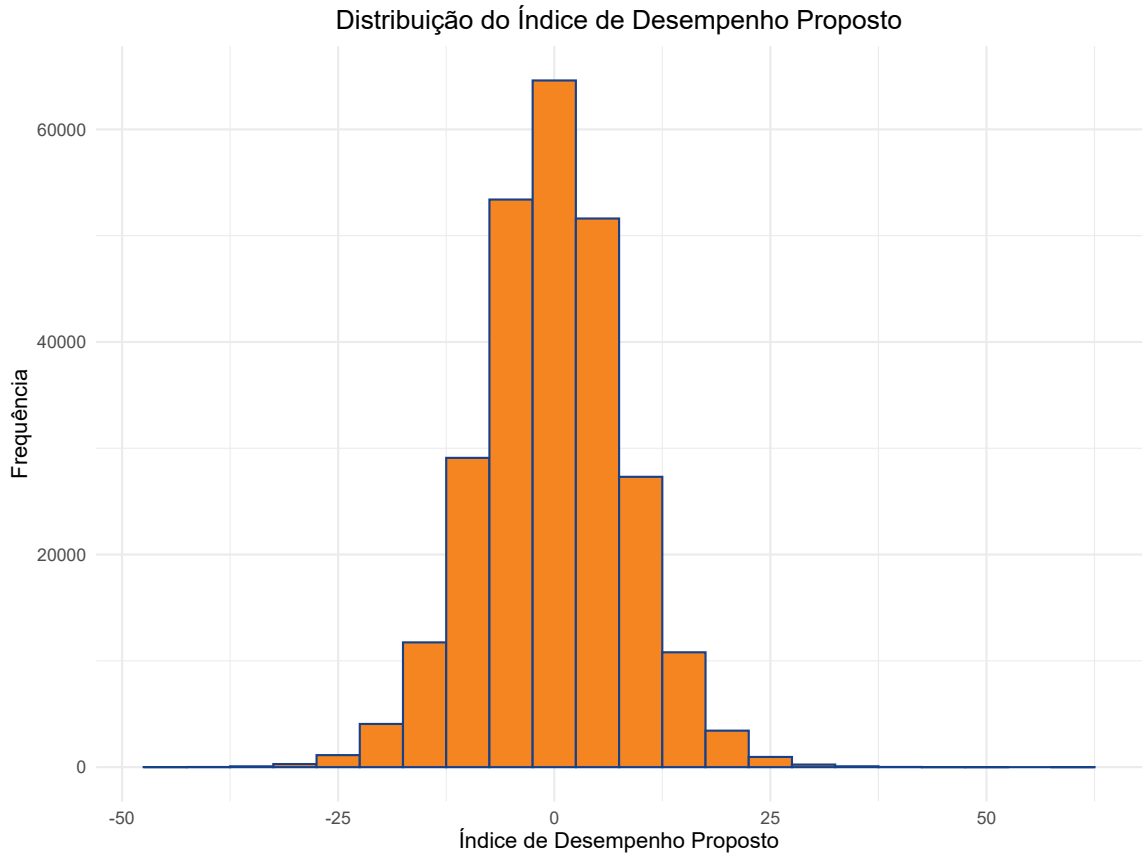

Figura 3.6: Distribuição do Índice de Desempenho Proposto.

Para complementar a análise, na Figura 3.6 podemos visualizar um histograma da variável de interesse, apresentando valores concentrados próximos da média (em torno de 0) e menores na cauda. Porém, é possível notar que a cauda da direita é um pouco mais pesada, indicando que há a possibilidade da distribuição não ser uma Normal.

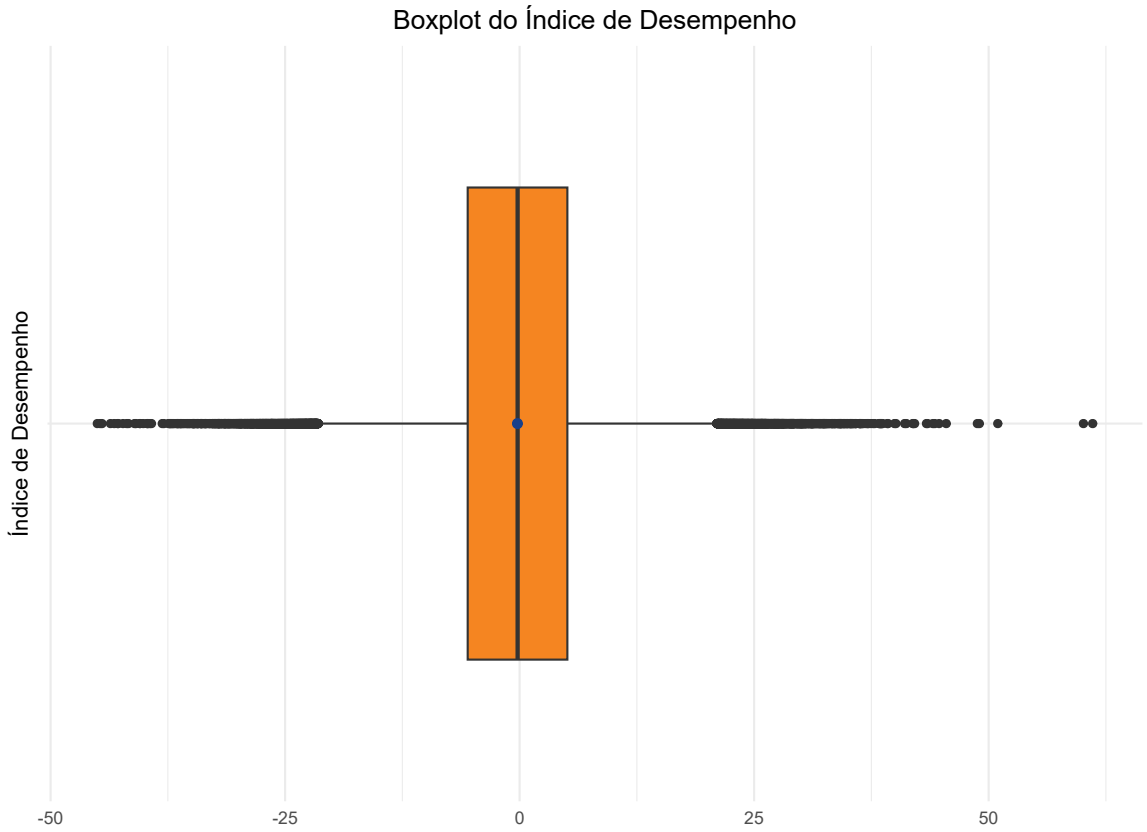

Figura 3.7: Boxplot do Índice de Desempenho Proposto.

Já pela Figura 3.7 podemos notar a variabilidade presente na variável de interesse, além da concentração de valores outliers em ambas as caudas, o que serve de ponto de atenção para o momento de modelagem.

3.2 Modelagens e Predições

Levando em consideração os *insights* obtidos na análise descritiva, os próximos passos para a etapa de predição são de ajustar os modelos considerando a variável de interesse, o índice proposto para medir o desempenho do jogador, e as covariáveis. Inicialmente, seguindo alguns passos importantes de *feature selection* foram excluídas *a priori*, por conta da sua correlação com a variável resposta ou por conta de serem variáveis categóricas com muitos níveis.

Além disso, foi testado um processo automática de seleção pelo pacote Boruta (Kursa e Rudnicki, 2010), que inicialmente apontou pela retirada da variável "Flag Mando" por não ter importância suficiente.

O primeiro modelo testado foi um MLG, que se caracteriza por sua flexibilidade na modelagem de relações não lineares entre variáveis independentes e a variável dependente. O modelo permite a combinação de uma função de ligação que relaciona a média da variável resposta a uma função linear dos preditores com uma distribuição da família exponencial para modelar a variabilidade dos dados. Essa abordagem é particularmente útil para analisar dados que não seguem as suposições tradicionais de normalidade e homocedasticidade.

Devido a característica da variável resposta, ser contínua com domínio em valores negativos, foi escolhida uma função de ligação identidade. Após o primeiro ajuste, foi conduzida uma análise de colinearidade, visto que havia uma possibilidade de *overfitting*, comprovada ao realizar alguns testes.

Para a validação dos primeiros modelos que foram ajustados, as etapas consistiram de análise de colinearidade entre as covariáveis, usando as funções `alias` e `VIF` no `R`, além de calcular a correlação entre as variáveis independentes, em busca de retirar aquelas com as maiores correlações absolutas entre si. O passo seguinte foi, após ajuste de um modelo que convergia nas iterações estipuladas e que apresentava resultados minimamente razoáveis, ajustar um modelo utilizando validação cruzada, no conjunto de variáveis que foi obtido após todas as análises descritas.

A validação cruzada é uma técnica robusta para avaliar o desempenho de modelos preditivos, sendo a validação cruzada *k-fold* sua técnica mais comum e a que foi utilizada no processo de modelagem. Nesta, os dados são divididos em k subconjuntos (*folds*) de tamanho aproximadamente igual. O modelo é treinado k vezes, cada vez utilizando $k-1$

folds para treinamento e o *fold* restante para validação. Isso resulta em k estimativas de desempenho, que são então agregadas para obter uma avaliação geral do modelo. A validação cruzada ajuda a mitigar o risco de *overfitting* (que estava acontecendo nos passos iniciais de modelagem, considerando o *pool* completo de variáveis), proporcionando uma avaliação mais precisa do poder preditivo do modelo e aumentando a confiabilidade das métricas de desempenho, como o Erro Quadrático Médio (MSE) e o Coeficiente de Determinação (R^2).

Tabela 3.9: Estimativas do modelo GLM por validação cruzada.

Covariável	Estimativa	Desvio Padrão	Estatística t	P(>|t|)
Intercepto	-11.5511953	0.0513570	-224.919	$<2.10^{-16}$
3PA	0.1295224	0.0029608	43.745	$<2.10^{-16}$
3P%	2.0655518	0.0264557	78.076	$<2.10^{-16}$
FT%	-0.2786279	0.0176066	-15.825	$<2.10^{-16}$
ORB	-0.2295344	0.0050495	-45.457	$<2.10^{-16}$
DRB	0.0965292	0.0026314	36.684	$<2.10^{-16}$
AST	0.2892024	0.0028474	101.567	$<2.10^{-16}$
BLK	0.7020433	0.0073120	96.013	$<2.10^{-16}$
PF	-0.1736304	0.0044166	-39.313	$<2.10^{-16}$
Posição: C-F	0.2711172	0.0320126	8.469	$<2.10^{-16}$
Posição: F	1.2323396	0.0257190	47.916	$<2.10^{-16}$
Posição: F-C	0.7924265	0.0281777	28.122	$<2.10^{-16}$
Posição: F-G	1.5690839	0.0345337	45.436	$<2.10^{-16}$
Posição: G	0.8054117	0.0283918	28.368	$<2.10^{-16}$
Posição: G-F	1.4026001	0.0316184	44.360	$<2.10^{-16}$
IdadeAnos	0.0381242	0.0014223	26.805	$<2.10^{-16}$
PER	0.6334441	0.0010006	633.063	$<2.10^{-16}$
eFG	1.1363657	0.0419626	27.080	$<2.10^{-16}$
TSA	-0.2209775	0.0014948	-147.829	$<2.10^{-16}$
FTR	0.0961110	0.0161488	5.952	$<2,6.10^{-10}$
TOV%	-0.0293903	0.0005353	-54.903	$<2.10^{-16}$

Como apresentado na Tabela 3.9 as variáveis explicativas no modelo final são: `3PA`, `3P%`, `FT%`, `ORB`, `DRB`, `AST`, `BLK`, `PF`, `Posição`, `Idade (Anos)`, `PER`, `eFG`, `TSA`, `FTR` e `TOV%`.

As covariáveis que entraram no modelo apresentaram todas um considerável nível de importância. Especialmente, para aquelas que trazem um prejuízo para o desempenho do atleta estão `PF` (faltas cometidas), `TSA` (métrica avançada para medir quantidade de

arremessos) e `TOV%` (porcentagem de erros). Já aquelas com importância mais significante em melhorar o desempenho estão `3P%` (porcentagem de arremessos de 3), `eFG` (métrica avançada de eficiência de arremesso) e algumas posições, predominantemente a que envolve jogadores *Forward*, ou Alas.

Tabela 3.10: Métricas de desempenho do modelo.

Métricas		
REQM	R^2	EMA
3.022419	0.8676759	2.365619

Na Tabela 3.9 temos algumas métricas de desempenho de ajuste do modelo para o conjunto de treinamento, sendo que o coeficiente de determinação (Draper e Smith, 1998) R^2 de $0,8677$ indica que aproximadamente $86,77\%$ da variabilidade da variável resposta é explicada pelo modelo. Esse valor sugere um bom ajuste do modelo aos dados, indicando que os preditores incluídos capturam uma proporção significativa da variância da variável dependente. Já a Raiz do Erro Quadrático Médio (REQM) de 3.0224 e o Erro Médio Absoluto (EMA) (Armstrong, 2001) de 2.3656 fornecem informações sobre a magnitude dos erros de predição, com valores baixos e que indicam um bom ajuste.

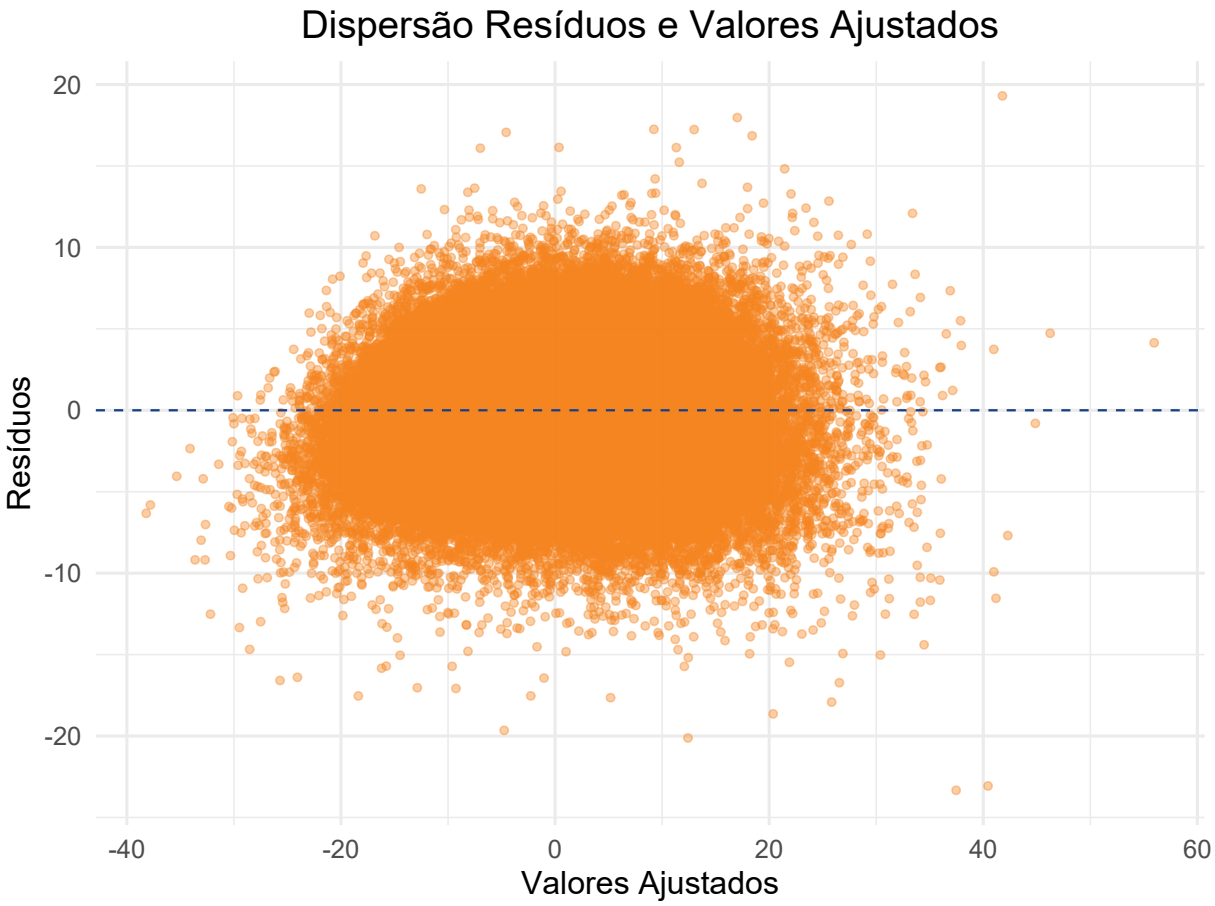

Figura 3.8: Dispersão dos resíduos do modelo.

Para refinar a análise do ajuste do modelo, na Figura 3.8 podemos observar a distribuição dos resíduos pelos valores ajustados do modelo. Não há evidência de um claro padrão, com apenas um leve aumento na dispersão nos valores extremos. Como destaques mais claros, é possível notar alguns valores *outliers*.

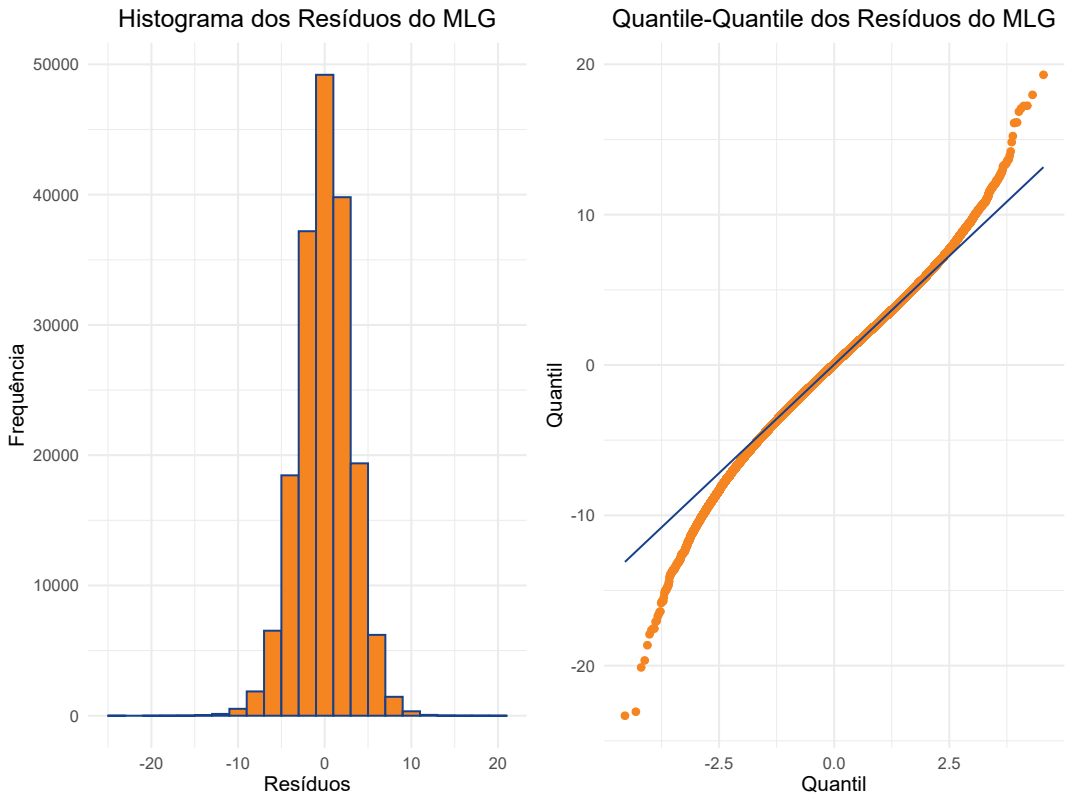

Figura 3.9: Histograma e QQ-Plot dos resíduos do MLG.

Na Figura 3.9 podemos notar que as suposições de normalidade dos resíduos parece satisfeita, pela forma da distribuição vista no histograma e do comportamento do gráfico QQ-Plot, com apenas dispersões nas caudas.

Porém, para trazer uma informação complementar a análise gráfica, foi realizado um teste Shapiro-Wilk de normalidade. Neste, o valor-p do teste foi de $2x10^{-}10$, o que traz indícios fortes contra a suposição de normalidade. Com isto em consideração, também serão testadas outras distribuições para o MLG.

Como uma outra alternativa para verificação da qualidade das predições do modelo, a última temporada disponível durante a captura dos dados, referente ao período de 2022 até 2023 foi separado do restante do conjunto de desenvolvimento e teste, para servir como uma partição OOT (*Out Of Time*), especificamente para capturar o objetivo dos modelos, que é prever adequadamente o desempenho nos jogos futuros.

Tabela 3.11: Métricas de desempenho das predições do MLG na temporada 2022-2023.

Métricas		
REQM	R^2	EMA
11,2224	0,8269	8,8268

Pela Tabela 3.11 podemos notar que no período mais recente a qualidade das predições do MLG são piores do que no período de desenvolvimento, apresentando indícios de que o modelo não conseguiu capturar as informações dos dados da melhor maneira possível.

Como uma nova alternativa para o modelo MLG apresentado, foi aplicada uma transformação sigmoide na variável resposta, para que seja exclusivamente positiva.

$$Y^* = \frac{1}{1 + exp(Y)},$$

sendo Y^* a variável resposta Y transformada.

Tabela 3.12: Métricas de desempenho das predições do MLG, distribuição Gama.

Métricas		
REQM	R^2	EMA
0,486	0,228	0,344

Pela Tabela 3.12 são apresentadas as métricas para o conjunto de treinamento do modelo, utilizando uma função de ligação inversa para a distribuição Gama. Embora as métricas de performance preditiva tenham melhorado, o R^2 apresenta valor muito baixo, indicando piora do modelo em capturar a variabilidade dos dados.

De maneira análoga, para a variável resposta transformada, foi ajustado um modelo com uma distribuição Poisson e função de ligação log.

Tabela 3.13: Métricas de desempenho das predições do MLG, distribuição Poisson.

Métricas		
REQM	R^2	EMA
0,359	0,426	0,272

Pela Tabela 3.13 são apresentadas as métricas para o conjunto de treinamento do

modelo. Embora as métricas de performance preditiva tenham melhorado, o R^2 apresenta valor muito baixo, indicando piora do modelo em capturar a variabilidade dos dados.

O principal problema ao aplicar uma transformação na variável resposta é sua perda de interpretabilidade, visto que sua construção é baseada na boa definição de interpretação do `BPM`.

Uma outra opção para buscar as melhores predições é ajustar um novo modelo aos dados, com uma técnica diferente. Sendo assim, será ajustado um modelo de árvore de decisão, especificamente um modelo de árvore de decisão para regressão.

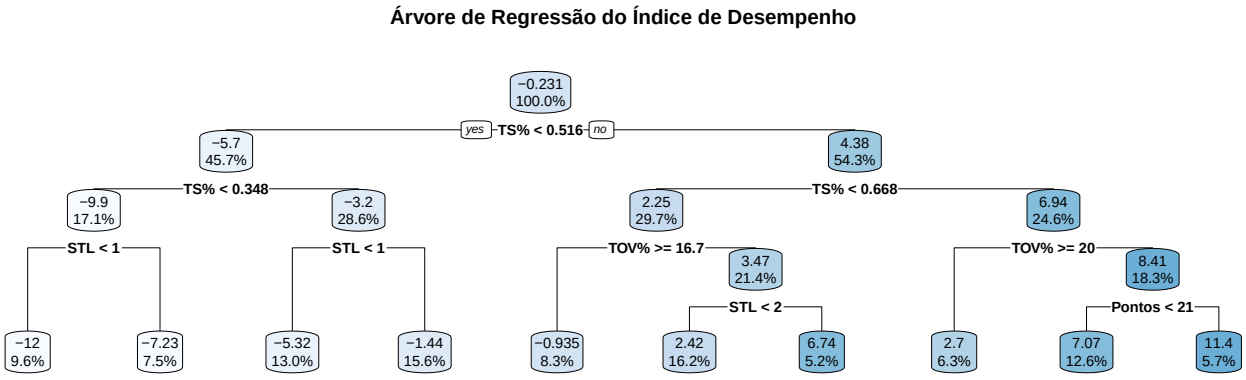

Figura 3.10: Nós e folhas do modelo de árvore de decisão.

Pela Figura 3.10 temos as decisões (nós) tomados pelo modelo de árvore. A primeira divisão ocorre com base no critério `TS%` < 0.516. Ou seja, para valores menores que 0.516 a árvore se ramifica à esquerda, para valores maiores que o corte, ramifica-se à direita. Em cada um dos nós, teremos duas informações: o valor predito, no topo da caixa azulada, e a quantidade de observações que está no nó específico, logo abaixo.

Assim, sucessivamente, são definidas as variáveis e seus pontos de corte em cada nó, sendo as 'decisoras': `TS%`, `TOV%` , `STL` e `Pontos` as preditoras responsáveis pelas ramificações nos nós da árvore.

Os nós finais (ou folhas) da árvore indicam diferentes valores finais da variável resposta com suas respectivas probabilidades (ou porcentagens de amostra). Valores positivos nas

folhas indicam desempenhos melhores para os jogadores, sendo que valores negativos indicam piora no desempenho. As caixas em azul mais escuro apresentam, gradativamente, os nós com melhor desempenho.

Podemos notar que, por exemplo, $5,7\%$ do conjunto de dados caiu na folha que representa os valores preditos mais altos que são equivalentes a $11,4$, que é um valor muito elevado, considerando que a média da base está próxima de zero. Para essa folha em específico, o modelo separou jogadores com *True Shooting* acima de $66,8\%$, com índice de turnover abaixo de 20 e mais de 21 pontos na partida.

Em contraste, para a folha que representa o pior desempenho, o *True Shooting* está abaixo de $34,8\%$ e o jogador, além do desempenho horrível arremessando, não conseguiu nenhum roubo de bola, uma variável importante de medição de desempenho defensivo para certas posições.

Em geral, por aparecer em diversos nós, temos que valores mais baixos de `eFG` tendem a estar associados a desempenhos negativos, enquanto que valores mais altos estão associados a melhores desempenhos.

Uma outra análise interessante é sobre o conjunto de variáveis que serviu como ponto de corte nos nós. Mesmo se a performance preditiva do modelo não for ideal, esse conjunto de variáveis pode servir como um foco inicial em estudos subsequentes.

Tabela 3.14: Métricas de desempenho das predições da Árvore de Regressão, conjunto de treinamento.

Métricas		
REQM	R^2	EMA
5,465503	0,5673204	4,248744

Pela Tabela 3.14 podemos notar pertinentemente que as predições do modelo de Árvore de Regressão não são melhores que o MLG, também apresentando um coeficiente de determinação R^2 bem abaixo, o que indica uma menor porcentagem da variabilidade dos

dados sendo explicada pelo modelo.

Seguindo o mesmo teste feito anteriormente, o modelo será ajustado para os dados referentes a temporada mais recente presente no conjunto de dados, referente ao período de 2022-2023, com o intuito de analisar as predições nesse período.

Tabela 3.15: Métricas de desempenho das predições da Árvore de Regressão, para a temporada de 2022-2023.

Métricas		
REQM	R^2	EMA
5,552201	0,5528062	4,36686

Analisando as predições na última temporada presente no conjunto de dados, que foi separada especificamente para as predições, a Tabela 3.15 apresenta os resultados encontrados para as métricas de precisão das predições. Como ponto positivo, as predições se mantém com a mesma qualidade nas duas partições, embora os resultados ainda sejam inferiores em relação a todos os MLG's que foram ajustados.

Capítulo 4

Comentários Finais

Neste estudo, foram desenvolvidos e comparados dois modelos estatísticos com o intuito de prever o desempenho dos jogadores da NBA a cada partida, sendo o desempenho medido por um índice específico que foi criado como uma variação de uma métrica avançada bastante utilizada considerando complementarmente o resultado de cada partida analisada. O primeiro modelo foi um Modelo Linear Generalizado (MLG) e o segundo uma Árvore de Regressão.

O MLG foi ajustado com distribuição Normal e uma função de ligação identidade, apresentando métricas de ajuste como REQM, R^2 e EMA com valores satisfatórios para considerar um bom ajuste do modelo. Porém, o teste de Shapiro-Wilk apresentou evidencias contrárias a suposição de normalidade, indicando um possível problema no ajuste. Os modelos ajustados utilizando as distribuições Gama e Poisson apresentaram métricas boas para predição, porém o ajuste medido pelo coeficiente R^2 é bem insatisfatório.

Por outro lado, com o modelo Árvore de Regressão, esperamos que seja capaz de capturar relações mais complexas entre a resposta e os preditores. Com isso, esperamos que a captura dessas relações ajude a apresentar predições mais precisas. Como foi realizada uma *feature selection* previamente ao ajuste dos dois modelos, esperamos não encontrar

problemas de *overfitting* para a árvore.

A fim, as métricas de qualidade das predições apontaram que o MLG apresentou uma performance superior ao modelo de Árvore de Regressão. Ainda assim, ambos modelos não apresentaram resultados tão bons na partição específica para previsão, que foi a temporada mais recente presente no conjunto de dados. Sendo assim, os modelos não se demonstraram perfeitos para o objetivo de predição por jogo, embora os resultados obtidos são promissores como um ponto inicial para modelos mais refinados.

Como uma possibilidade de abordagem futura, além da expansão do conjunto de treinamento para todos (ou uma amostra maior) os jogadores e outras métricas que podem agregar na predição, pode ser de interesse testar diferentes técnicas de modelagem. Ademais, a escolha do índice que será utilizado para mensurar o desempenho já é um caso especial, visto que em diversos contextos de predição esportiva (BBall, 2024) muitos analistas despendem considerável tempo e recurso em busca de gerar um índice novo, com o intuito de apresentar e medir desempenho de maneira inovadora.

Referências Bibliográficas

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. John Wiley And Sons.

Alves, C. C. (2011). Basketball court. https://esportecastroalves.blogspot.com/2011/05/quadra-de-basquete.html?m=1. Accessed: 2024-08-28.

Armstrong, J. S. (2001). *Principles of Forecasting: A Handbook for Researchers and Practitioner*. Springer.

BBall, I. (2024). Lebron: The man, the myth, the metric? https://www.bball-index.com/lebron-introduction/. Accessed: 2024-08-28.

Brasil, A. O. (2016). Posição jogadores. https://alley-oopbrasil.blogspot.com/2016/11/as-principais-posicoes-do-basquete.html?m=1. Accessed: 2024-08-28.

BREIMAN, L., FRIEDMAN, J., OLSHEN, R. e STONE, C. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.

Casella, G. e Berger, R. L. (2002). *Statistical Inference (2nd ed.)*. Duxbury.

Choi, K. M. (2017). Medical cost personal datasets. https://www.kaggle.com/mirichoi0218/insurance. Accessed: 2024-08-28.

CLARKE, B., FOKOUÉ, E. e ZHANG, H. (2009). *Principles and Theory for Data Mining and Machine Learning*. Springer.

DINIZ, C. A. R. e LOUZADA NETO, F. (2012). *Modelagem Estatística Para Risco de Crédito*. 20º SINAPE - Simpósio Nacional de Probabilidade e Estatística.

Dobson, A. J. e Barnett, A. G. (2008). *An Introduction to Generalized Linear Models*. CRC Press.

Draper, N. R. e Smith, H. (1998). *Applied Regression Analysis*. Wiley.

Haefner, J. (2021). What is effective field goal percentage? https://www.breakthroughbasketball.com/stats/effective-field-goal-percentage.html. Accessed: 2024-09-12.

Head, S. (2024). Stat head. https://www.stathead.com. Accessed: 2024-08-28.

HOLLINGER, J. (2004). *Pro Basketball Prospectus*. Potomac Books, Incorporated.

IZBICKI, R. e SANTOS, T. M. (2020). *Aprendizado de Máquina: uma abordagem estatística*.

JESUS, M. F. N. d. (2015). Estudo comparativo entre as funções de ligação logit e probit : estimando parâmetros. *Monografia (Bacharelado em Estatística) - Departamento de Estatística e Ciências Atuarias, Centro de Ciências Exatas e Tecnologia, Universidade Federal de Sergipe*.

Kursa, M. B. e Rudnicki, W. R. (2010). Feature selection with the boruta package. *Journal of Statistical Software, 36(11), 1–13. https://doi.org/10.18637/jss.v036.i11*.

Learn, S. (2024). Cross-validation: evaluating estimator performance. https://

scikit-learn.org/stable/modules/cross_validation.html. Accessed: 2024-08-28.

McCULLAGH, P. e NELDER, J. (1989). *Generalized linear models*. 2nd Edition London: Chapman & Hall.

Memorial, N. (2024). Nba hall of fame. https://www.hoophall.com/. Accessed: 2024-08-28.

Midfield, T. (2024). Who is john hollinger. https://themidfield.com/nba-betting/john-hollinger-nba-metrics/. Accessed: 2024-09-12.

NBA (2022). Curry record breaking 3. https://www.nba.com/news/stephen-curry-tracker-all-time-3s-record. Accessed: 2024-08-28.

NELDER, J. A. e WEDDERBURN, R. W. M. (1972). *Generalized Linear Models*. Wiley.

of Baskteball, L. (2024). 2-0 playoff lead. https://www.landofbasketball.com/statistics/playoff_series_2_0.htm. Accessed: 2024-08-28.

QUINLAN, J. (1986). *Introduction of Decision Trees*. Kluwer Academic Publishers.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Reference, B. (2024a). Basketball reference. https://www.basketball-reference.com. Accessed: 2024-08-28.

Reference, B. (2024b). First baa game. https://www.basketball-reference.com/boxscores/194611010TRH.html. Accessed: 2024-08-28.

Reference, B. (2024c). Basketball reference glossary. https://www.basketball-reference.com/about/glossary.html. Accessed: 2024-09-12.

SEGDWICK, P. (2015). A comparison of parametric and non-parametric statistical tests. *BMJ 2015;350:h2053*.

STROBL, C., MALLEY, J. e TUTZ, G. (2009). *An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests*. Psychol Methods.

T Hastie, R Tibshirani, J. F. (2001). *Elements of Statistical Learning*. Springer.

THERNEAU, T. M. e ATKINSON, E. J. (2019). *An Introduction to Recursive Partitioning Using the RPART Routines*.

WCBV5 (2024). Celtics titles. https://www.wcvb.com/article/boston-celtics-nba-championship-history/60973137. Accessed: 2024-08-28.

Apêndice A

Códigos

Os códigos utilizados para desenvolvimento do estudo podem ser encontrados em: https://drive.google.com/drive/folders/1dSSH0sgqcRMLRZDwbCLmVlXeQ5G7RBmS?usp=drive_link. Caso não esteja mais disponível, pode ser solicitado pelo e-email: albertotjr@gmail.com.