INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Previsão de Resultados da NBA: Da Análise das Dinâmicas da Liga à Implementação de Modelos Preditivos

FÁBIO DANIEL CERQUEIRA PIRES
Setembro de 2024

POLITÉCNICO
DO PORTO

ISEP INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO

Previsão de Resultados da NBA

Da Análise das Dinâmicas da Liga à Implementação de Modelos Preditivos

Fábio Daniel Cerqueira Pires

Dissertação para obtenção do Grau de Mestre em Engenharia Informática, Área de Especialização em Sistemas de Informação e Conhecimento

Orientador: Doutora Fátima Rodrigues

Júri:
Presidente:
Doutor Bruno da Silva

Vogais:
Doutora Isabel Praça

Porto, 4 de setembro de 2024

Declaração de Integridade

Declaro ter conduzido este trabalho académico com integridade.

Não plagiei ou apliquei qualquer forma de uso indevido de informações ou falsificação de resultados ao longo do processo que levou à sua elaboração.

Portanto, o trabalho apresentado neste documento é original e de minha autoria, não tendo sido utilizado anteriormente para nenhum outro fim.

Declaro ainda que tenho pleno conhecimento do Código de Conduta Ética do P.PORTO.

ISEP, Porto, 4 de setembro de 2024

Resumo

A popularidade global da NBA e o crescimento exponencial das apostas desportivas em Portugal destacam a necessidade de modelos preditivos nesta área. Utilizando a metodologia CRISP-DM, esta tese foca-se no desenvolvimento de um modelo eficaz para prever resultados de jogos da NBA. O estudo identifica padrões cruciais para previsões precisas, analisando dados históricos, estatísticas de jogadores e resultados de jogos.

As seis fases da metodologia incluem a compreensão dos objetivos do negócio, a exploração de dados, a preparação meticulosa dos dados, a seleção do modelo, a avaliação rigorosa do mesmo e a implementação final num site acessível aos utilizadores interessados. Na fase de compreensão dos objetivos do negócio, foram definidos os requisitos e as metas do projeto. Durante a exploração de dados, foram analisados e compreendidos os dados disponíveis. A preparação dos dados envolveu a limpeza e transformação dos mesmos para garantir a sua qualidade. Na fase de seleção do modelo, foram treinados diversos modelos, recorrendo a vários algoritmos, com o objetivo de obter o melhor desempenho possível. A avaliação dos modelos foi feita com base em várias métricas, sendo que o modelo escolhido atingiu uma taxa de acerto de 64,4% e F1=72,4%. Finalmente, o modelo foi implementado num website de fácil utilização, através da *framework Streamlit* .

A implementação num website aborda a atual falta de ferramentas eficazes de apoio à decisão para prever jogos da NBA, contribuindo para a evolução do cenário de apostas desportivas em Portugal.

Palavras-chave: NBA, Apostas desportivas, Machine Learning, CRISP-DM, Classificação

Abstract

The NBA's global popularity and the exponential growth of sports betting in Portugal highlight the need for predictive models in this area. Using the CRISP-DM methodology, this thesis focuses on developing an effective model for forecasting NBA game outcomes. The study identifies patterns critical for precise predictions by analyzing historical data, player statistics, and game results.

The six phases of the methodology include understanding business goals, data exploration, meticulous data preparation, model selection, its rigorous evaluation, and final deployment on an accessible website for the interested users. In the business objectives understanding phase, the project's requirements and goals were defined. During data exploration, the available data was analyzed and insights about it were obtained. Data preparation involved cleaning and transforming the data to ensure its quality. In the model selection phase, multiple models were trained using various algorithms with the objective of obtaining the best possible performance. The models were evaluated based on multiple metrics, with the chosen model achieving an accuracy rate of 64.4% and F1=72,4%. Finally, the model was implemented on a user-friendly website using the Streamlit framework.

The implementation on a website addresses the current lack of effective decision support tools for NBA game predictions, contributing to Portugal's evolving sports betting landscape.

Keywords: NBA, Sports betting, Machine Learning, CRISP-DM, Classification

Agradecimentos

Gostaria de expressar a minha sincera gratidão a todas as pessoas que, de uma forma ou de outra, me apoiaram e contribuíram para a concretização deste projeto, ajudando-me a superar os desafios ao longo deste percurso.

Em particular, quero agradecer à minha orientadora, Doutora Fátima Rodrigues, pela sua inestimável orientação, paciência e disponibilidade ao longo de todo o processo. O seu conhecimento, os seus conselhos e o seu *feedback* foram essenciais para a conclusão deste trabalho.

Agradeço também aos meus pais, pelo apoio e encorajamento incondicional em todos os momentos. Sem eles, nada disto teria sido possível.

A todos os meus colegas e amigos que, com as suas palavras de incentivo, partilharam comigo esta jornada, deixo igualmente o meu profundo agradecimento.

Finalmente, expresso a minha gratidão a todos os professores e membros da instituição que, de alguma forma, contribuíram para o meu crescimento académico e pessoal durante este percurso.

A todos, o meu muito obrigado!

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

AD	Árvore de Decisão.
ETEI	*Extended Team Efficiency Index*.
GB	*Gradient Boosting*.
GCN	*Graph Convolution Networks*.
GM	*General Manager*.
KNN	*K-Nearest Neighbors*.
LASSO	*Least Absolute Shrinkage and Selection Operator*.
LGBM	*Light Gradient-Boosting Machine*.
ML	*Machine Learning*.
NB	*Naive Bayes*.
NBA	*National Basketball Association*.
PCA	*Principal Component Analysis*.
PEM	Princípio da Entropia Máxima.
RF	*Random Forest*.
RL	Regressão Logística.
RNA	Rede Neuronal Artificial.
SRIJ	Serviço de Regulação e Inspeção de Jogos.
SVM	*Support Vector Machine*.
TAB	Taxa de Acerto Balanceada.
VCH	*Voting Classifier Hard*.
VCS	*Voting Classifier Soft*.
WNBA	Women's National Basketball Association.
XGBoost	*Extreme Gradient Boosting*.

Capítulo 1

Introdução

Neste capítulo, é explorado o contexto e o problema subjacente a esta tese. São definidos os objetivos da tese e é explicada a metodologia de trabalho adotada, além de abordar questões éticas pertinentes.

1.1 Contexto

O *basketball*, em particular a *National Basketball Association* (NBA), é um dos desportos mais populares no mundo, tendo-se tornado numa indústria de milhões de dólares nas últimas décadas. Apesar do cenário das apostas desportivas em Portugal ser relativamente recente, teve início em 2015 com a chegada da *Placard*, tem vindo a crescer exponencialmente. Só no primeiro ano de existência, esta casa de apostas conseguiu alcançar o 3ž lugar no top de vendas dos Jogos Santa Casa, totalizando cerca de 1 milhão de apostadores e 300 milhões em vendas. A NBA esteve entre as competições que registaram maior volume de vendas o que comprova a popularidade do desporto (Lusa 2016).

Passados quase 10 anos, o número de casas de apostas legalizadas em Portugal teve um aumento exponencial. Segundo o Serviço de Regulação e Inspeção de Jogos (SRIJ), entidade responsável por atribuir as licenças, existem atualmente 17 casas de apostas legalizadas em Portugal, sendo que 11 delas possuem a modalidade de apostas desportivas (SRIJ 2023). Entre 2018 e 2022, foram movimentados 4000 milhões de euros em apostas desportivas, segundo o SRIJ (DN/Lusa 2023) .

Este aumento exponencial das apostas desportivas está diretamente relacionado com o forte *marketing* que estas entidades praticam. Um dos exemplos deste forte *marketing* é um anúncio da *Betclic* (Betclic 2021) que inclui Neemias Queta, o primeiro português a jogar na NBA (Lusa 2021). Neste caso específico, a plataforma capitalizou de forma notável o entusiasmo gerado em Portugal pelo feito extraordinário do atleta. Com este anúncio, a *Betclic* não só solidificou a sua presença no mercado, como também estabeleceu uma ligação emocional com os apostadores portugueses. O anúncio, cuidadosamente elaborado, explorou a admiração coletiva e o orgulho patriótico associados à conquista de Neemias, de forma a atrair potenciais novos apostadores. Recentemente, a *Betclic* realizou um novo anúncio (Betclic 2023), desta vez com Ticha Penicheiro, atleta portuguesa incluída no *Hall of Fame* da WNBA (FPB 2019), como protagonista.

Atualmente, diversas fontes online, como *websites* e aplicações móveis, oferecem previsões gratuitas para resultados de jogos de *basketball*. Exemplos incluem o *Vitibet* (*Vitibet* s.d.) e o *Forebet* (*Forebet* s.d.), que consideram a forma das equipas e a força em casa/fora. Contudo, esses critérios podem ser insuficientes para determinar resultados precisos. O

Sportytrader (*SportyTrader* s.d.) sugere apostas com base em palpites de especialistas, indo além da simples previsão de vencedores, podendo incluir, por exemplo, desempenho individual de jogadores. Quanto a aplicações móveis, existe um infindável número de plataformas para este propósito, sendo que os critérios são praticamente os mesmos utilizados nos *websites* mencionados. Esta análise permitiu concluir que existe uma carência de boas soluções de apoio à decisão a nível de apostas em jogos da NBA no mercado, pelo que é essencial haver uma melhoria na forma como se tenta prever o resultado destes eventos.

1.2 Problema

A prática de apostas desportivas, especialmente no cenário altamente dinâmico e imprevisível da NBA, apresenta desafios consideráveis. O cerne deste desafio reside na capacidade de antecipar com precisão os resultados dos jogos, dada a multiplicidade de fatores que influenciam o desempenho das equipas e dos jogadores. Além disso, a coleta de dados de várias fontes, a integração desses dados e a garantia da sua qualidade e atualização constante constituem problemas adicionais. A atual falta de ferramentas confiáveis que consigam tratar e analisar esses dados de forma eficiente para orientar as decisões dos apostadores intensifica ainda mais esse problema. Deste modo, torna-se imperativo desenvolver métodos inovadores e eficazes para a coleta, processamento e análise de dados, a fim de fornecer aos apostadores informações mais precisas e úteis. Esta tese visa desenvolver uma solução eficaz que possa beneficiar os entusiastas de apostas desportivas, especificamente no contexto da NBA.

1.3 Objetivos

O objetivo principal deste projeto é desenvolver um modelo de previsão de resultados de jogos da NBA robusto e disponibilizá-lo de forma fácil aos apostadores através de um *website*.

Para alcançar este objetivo principal, os seguintes objetivos específicos foram definidos:

- Recolher dados detalhados relacionados com jogos realizados em épocas anteriores.
- Analisar o desempenho dos jogadores ao longo das suas carreiras de forma a identificar padrões e tendências que possam contribuir para previsões mais precisas.
- Investigar o impacto das mudanças de equipa no rendimento dos jogadores para uma compreensão mais profunda das dinâmicas da NBA.
- Analisar o desempenho das equipas tanto em casa como fora, para entender as variantes que afetam o seu sucesso em diferentes contextos.
- Associar o número de vitórias de uma equipa com estatísticas de desempenho específicas em casa e fora, proporcionando conhecimentos valiosos sobre os elementos-chave para o sucesso na NBA.
- Realizar uma análise sequencial para explorar a evolução das posições em campo e do jogo na NBA ao longo do tempo, identificando tendências e adaptações estratégicas. Além disso, estender a análise à versatilidade dos jogadores, examinando como as suas habilidades evoluíram e se adaptaram às exigências do jogo.

1.4 Contribuições

Esta tese visa contribuir significativamente para o campo da previsão de resultados desportivos, com especial foco na NBA, através do desenvolvimento de um modelo preditivo robusto. As principais contribuições desta investigação incluem:

- **Desenvolvimento de um modelo preditivo consistente**: Criação de um modelo capaz de prever com precisão os resultados das partidas da NBA, mantendo uma elevada taxa de acerto ao longo das jornadas e garantindo uma boa percentagem de acerto nos dois resultados possíveis de uma partida de NBA.
- **Disponibilização através de uma plataforma web**: Implementação do modelo preditivo numa plataforma *web* acessível aos utilizadores, facilitando a utilização prática das previsões.
- **Análise do desempenho dos jogadores**: Fornecimento de respostas concretas sobre o desempenho dos jogadores ao longo das suas carreiras, identificando padrões e tendências que possam surgir com o tempo.
- **Impacto das mudanças de equipa**: Investigação detalhada sobre o impacto das mudanças de equipa nos jogadores, revelando como estas transições afetam o desempenho individual e coletivo.
- **Desempenho das equipas em diferentes contextos**: Análise do desempenho das equipas em jogos em casa e fora, elucidando os fatores cruciais que influenciam o sucesso na NBA.
- **Evolução das posições e do jogo**: Exploração da evolução das posições em campo e das estratégias de jogo ao longo do tempo, oferecendo conhecimentos valiosos sobre as tendências e adaptações no cenário do *basketball* profissional.
- **Versatilidade dos Jogadores:** Análise da versatilidade dos jogadores ao longo das épocas, contribuindo para a compreensão das mudanças no estilo de jogo e das habilidades específicas demandadas no *basketball* moderno.

Estas contribuições pretendem não só avançar o conhecimento académico na área, mas também proporcionar ferramentas práticas e conhecimentos estratégicos que possam ser aplicados por profissionais do desporto e entusiastas da NBA.

1.5 Ética

É imperativo reconhecer e abordar questões éticas de forma a garantir a integridade, transparência e responsabilidade deste projeto. Em seguida, exploraremos os princípios éticos que regem o seu desenvolvimento, visando assegurar práticas éticas ao longo de toda a sua execução.

Um dos princípios fundamentais deste projeto é a transparência. Com base neste princípio, os utilizadores finais devem estar informados das limitações inerentes ao modelo de previsão de resultados, reconhecendo a complexidade dos jogos da NBA e a influência de inúmeros fatores imprevisíveis. Deve ser claramente comunicado que o modelo não garante uma taxa de acerto de 100%, podendo ocasionalmente falhar no prognóstico de um jogo.

Além disso, há um compromisso inequívoco em promover práticas de jogo responsável. A plataforma onde o modelo será disponibilizado deve não apenas informar sobre os riscos associados às apostas desportivas, mas também fornecer informações abrangentes para permitir que os utilizadores tomem decisões de forma informada. Os utilizadores serão encorajados a apostar de forma responsável, compreendendo que as previsões não garantem lucros.

Consciente de que este projeto pode influenciar as decisões dos apostadores, é reconhecida a possibilidade de impactos financeiros nas casas de apostas.

A recolha de dados para este projeto será conduzida de maneira ética, respeitando rigorosamente a privacidade dos jogadores da NBA. Será utilizado exclusivamente um dataset público, garantindo a conformidade com os mais elevados padrões éticos na obtenção e utilização de informações.

É reconhecida a importância de estar em conformidade com todas as leis e regulamentações locais relacionadas com apostas e jogos de azar.

1.6 Metodologia

Esta tese será desenvolvida recorrendo à metodologia CRISP-DM.

A metodologia CRISP-DM (Wirth e Hipp 2000) pode ser definida como uma abordagem abrangente para o desenvolvimento de projetos de *data mining*. Fornece um processo independente do setor e da tecnologia utilizada, com o objetivo de tornar este tipo de projetos mais eficientes, confiáveis, gerenciáveis e rápidos. Dentre as principais vantagens desta metodologia destacam-se a sua flexibilidade e aplicabilidade.

É composta por 6 fases inter-relacionadas, passando a enumerar:

1. **Entendimento do Negócio:** É a fase inicial do projeto e foca-se em entender os objetivos e requisitos do projeto numa perspetiva do negócio. Além disso, é crucial adquirir um conhecimento profundo das regras do jogo, posições dos jogadores, dinâmica de jogo e outros elementos essenciais para uma análise precisa. Este conhecimento detalhado será fundamental para traduzir eficientemente os requisitos de negócio na definição do problema. Para além disso, nesta fase deve ser esboçado um plano preliminar para o projeto. No contexto desta tese, será realizada uma revisão aprofundada da literatura sobre previsão de resultados desportivos, com especial ênfase em estudos relacionados com a NBA, com o objetivo de identificar fatores-chave que influenciam o desempenho das equipas e jogadores. Esta análise incluirá não apenas os aspetos estatísticos, mas também as nuances do jogo, tais como estratégias táticas, estilos de jogo predominantes e quaisquer fatores externos que possam impactar o desempenho desportivo.
2. **Entendimento dos Dados:** Esta fase inicia-se com uma recolha inicial dos dados. Após esta recolha, são realizadas atividades que visam aumentar a familiaridade com os dados, identificar problemas de qualidade nos dados, descobrir primeiras perceções sobre os dados ou até mesmo detetar subconjuntos de dados interessantes de forma a elaborar hipóteses relativas a informações escondidas. É importante realçar que esta fase tem uma ligação forte com a primeira fase da metodologia CRISP-DM, visto que, de modo a definir o problema e elaborar o plano, deve já existir algum conhecimento sobre os dados disponíveis. No contexto desta tese, além da análise geral dos dados,

nesta fase serão recolhidos dados históricos da NBA relevantes para a previsão, incluindo resultados de jogos anteriores e estatísticas dos jogadores. A análise destes dados será conduzida de forma abrangente, abordando vários anos, e será complementada por representações gráficas para identificar tendências ao longo do tempo. Isto permitirá não só compreender as variações sazonais, mas também destacar eventuais mudanças significativas nas dinâmicas do jogo ao longo das diferentes épocas. A análise gráfica será essencial para determinar se existem diferenças substanciais entre as várias épocas da NBA. Esta análise detalhada orientará a determinação do número de anos necessários para o desenvolvimento dos modelos de previsão, assegurando que a modelação leva em consideração as nuances temporais relevantes. Ao considerar as características distintas do *basketball* e os fatores críticos para o sucesso nas partidas, espera-se que esta fase forneça conhecimentos valiosos para as etapas seguintes do projeto.

3. **Preparação dos Dados:** Esta fase engloba todas as atividades que visam a construção do *dataset* final que irá ser usado nos modelos. Estas atividades incluem a seleção de tabelas, registos e características, a limpeza dos dados e outras transformações que possam ser necessárias. Considerando as características específicas do *basketball*, no contexto desta tese, serão realizadas seleções criteriosas de tabelas, registos e características relevantes para o desempenho das equipas e jogadores. A limpeza dos dados será uma parte crucial deste processo, abordando desafios específicos que podem surgir no contexto da NBA, como tratamento de dados ausentes, *outliers* e outras peculiaridades. Além disso, serão implementadas transformações que sejam necessárias para melhorar a qualidade e a adequação dos dados ao desenvolvimento dos modelos de previsão. Isto pode incluir o escalonamento de dados, a criação de novas variáveis que capturem aspetos específicos do jogo e a aplicação de técnicas para lidar com possíveis distorções nos dados. Esta fase garantirá que o *dataset* final é otimizado para a previsão de resultados da NBA, levando em consideração os fatores que mais impactam o desempenho das equipas e jogadores.

4. **Modelação:** Nesta fase, são selecionadas e aplicadas várias técnicas de modelação, assim como existe uma calibragem para valores ótimos dos seus parâmetros. É importante realçar que esta fase tem uma ligação forte com a terceira fase da metodologia, visto que, é frequente que sejam detetados problemas nos dados nesta fase, o que leva a uma repetição da preparação dos dados. No contexto desta tese, nesta fase, serão selecionadas e aplicadas técnicas de modelação adaptadas à natureza específica da previsão de jogos da NBA. Considerando as características únicas do *basketball*, será dada atenção especial à escolha de modelos que capturem efetivamente os padrões e fatores determinantes no desempenho das equipas e jogadores. Durante a modelação, haverá uma calibragem cuidadosa para garantir valores ótimos dos parâmetros, levando em consideração as particularidades do contexto da NBA. É essencial destacar que o processo de modelação será sensível às nuances deste desporto, reconhecendo a imprevisibilidade que lhe é associada. Além disso, os modelos serão adaptados para lidar com possíveis fatores externos que podem influenciar os resultados.

5. **Avaliação:** Nesta fase da metodologia, já foram construídos um ou mais modelos com aparente qualidade, de uma perspetiva de análise de dados. Antes de passar para a fase final, é crucial não só fazer uma avaliação rigorosa ao modelo, como também, uma revisão dos passos realizados no seu desenvolvimento, de forma a garantir que todos os objetivos são cumpridos. No final desta fase, a decisão se o modelo irá ser utilizado

ou não deve ser tomada. No contexto desta tese, nesta fase o modelo será avaliado através da análise de métricas de desempenho específicas para previsão de resultados, como a taxa de acerto na previsão de vitórias e derrotas e outras métricas específicas do contexto desportivo. Além disso, será conduzida uma revisão abrangente dos passos realizados durante o desenvolvimento do modelo. Isto garantirá que todos os objetivos estabelecidos na primeira fase foram atingidos e que o modelo está alinhado com as necessidades específicas da previsão de jogos da NBA. A decisão de usar ou não o modelo será baseada não apenas na taxa de acerto das previsões, mas também na validação do processo e na conformidade com os objetivos iniciais.

6. **Implantação:** Esta é a fase final da metodologia. Nesta fase, o conhecimento adquirido é organizado e apresentado ao utilizador final. No contexto desta tese, esta fase refere-se à disponibilização do modelo de previsão de resultados da NBA num *website*.

1.7 Organização do Documento

Este documento está dividido em sete capítulos. O primeiro aborda de maneira geral o problema em questão e os objetivos que se pretendem alcançar. O segundo capítulo apresenta o estado da arte relacionado ao problema, onde também é feita uma análise crítica das abordagens existentes para o resolver, considerando o contexto previamente apresentado. No terceiro capítulo, é apresentado o conjunto de dados utilizado, com uma explicação detalhada do seu tratamento. No quarto capítulo é feita uma exploração dos dados, através de análises gráficas. O quinto capítulo demonstra o processo de treino dos modelos, explicando a metodologia adotada e a forma como o modelo final foi disponibilizado num *website*. O sexto capítulo foca-se na análise dos resultados obtidos, destacando os pontos fortes do modelo, enquanto se analisa, de forma crítica, os aspetos a melhorar. Finalmente, o sétimo capítulo apresenta as conclusões deste trabalho, discutindo o impacto dos resultados alcançados e sugerindo potenciais desenvolvimentos futuros.

Capítulo 2

Estado da Arte

2.1 Área de Negócio

Basketball é um desporto jogado entre duas equipas, cada uma composta por 5 jogadores, em partidas de 48 minutos divididos em quartos de 12 minutos. O objetivo principal é marcar mais pontos que a equipa adversária. Para isso, existem duas maneiras de pontuar: um cesto dentro da linha de 3 pontos, que vale 2 pontos, e um cesto fora da linha de 3 pontos, que vale 3 pontos. Além disso, os ressaltos, onde os jogadores competem para capturar a bola depois de um lançamento falhado, são cruciais para manter ou ganhar a posse de bola. Os bloqueios, quando um jogador impede o lançamento de um adversário, desempenham um papel fundamental na defesa de uma equipa. Quando uma equipa sofre uma falta, é recompensada com um lance livre, que, caso seja concretizado, vale 1 ponto. Cada equipa dispõe de 24 segundos para atacar, sendo que, caso esse intervalo seja ultrapassado, a posse de bola é atribuída à outra equipa (NBA 2023). A Figura 2.1 fornece não só uma visão do formato de um campo de *basketball*, como também das linhas que o limitam.

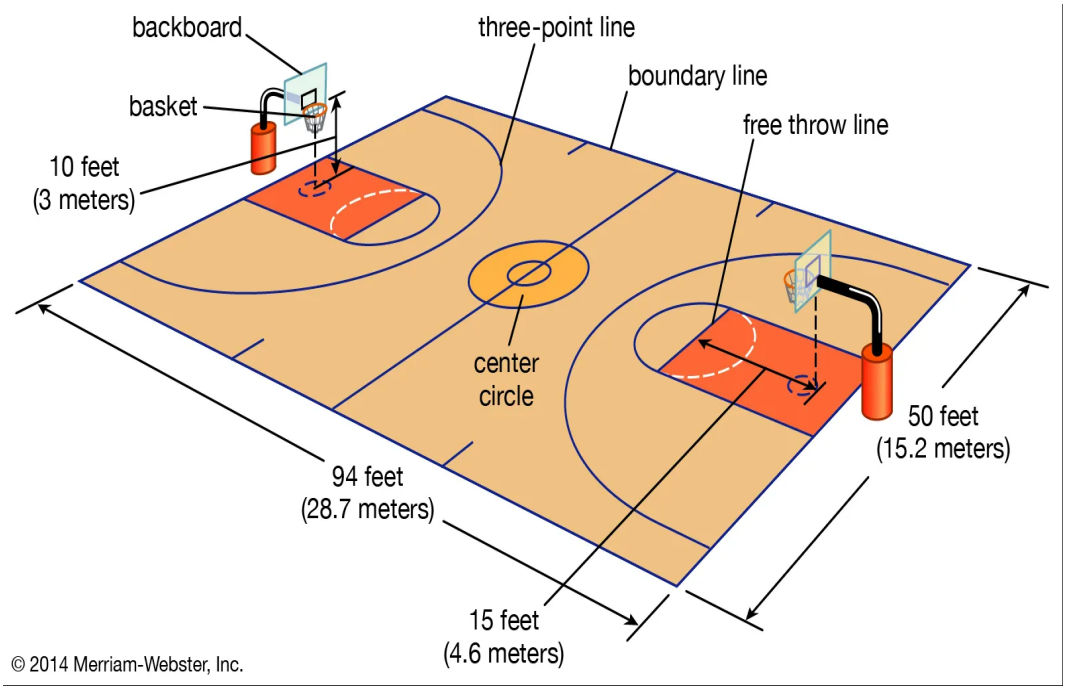

Figura 2.1: Básicos do *basketball* (Mokray, Donald e Logan 2023)

O objetivo de uma aposta desportiva é fazer um prognóstico correto de forma a obter lucro financeiro. Deste modo, uma aposta deste cariz, não deve ser feita de forma inconsciente e não ponderada.

Relativamente ao *basketball*, há diferentes tipos de apostas que um indivíduo pode realizar. Apesar da mais comum ser apostar na equipa vencedora, outros tipos de apostas, também, podem ser feitos, tais como se o número total de pontos ultrapassa um valor ou não ou se um jogador atinge um determinado número numa estatística de jogo, como por exemplo pontos, assistências, bloqueios ou ressaltos (Bwin 2023). Um ponto importante a destacar é que na NBA apenas existem dois resultados possíveis, pois caso um jogo termine empatado, o jogo irá a prolongamento até ser encontrado um vencedor.

O retorno de uma aposta desportiva é proporcional à odd da aposta, isto é, quanto maior a odd maior o retorno. Porém, é importante ressalvar que o mesmo acontece para o risco, ou seja, apostas com odds maiores têm um maior risco. O valor ganho pelo apostador é calculado multiplicando o valor apostado inicialmente pela odd da aposta.

2.2 Metodologia de Pesquisa

A revisão da literatura que se vai apresentar segue a metodologia PRISMA (Sohrabi et al. 2021) que se baseia numa série de procedimentos a seguir para que as pesquisas efetuadas possam ser replicadas e assim comparar as conclusões extraídas. Esta metodologia destaca a necessidade de formular perguntas de pesquisa que contenham os principais pontos sobre os domínios em questão.

Esta pesquisa é orientada pela questão central: "Como desenvolver um modelo de previsão eficaz de resultados de jogos da NBA?". A formulação inicial da pesquisa foi conduzida por meio da seguinte query :

('NBA games' OR 'NBA match prediction' OR 'NBA game outcome prediction')
AND
('predictive model' OR 'machine learning' OR 'data analysis' OR 'statistical model')

A *query* incluiu termos como "NBA games", "NBA match prediction"e "NBA game outcome prediction", combinados com "predictive model", "machine learning", "data analysis"e "statistical model". A revisão da literatura subsequente refinou e focalizou a pesquisa com base nessas consultas específicas.

Para a recolha de informações relevantes, a pesquisa explorou duas bases de dados, b-on e ACM. Estas bases de dados foram escolhidas devido à sua abrangência e relevância para a pesquisa, proporcionando dados essenciais para o desenvolvimento do modelo de previsão de resultados da NBA. Foi aplicado um filtro temporal, restringindo a pesquisa a artigos publicados no período de 2013 a 2023. Para além disso, foram consideradas apenas revistas académicas e materiais de conferência.

Além disso, foram estabelecidos critérios de inclusão e exclusão para garantir a qualidade e relevância dos estudos selecionados. Os critérios estão detalhados nas tabelas 2.1 2.2 respetivamente.

Tabela 2.1: Critérios de inclusão

Inclusão	Critério de Inclusão
IC1	A fonte foca-se na previsão de resultados de jogos da NBA.
IC2	A fonte aborda métodos estatísticos ou algoritmos relevantes para a previsão de jogos da NBA.
IC3	A fonte fornece análises de desempenho de jogadores ou equipas relacionadas à previsão de jogos da NBA.
IC4	A fonte discute fatores-chave, como vantagem de jogar em casa, estatísticas de equipa e outros elementos relevantes para a previsão de jogos da NBA.

Tabela 2.2: Critérios de exclusão

Exclusão	Fator de Exclusão
EC1	A fonte não está escrita em inglês.
EC2	A fonte é um procedimento que não pertence a uma conferência com classificação CORE A ou CORE B.
EC3	A fonte não é suficientemente clara ou aplicável aos objetivos da previsão de resultados de jogos da NBA.

O processo de seleção seguiu os seguintes passos:

1. Eliminação de artigos duplicados
2. Filtragem por *abstracts*: os *abstracts* dos artigos foram lidos e analisados para determinar a relevância em relação à pesquisa. Os artigos que não estavam relacionados com previsão de resultados da NBA foram excluídos nesta fase.
3. Leitura completa dos artigos.
4. Aplicação dos critérios de inclusão e exclusão: Os critérios de inclusão e exclusão detalhados nas tabelas 2.1 e 2.2 respetivamente foram aplicados nesta etapa para garantir a seleção de estudos pertinentes à pesquisa.

Na Figura 2.2, é possível observar o fluxo deste processo, incluindo o número de registos eliminados em cada um dos passos.

É importante referir que as referências bibliográficas dos artigos selecionados foram também consultadas durante a escrita deste documento, de forma a obter conhecimento adicional. Neste estágio, foram aceites outros tipos de fontes, principalmente livros, que forneciam conhecimentos significativos para o desenvolvimento do modelo de previsão.

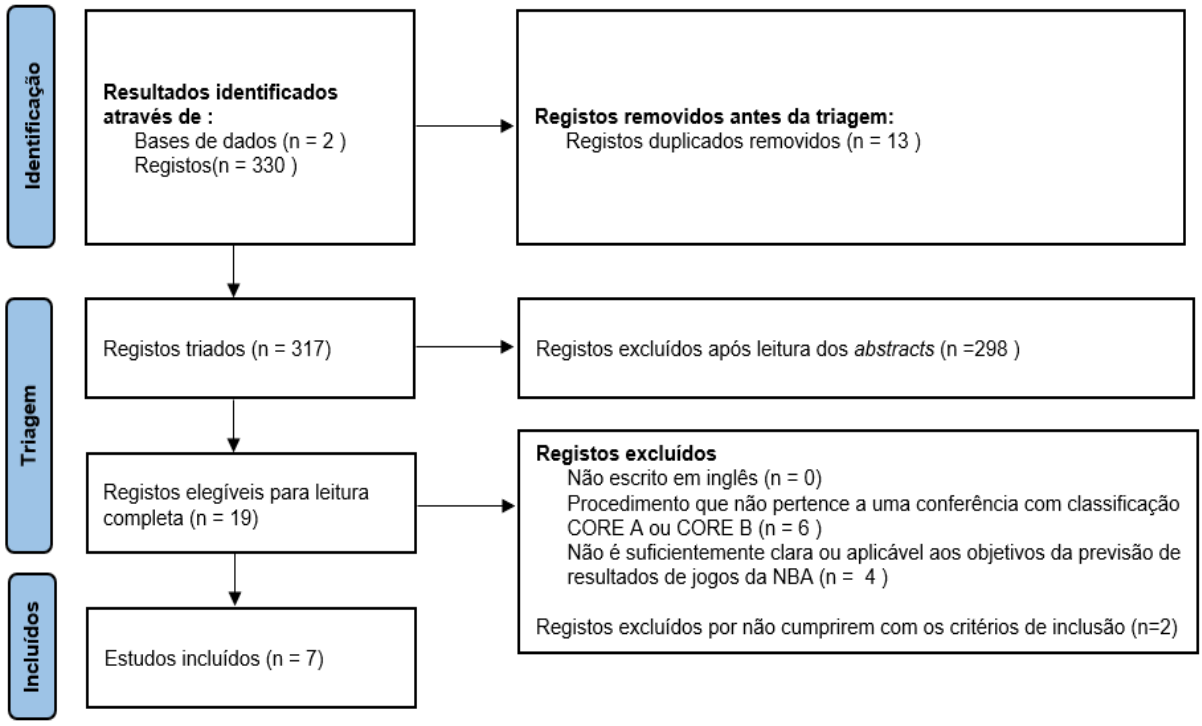

Figura 2.2: Processo de seleção de estudos

2.3 Estado da Arte em Abordagens Existentes

Um estudo recente (Horvat, Job et al. 2023) propôs um modelo de *Machine Learning* (ML) que combina estatísticas da liga e um índice de eficiência de equipa personalizado, *Extended Team Efficiency Index* (ETEI), criado através da expansão do índice tradicionalmente utilizado na NBA, com o objetivo de prever resultados dos jogos da liga. Para este caso de estudo foram utilizados dados sobre 6567 jogos de 5 épocas da NBA. O modelo criado pelos autores superou os modelos convencionais de ML, atingindo uma taxa de acerto média de 66% para mais de 2500 jogos. A taxa de acerto máxima obtida foi de 78%, porém esta foi atingida com apenas uma época para treino e outra época para teste do modelo, o que suscita algumas preocupações sobre a possibilidade de *overfitting*. Além disso, a discrepância que se faz notar entre a taxa de acerto máxima alcançada e a taxa de acerto média indica uma variabilidade significativa no desempenho do modelo em diferentes ambientes, sugerindo a necessidade de uma investigação adicional através de, por exemplo, outros indicadores como *recall* e *F1-score*.

(Ozkan 2020) propôs um inovador sistema híbrido inteligente que combina os pontos fortes das Redes Neuronais Artificiais e dos métodos da lógica *fuzzy* para prever o resultado de jogos de *basketball*. O estudo usa dados da época 2015/2016 da liga turca de *basketball*. Para a implementação prática deste sistema, foi adotada uma estrutura conectada em cascata, onde o Sistema de Inferência *Fuzzy* está conectado a uma Rede Neuronal Artificial (RNA). Inicialmente, é determinada a equipa favorita com base nos valores de ambas as equipas, que são depois utilizados como *input* da rede neuronal. O autor compara o modelo de rede neuronal tradicional, que alcançou valores de taxa de acerto, sensibilidade e especificidade de 70.8%, 54.5% e 84.6% respetivamente, com o modelo híbrido de rede neuronal com lógica *fuzzy* que, para esses critérios alcançou valores de 79.2%, 72.7% e 79.1%. É concluído que

o modelo híbrido teve um desempenho superior ao modelo composto apenas por uma rede neuronal, comprovando a sua eficiência em prever resultados de jogos de *basketball*.

Este modelo apresentou a taxa de acerto mais alta de todos os estudos, com uns significativos 79,2%, o que o torna bastante promissor. No entanto, é essencial abordar as restrições a nível de tamanho do *dataset* utilizado, que contava com apenas 240 jogos. Para além disso, visto que os dados usados para treino do modelo pertencem à mesma época que os dados usados para testar o modelo, este não tem em consideração possíveis mudanças que possam impactar significativamente o desempenho das equipas. Estes dois fatores suscitam alguma preocupação sobre o risco de *overfitting*, o que leva a crer que com um *dataset* que contenha mais jogos e de diferentes épocas, a taxa de acerto sofresse uma queda. Outra crítica que se pode apontar a este estudo é o facto de não ter incluído parâmetros individuais, tais como o talento dos jogadores, as suas *performances* e motivação.

(Zhao, Du e G. Tan 2023) propuseram um método baseado em *Graph Convolution Networks* (GCN) para prever resultados na NBA, convertendo dados estruturados em grafos não estruturados. O estudo utiliza um grafo homogéneo, onde os nós representam equipas e as arestas representam jogos passados/futuros, conectando as equipas entre si através de relações de jogos. Foram exploradas três abordagens distintas, cada uma combinando GCN com uma técnica específica de ML. Na primeira, foi usado *Principal Component Analysis* (PCA) com o objetivo de reduzir a dimensão dos dados. O segundo método empregou a técnica LASSO para a seleção de características, enquanto o terceiro adotou *Random Forest* (RF) para seleção de características. O estudo utilizou um *dataset* com dados estatísticos da NBA de 2012 a 2018. A conclusão destaca que a combinação mais eficaz foi CGN com RF, alcançando uma taxa de acerto média de 71,54%, superando a combinação GCN e LASSO. Por outro lado, a combinação PCA e CGN não atingiu os resultados esperados, sugerindo uma perda de dados essenciais. Adicionalmente, o modelo proposto superou os modelos *baseline*.

O modelo apresentou uma taxa de acerto média de 71,54%. É importante referir que, apesar de o *dataset* utilizado conter 6 épocas da NBA, este foi dividido em 6 *datasets* diferentes, um *dataset* para cada época. Deste modo, as críticas feitas ao modelo anterior também se aplicam a este estudo relativamente ao tamanho dos *datasets*, do risco de *overfitting* e da não consideração por possíveis mudanças que impactem o desempenho das equipas. Apesar destes pontos, o modelo revela-se consistente, visto que não se identificam variações excessivamente altas na taxa de acerto do modelo de época para época, sendo o mínimo 69.51% e o máximo 73,78%.

Ao contrário da maioria dos estudos, que se focam em prever o vencedor de um jogo, um estudo de 2021 (Chen et al. 2021) focou-se em prever o marcador final exato de um jogo da NBA, utilizando estatísticas das equipas durante a temporada 2018-2019. Para atingir este objetivo, foram explorados dois tipos principais de modelos: os modelos individuais (S-) e os modelos de duas etapas (T-). Os modelos individuais, como S-ELM, S-MARS, S-XGBoost, S-SGB, S-KNN, analisaram diretamente as estatísticas do jogo para prever o resultado final. Por outro lado, os modelos de duas etapas, como T-ELM, T-MARS, T-XGBoost, T-SGB, T-KNN, incorporaram um processo mais complexo. Primeiramente são identificadas as características cruciais a partir das estatísticas do jogo, sendo esta seleção feita através da utilização dos algoritmos MARS, XGBoost e SGB. Cada um destes algoritmos gera os seus melhores subconjuntos, que são depois combinados de forma a proporcionar uma seleção estável e eficaz de características. Em seguida, essas características são utilizadas para prever o marcador final. O modelo T-XGBoost destacou-se, apresentando o melhor desempenho

de forma consistente, especialmente quando considerado um intervalo específico de jogos (*game-lag* = 4). Além disso, o estudo enfatiza a importância da escolha apropriada do número de jogos anteriores utilizados para a obtenção de resultados mais precisos.

O modelo apresentou um MAPE de 0.0818, indicando uma notável capacidade de previsão. No entanto, é importante observar que o conjunto de dados utilizado abrange apenas uma temporada da NBA, suscitando críticas relacionadas ao tamanho limitado do conjunto de dados, ao risco de *overfitting* e à falta de consideração de possíveis mudanças que poderiam impactar o desempenho das equipes, conforme discutido em estudos anteriores. Outro ponto de crítica deste estudo, prende-se no facto de que, no modelo de previsão não foram incluídos dados sobre a equipa adversária. A ausência destas informações pode limitar a capacidade do modelo de capturar nuances importantes que possam afetar os resultados dos jogos, diminuindo assim a sua aplicabilidade prática e generalização para diferentes situações.

(Cheng et al. 2016) propuseram a criação de um modelo (NBAME) para previsão de resultados de jogos de *playoff* da NBA, recorrendo ao Princípio da Entropia Máxima (PEM). Os autores abordam as limitações associadas a outros métodos de ML, como redes neuronais, Árvore de Decisão (AD) e *Naive Bayes* (NB), que enfrentam desafios como *overfitting* devido a conjuntos de dados limitados, falta de independência entre características e incapacidade de fornecer valores de probabilidade interpretáveis. O modelo sugerido foi capaz de prever o vencedor da partida com 74,4% de taxa de acerto, demonstrando ter uma *performance* superior a outros algoritmos de ML (NB, Regressão Logística (RL), Back Propagation (BP) Neural Networks, RF, cuja taxa de acerto máxima que conseguiram atingir foi de 70,6%.

O modelo atingiu uma taxa de acerto máxima de 74,4%, o que é de facto notável. Porém, é importante referir que, apesar de o *dataset* utilizado conter 8 épocas da NBA, o modelo foi sempre treinado para temporadas individuais, levantando dúvidas relativamente a *overfitting*, devido à quantidade limitada de dados em cada temporada. O modelo NBAME mostra sensibilidade ao ajuste do limiar de decisão (*threshold*). Quando este critério é aumentado, o modelo é capaz de prever um menor número de jogos, porém com maior precisão. Isto significa que é possível ajustar o modelo consoante a prioridade que damos ao número de jogos a prever ou à precisão que pretendemos. Esta flexibilidade é considerada valiosa, especialmente em contextos comerciais.

(Horvat, Hava e Srpak 2020) realizaram um estudo cujo objetivo principal foi identificar a melhor combinação de algoritmo de ML, método de validação e preparação de dados para prever resultados de jogos da NBA. Utilizando dados de nove temporadas consecutivas, de 2009/2010 a 2017/2018, os autores aplicaram métodos supervisionados de ML, empregando sete algoritmos distintos: RL, NB, AD, rede neuronal *multilayer perceptron, RF, K-Nearest Neighbors (KNN)* e *LogitBoost*. Duas técnicas diferentes de preparação de dados foram aplicadas: dados disjuntos, onde os conjuntos de treino e teste estavam completamente separados, sem compartilhar qualquer informação, e dados atualizados, onde, após cada previsão, os dados de resultados conhecidos da fase de teste eram adicionados ao conjunto de treino. Inicialmente, os resultados foram validados por meio de dois métodos, *Train & Test validation* e validação cruzada, utilizando os dados disjuntos. A análise revelou que validação cruzada foi mais eficaz. O método de validação *Train & Test validation* apresentou resultados satisfatórios. Em seguida, foi analisado o cenário de *Train & Test validation* com dados atualizados. Foi concluído que, para o método de validação *Train & Test validation*, o uso de dados atualizados apresentava melhores resultados. Ao analisar os algoritmos de ML utilizados, observou-se que, em geral, os melhores resultados foram alcançados ao empregar o algoritmo KNN, enquanto os piores resultados foram obtidos ao utilizar AD. Pelos dados

demonstrados, a melhor combinação de método de preparação, algoritmo de ML e método de avaliação foi o uso de dados atualizados, KNN e *Train & Test*, alcançando uma taxa de acerto média de 60.01% e máxima de 60,82%, que foi atingida utilizando uma temporada para treino e duas temporadas para teste.

Apesar do estudo fazer uma boa análise da importância da escolha de um método de validação no treino de um modelo de previsão, os resultados alcançados foram apenas razoáveis, visto que a taxa de acerto média atingida foi apenas de 60.01%. Uma lacuna significativa observada é à ausência do treino de modelos para o cenário de validação cruzada com dados atualizados. Enquanto que os resultados para os outros cenários foram apresentados de maneira abrangente, a falta de uma análise específica para validação cruzada com dados atualizados limita a compreensão completa do desempenho desse cenário.

(Zheng 2022) propôs a introdução de fatores internos e externos, tais como a classificação ELO das equipas, *performance* média das equipas em jogos recentes, fator casa e cansaço dos jogadores devido a jogos consecutivos com o objetivo de prever resultados de jogos da NBA. Para este objetivo foram usados dados históricos desde a época 2012-13 até à época 2020-21. Neste estudo, foi analisada a *performance* de vários modelos de ML como Redes Neuronais, *Support Vector Machine* (SVM), RL, RF e NB. Foi possível atingir um valor de taxa de acerto de 67,98% com o algoritmo RF. Através da análise dos dados obtidos, o autor concluiu que as novas variáveis introduzidas tiveram um impacto positivo na precisão dos modelos treinados e que, para o fim de prever resultados de jogos da NBA, os melhores algoritmos eram RF e NB.

Em suma, os estudos analisados apresentam uma ampla variação nas taxas de acerto dos modelos, tendo sido o valor mínimo atingido 60.01% e o máximo 79.2%. Uma comparação objetiva e fiável entre os diferentes estudos seria difícil de realizar e muito discutível, devido aos diferentes dados utilizados e às diferentes condições experimentais de cada trabalho. No entanto, estes valores enfatizam a dificuldade em obter modelos com taxas de acerto elevadas e consistentes na previsão de resultados de jogos da NBA.

A tabela 2.3 apresenta de forma resumida as abordagens estudadas neste estado da arte.

Após a análise dos diversos modelos abordados sobre previsão de resultados da NBA, torna-se evidente a complexidade e a diversidade de abordagens existentes. Cada algoritmo apresentado possui vantagens e limitações próprias, o que enfatiza a necessidade de uma escolha criteriosa do modelo mais adequado para o contexto específico. As variáveis consideradas também desempenham um papel crucial na eficácia destes modelos, revelando a importância de uma análise abrangente e atualizada das características relevantes para a previsão de resultados da NBA.

Além disso, os resultados obtidos oferecem conhecimentos valiosos sobre a aplicabilidade prática destes modelos, indicando áreas de sucesso e possíveis oportunidades de aprimoramento. À medida que avançamos na pesquisa sobre previsão de resultados desportivos, é imperativo considerar não apenas os aspetos técnicos dos algoritmos, mas também a adaptabilidade e a interpretabilidade dos modelos, garantindo a sua utilidade no contexto dinâmico da NBA.

Em resumo, é destacada a complexidade do desafio de previsão de resultados desportivos, ao mesmo tempo que é realçada a promissora evolução dos modelos existentes.

Tabela 2.3: Resumo das abordagens existentes

Artigo	*Dataset*	Atributos	Algoritmos	Resultados
(Horvat, Job et al. 2023)	2013-2018	I NBA I CMPR Básicos(ex:pontos, assistências) Derivados(ex:eficiência de lançamento) Avançados(ex:eficácia defensiva) Classificação (ex:posição na tabela)	*ETEI*	Taxa de Acerto Média: 66% Taxa de Acerto Máxima: 78%
(Ozkan 2020)	2015/2016	Equipa da casa Equipa visitante Média de pontos nos últimos 4 jogos (casa) Média de pontos nos últimos 4 jogos (visitante) Média de pontos na liga (casa) Média de pontos na liga (visitante) Média de pontos sofridos últimos 4 jogos (casa) Média de pontos sofridos últimos 4 jogos (visitante) Jornada Resultado	Lógica *Fuzzy* RNA	Taxa de Acerto: 79,2% Sensibilidade: 72,7% Especificidade: 79,1%
(Zhao, Du e G. Tan 2023)	2012-2018	*Team Efficiency Differential* *Team Defensive Rating* *Team Floor Impact Counter*	CGN RF	Taxa de Acerto Média:71,54% Taxa de Acerto Máxima:73,78%
(Chen et al. 2021)	2018/2019	Tentativas de lançamento (2pts, 3pts, lances livres) Percentagem de lançamento (2pts, 3pts, lances livres) Ressaltos (ofensivos e defensivos) Assistências Roubos de Bola Bloqueios *Turnovers* (perdas de bola) Faltas pessoais Pontuação da equipa	MARS *XGBoost* SGB	MAPE:0.0818
(Cheng et al. 2016)	2007-2015	Cestos (3pts, 2pts, lances livres) casa/visitante Tentativas de lançamento (3pts, 2pts, lances livres) casa/visitante Ressaltos (ofensivos e defensivos) casa/visitante Assistências casa/visitante Roubos de bola casa/visitante Bloqueios casa/visitante *Turnovers* casa/visitante Faltas pessoais casa/visitante Pontos casa/visitante Resultado	PEM	Taxa de Acerto:74,4%
(Horvat, Hava e Srpak 2020)	2009-2018	Cestos (3pts, 2pts, lances livres) Tentativas de lançamento (3pts, 2pts, lances livres) Ressaltos (ofensivos e defensivos) Assistências Roubos de bola *Turnovers* Bloqueios Faltas cometidas	*Train & Test* KNN	Taxa de Acerto Média:60,01% Taxa de Acerto Máxima:60,82%
(Zheng 2022)	2012-2021	Estatísticas médias de jogos recentes(casa e visitante) ELO Vantagem casa Cansaço (casa) Cansaço (visitante) Diferenças entre equipas	RF	Taxa de Acerto:67,98%

2.4 Algoritmos

De acordo com os estudos analisados para modelos de previsão de resultados da NBA, serão utilizados algoritmos de classificação simples, como RL, AD, NB, KNN e SVM, além de técnicas mais elaboradas de *ensemble learning* como *Voting Classifier*, *Stacking*, *Bagging* e *Boosting*. Nas subsecções seguintes são descritos os algoritmos utilizados nesta tese.

2.4.1 Regressão Logística

RL é uma generalização de regressão linear e é utilizada na previsão de variáveis dependentes, quer sejam elas binárias (variáveis com apenas dois resultados possíveis, como "sim"ou "não", "1"ou "0", "verdadeiro"ou "falso") ou multiclasse (variáveis com mais de dois resultados possíveis, como por exemplo tipo de filme, "Comédia", "Terror"ou "Ação"), sendo as respostas discretas. Este modelo estatístico utiliza funções *sigmoides*, caracterizadas por posicionar números reais no intervalo entre 0 e 1, para modelar variáveis dependentes binárias. A semelhança com a regressão linear é evidente, uma vez que ambas calculam saídas para entradas específicas usando coeficientes ou pesos. Contudo, a RL produz exclusivamente saídas binárias, 0 ou 1. Este método é particularmente útil na classificação de dados de baixa dimensionalidade com fronteiras não lineares. Para resolver problemas

de classificação binária, a função linear (2.1) é estendida pela função logística (2.2) para formar uma nova função (2.3) (Horvat, Hava e Srpak 2020).

$$y(X) = W^T \cdot X + w_0 \tag{2.1}$$

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

$$y(X) = F(\theta^T \cdot X + w_0) = f(z) \tag{2.3}$$

Sendo o parâmetro z definido pela equação 2.4

$$z = \theta^T \cdot X + w_0 \tag{2.4}$$

O retorno da nova função é um valor contido no intervalo [0,1] e é interpretado como a probabilidade de a variável alvo ser 1 dado o exemplo X P(C1/X). Quando este valor é maior ou igual a 0.5, X é classificado como 1, caso contrário, X é classificado como 0 (Cao 2012).

2.4.2 Árvore de Decisão

Segundo (Horvat, Hava e Srpak 2020) AD é um algoritmo de classificação que replica a estrutura de uma árvore. As folhas representam as classes e os ramos representam as combinações de características. Este algoritmo pertence ao conjunto de algoritmos de aprendizagem supervisionada e é aplicado tanto em problemas de regressão como em problemas de classificação, sendo o objetivo principal a criação de um modelo de treino capaz de prever as variáveis-alvo com base em regras de decisão aprendidas durante o treino do modelo.

Uma AD é constituída por 3 tipos de nós (P.-N. Tan, Steinbach, Karpatne et al. 2018), como é possível verificar na Figura 2.3 :

- **Nó raiz:** não possui qualquer ramo de entrada e pode conter 0 ou mais ramos de saída.
- **Nós internos:** possuem exatamente um ramo de entrada e dois ou mais ramos de saída. Os nós não terminais, onde se incluem este tipo de nós e o nó raiz, contêm condições de teste de atributos de forma a separar registos que apresentam diferentes características.
- **Folhas:** possuem exatamente um ramo de entrada e nenhum ramo de saída. A cada um deste nós é atribuída uma classe.

A classificação de um registo inicia-se no topo da árvore, conhecido como nó raiz, onde uma condição de teste é aplicada ao registo em análise. Consequentemente, o percurso prossegue ao longo do ramo apropriado, com base no resultado do teste anterior. Este caminho pode conduzir a um novo nó interno, onde o processo é repetido com uma nova condição de teste, ou diretamente a uma folha. Assim que uma folha é alcançada, a classe associada a essa folha é atribuída ao registo, concluindo assim o processo de classificação (P.-N. Tan, Steinbach, Karpatne et al. 2018). A Figura 2.4 representa um exemplo de classificação de um animal em mamífero ou não mamífero.

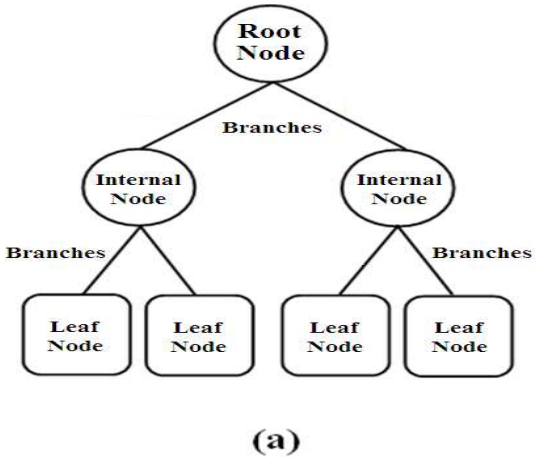

Figura 2.3: Estrutura de uma AD (Sá et al. 2016)

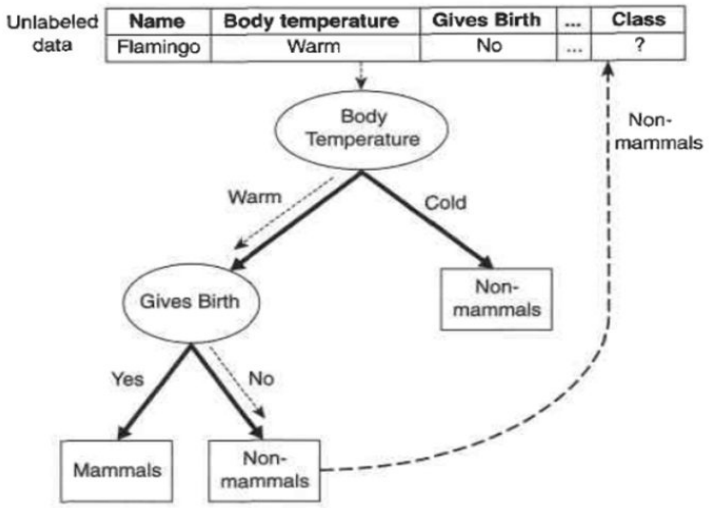

Figura 2.4: Processo de classificação numa AD de decisão(P.-N. Tan, Steinbach, Karpatne et al. 2018)

2.4.3 Naive Bayes

NB é um conjunto de técnicas simples de classificação que tem por base o teorema de *Bayes* (2.5) (Horvat, Hava e Srpak 2020), que afirma que a probabilidade de Y ocorrer dado que X ocorreu ($P(Y|X)$) é proporcional à probabilidade de X ocorrer dado que Y ocorreu ($P(X|Y)$) multiplicada pela probabilidade original de Y ocorrer ($P(Y)$), e então dividida pela probabilidade original de X ocorrer ($P(X)$). (P.-N. Tan, Steinbach e Kumar 2013).

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)} \tag{2.5}$$

No contexto de classificação, o algoritmo NB, calcula a probabilidade de um registo X pertencer a uma classe Y. É importante realçar que este algoritmo opera sobre a premissa de total independência entre as variáveis envolvidas, ou seja o valor de uma variável não

afeta nem é afetado pelo valor de outra variável (Horvat, Hava e Srpak 2020) (P.-N. Tan, Steinbach e Kumar 2013).

2.4.4 K-Nearest Neighbors

KNN é um algoritmo supervisionado não paramétrico de ML utilizado para resolver problemas de classificação e regressão, através da atribuição de pesos às contribuições dos vizinhos, onde os mais próximos têm uma influência maior do que os demais (Chen et al. 2021) (Horvat, Hava e Srpak 2020).

A essência deste algoritmo reside no valor da métrica de distância, sendo a distância euclidiana a métrica mais comum (Chen et al. 2021) (Horvat, Hava e Srpak 2020). O KNN identifica um grupo de k objetos no conjunto de treino que estão mais próximos do objeto de teste, facilitando a atribuição de uma *label* com base na prevalência de uma classe específica nessa "vizinhança". Ao contrário de métodos baseados em modelos, nos quais é feito um treino antes da previsão, este algoritmo reduz a etapa de treino e realiza tarefas de classificação, calculando a distância entre os pontos de dados de teste e todos os pontos de dados de treino para obter os vizinhos mais próximos e, em seguida, prosseguir com a classificação KNN (Chen et al. 2021).

2.4.5 Support Vector Machine

SVM é um algoritmo utilizado para classificação e regressão, fundamentado na teoria de aprendizagem estatística. Uma das suas grandes vantagens é a capacidade de lidar com grandes quantidades de dados, evitando problemas relacionados com alta dimensionalidade. (P.-N. Tan, Steinbach, Karpatne et al. 2018) (Cervantes et al. 2020).

O objetivo deste algoritmo é encontrar o hiperplano que melhor separa as diferentes classes de dados, maximizando a distância entre elas (Vidhya s.d.). Este hiperplano funciona como uma linha divisória que divide os dados em diferentes categorias. Por outras palavras, o SVM procura a linha que deixa a maior distância possível entre as categorias de dados, garantindo que a separação seja a mais clara e robusta possível. Isto é possível devido à sua extraordinária capacidade de generalização, isto é, a capacidade de encontrar sempre a melhor solução para separar os dados, e ao seu poder de descriminição, ou seja, a sua capacidade de distinguir diferentes classes de dados de forma eficaz (Cervantes et al. 2020).

Apesar das suas vantagens, SVM apresenta algumas limitações, como a complexidade computacional em conjuntos de dados muito grandes, o que resulta em tempos de treino significativos (Cervantes et al. 2020).

2.4.6 Voting Classifier

Voting Classifier é uma técnica de ML que combina as previsões de múltiplos modelos individuais com o objeto de tomar uma decisão final sobre a classe de saída. (Khatun et al. 2023). Existem dois tipos principais de classificador, *Voting Classifier Hard* (VCH) e *Voting Classifier Soft* (VCS) (Shariah et al. 2022).

Em VCH, cada modelo individual faz uma previsão e a classe final é determinada pela maioria dos votos (Shariah et al. 2022) (Medium 2023b).

Por outro lado, em VCS, as previsões dos modelos individuais são ponderadas com base na confiança das previsões de cada modelo. Cada modelo atribui probabilidades para cada classe

e a classe final é determinada pela média ponderada dessas probabilidades. Isso permite que modelos mais confiantes tenham um peso maior na decisão final (Sherazi, Bae e Lee 2021) (Medium 2023b). Por exemplo, suponhamos que temos três classificadores, cada um treinado para classificar imagens de gatos e veados. O primeiro classificador prevê que a imagem é de um gato com uma probabilidade de 70% e de um veado com uma probabilidade de 30%; o segundo classificador prevê que a imagem é de um gato com uma probabilidade de 40% e de um veado com uma probabilidade de 60%; e o terceiro classificador prevê que a imagem é de um gato com uma probabilidade de 90% e de um veado com uma probabilidade de 10%. A previsão seria que a imagem é de um gato, pois a probabilidade total da classe gato é (0,7 + 0,4 + 0,9) / 3 = 0,666 e a probabilidade da classe veado é (0,3 + 0,6 + 0,1) / 3 = 0,333. A imagem abaixo ilustra o mesmo (Medium 2023b).

A Figura 2.5 e a Figura 2.6 ilustram, respetivamente, o funcionamento de VCH e VCS.

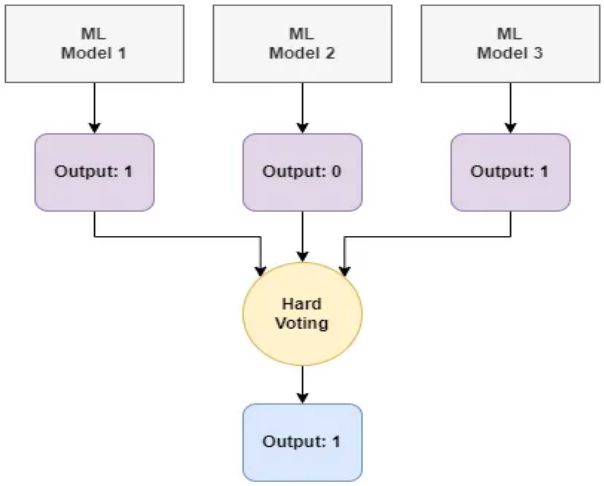

Figura 2.5: Exemplo do funcionamento de VCH (Medium 2023b)

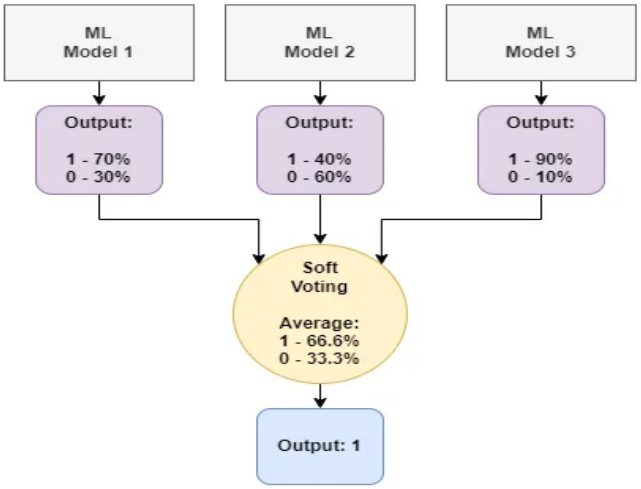

Figura 2.6: Exemplo do funcionamento de VCS (Medium 2023b)

2.4.7 Stacking

Stacking é uma técnica de *ensemble learning* que permite combinar vários modelos de classificação através de um meta-classificador. Este método consiste em duas fases de aprendizagem. Inicialmente, os classificadores base são treinados utilizando o conjunto de dados original. As previsões resultantes destes classificadores são então utilizadas como novos dados de entrada para a fase seguinte, onde um meta-classificador é treinado utilizando estas previsões, produzindo assim a previsão final (Alexandropoulos et al. 2019) (Ali et al. 2022).

A principal vantagem desta técnica é o facto de tirar proveito da diversidade dos erros de previsão dos vários classificadores base, pois diferentes algoritmos de aprendizagem apresentam diferentes vieses e, consequentemente, erram em pontos distintos (Alexandropoulos et al. 2019). Desta forma, ao combinar os modelos, as falhas individuais são mitigadas, resultando num modelo mais robusto e com melhor desempenho generalizado. Para garantir que as previsões dos classificadores base sejam confiáveis, o conjunto de dados de treino é frequentemente dividido em subconjuntos, onde uma parte é utilizada para treinar os classificadores base e outra para gerar as previsões que serão usadas pelo meta-classificador (Ali et al. 2022). Esta abordagem garante que o meta-classificador não é treinado com dados utilizados anteriormente pelos classificadores base, evitando assim *overfitting*.

Em resumo, *stacking* melhora o desempenho de modelos de aprendizagem, combinando de forma inteligente as previsões de múltiplos classificadores base, através de um meta-classificador, resultando em previsões mais precisas e robustas.

Na Figura 2.7 está ilustrado o funcionamento desta técnica.

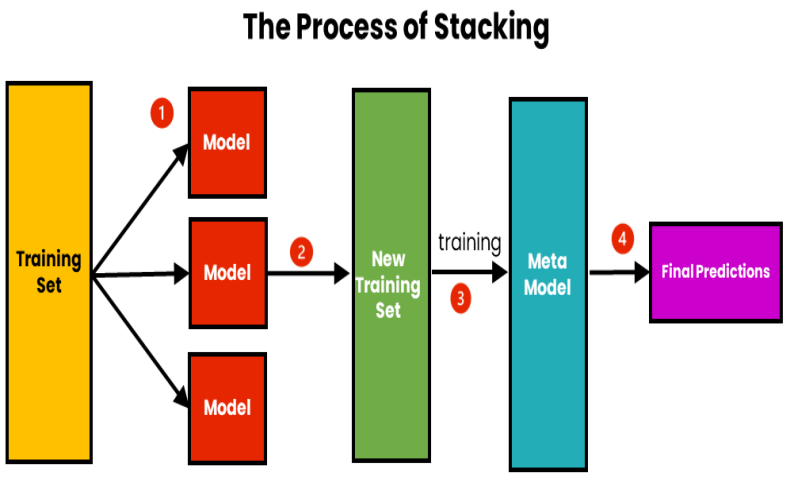

Figura 2.7: Exemplo do funcionamento de *Stacking* (Medium s.d.[b])

2.4.8 Bagging

Bagging é uma técnica de *ensemble learning* que tem como objetivo aumentar o desempenho de algoritmos base, como por exemplo AD, através da criação de subgrupos de dados, denominados *bags*. Um modelo é treinado em cada um desses subconjuntos e, as previsões individuais desses modelos são agregadas, por votação ou por média, de forma a retornar uma previsão final (Ngo, Beard e Chandra 2022) (learn s.d.). Esta técnica ajuda a diminuir a variação do modelo final, tornando-o mais robusto e preciso (Ngo, Beard e Chandra 2022).

RF é um exemplo da implementação desta técnica, onde são usadas múltiplas árvores de decisão (Ngo, Beard e Chandra 2022).

A Figura 2.8 ilustra o funcionamento da técnica *Bagging*.

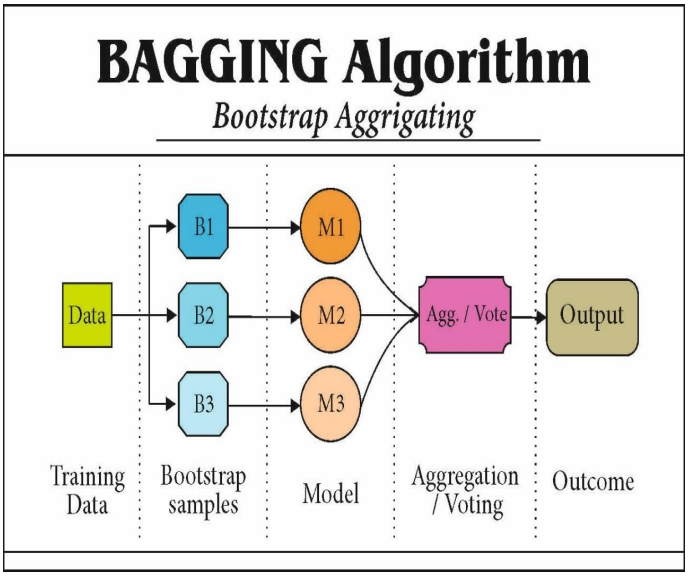

Figura 2.8: Exemplo do funcionamento de *Bagging* (Medium s.d.[a])

2.4.9 Random Forest

RF é um algoritmo de ML amplamente utilizado para tarefas de classificação e regressão que agrega previsões de múltiplos modelos para gerar uma previsão final. Além disso, este algoritmo pode ser empregue para extração de características, identificando as mais significativas dentro de um conjunto de dados. O funcionamento do algoritmo baseia-se na criação de múltiplas árvores de decisão, sendo que cada árvore é treinada com um subconjunto aleatório de dados e características. Durante o treino do modelo, são criadas várias árvores, cada uma treinada com um número definido de amostras do conjunto de treino original. A ideia é treinar cada árvore com conjuntos diferentes de amostras para reduzir a variância para toda a floresta, sem aumentar o viés. Embora cada árvore possa ter uma variância alta em relação a um conjunto específico de dados de treino, a previsão mais frequente das árvores individuais é a previsão final do *Random Forest* (Horvat, Hava e Srpak 2020) (Zhao, Du e G. Tan 2023).

Durante o treino do modelo, é calculada a importância de cada característica com base na sua contribuição para a precisão global do modelo. Características com maior contribuição recebem pontuações mais elevadas. A análise dessas pontuações é útil para identificar as características mais importantes no conjunto de dados, sendo particularmente benéfica em situações de dados em alta dimensão ou quando se deseja simplificar o modelo sem comprometer a sua precisão (Zhao, Du e G. Tan 2023).

O funcionamento do algoritmo RF é demonstrado na Figura 2.9. O algoritmo começa por gerar diferentes árvores de decisão, de forma aleatória de forma a conterem características diferentes. No final as previsões são combinadas e as decisões são feitas com base na maioria dos votos (P.-N. Tan, Steinbach e Kumar 2013). Por exemplo, se tivermos 4 árvores de

decisão e 3 delas apontarem que uma equipa com uma média de pontos por jogo superior a 110 ganha, a árvore final também considerará que essa equipa sairá vitoriosa na partida.

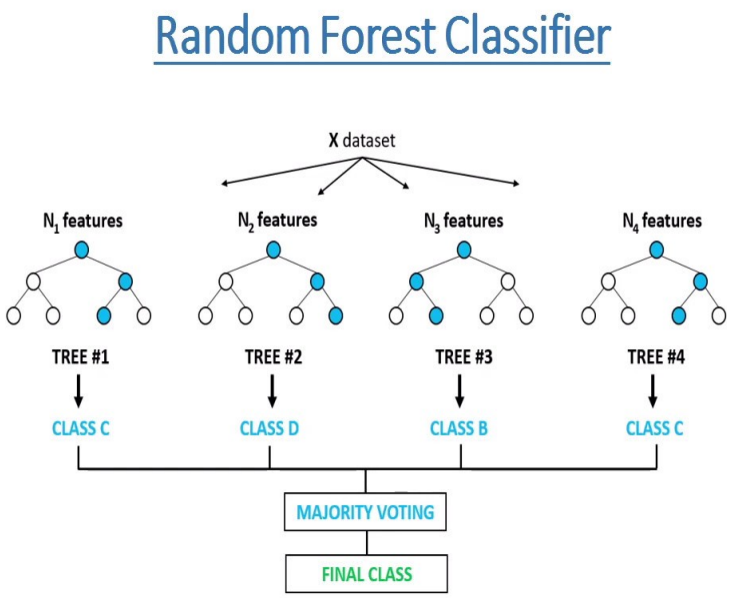

Figura 2.9: Exemplo de funcionamento de RF (Khushaktov 2023)

2.4.10 LASSO

Least Absolute Shrinkage and Selection Operator (LASSO) é um algoritmo utilizado para selecionar as características mais importantes num conjunto de dados com muitas variáveis, identificando aquelas que têm impacto significativo no modelo. Ao adicionar uma penalização à função de custo da RL, o LASSO reduz os coeficientes das variáveis menos relevantes a zero, resultando num modelo que mantém apenas as características mais significativas. Esta técnica de regularização ajuda a evitar *overfitting*, eliminando variáveis irrelevantes e reduzindo a complexidade do modelo. Assim, o LASSO cria modelos mais simples e interpretáveis, melhorando a precisão das previsões em novos dados (Zhao, Du e G. Tan 2023).

2.4.11 Gradient Boosting

Gradient Boosting (GB) é um método de *ensemble learning*, utilizado em problemas de regressão e classificação, que se distingue por construir um modelo robusto através da combinação sequencial de vários modelos fracos (Natekin e Knoll 2013) (Aziz et al. 2020). O processo de aprendizagem ajusta novos modelos de forma a fornecer estimativas mais precisas, alinhando-os com o gradiente negativo da função de perda. Por outras palavras, cada novo modelo é desenvolvido para corrigir os erros do modelo anterior, sendo altamente correlacionado com o gradiente da função de perda escolhida (Natekin e Knoll 2013). Este método permite a criação de um estimador mais forte, a partir da combinação ponderada de

modelos fracos, otimizando a previsão. O comportamento do GB está ilustrado na Figura 2.10.

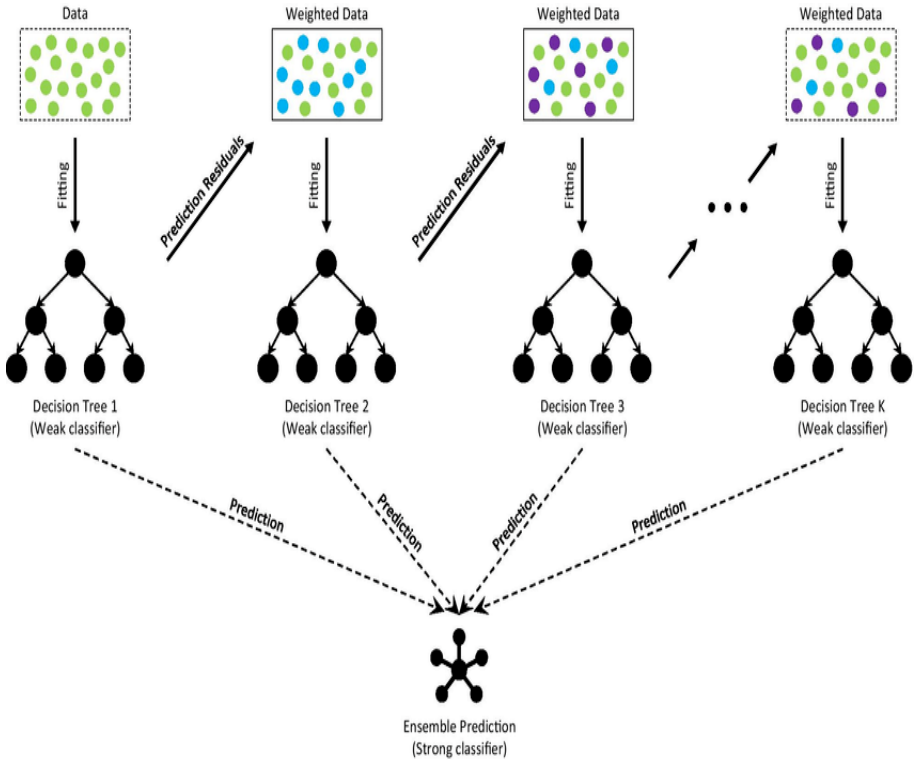

Figura 2.10: Exemplo do funcionamento de GB (Deng et al. 2021)

2.4.12 XGBoost

Extreme Gradient Boosting (XGBoost) é um algoritmo de ML supervisionado, baseado em árvores de decisão, desenvolvido a partir do conceito *Gradient Boosting*. Este conceito é uma técnica de aprendizagem que consiste no desenvolvimento de vários modelos, onde cada modelo visa corrigir ou melhorar as deficiências do modelo anterior, melhorando assim a precisão geral do modelo final (Horvat, Hava e Srpak 2020).

XGBoost utiliza uma fórmula especial para aprender, constituída por dois componentes importantes: uma função que mede o quão erradas são as previsões relativamente aos valores reais, denominada função de perda, e um termo de regularização para combater *overfitting*, através da atribuição de uma penalização com base na complexidade do modelo (Ting et al. 2020).

Na Figura 2.11 é apresentado um exemplo simples do funcionamento deste algoritmo. O objetivo deste exemplo é classificar se uma pessoa irá gostar de um hipotético videojogo. Os membros de uma família são classificados e é atribuída a cada um uma pontuação em diferentes árvores. Na primeira árvore, é avaliada a idade do membro da família, enquanto que na segunda árvore é verificado se esse membro da família utiliza o computador diariamente. A previsão final é calculada pela soma das previsões das duas árvores (Developers s.d.).

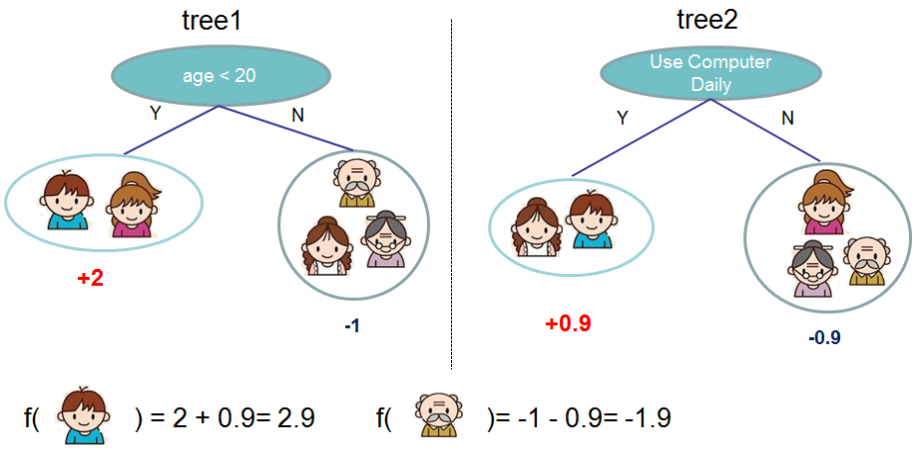

Figura 2.11: Exemplo de funcionamento *XGBoost* (Developers s.d.)

2.4.13 AdaBoost

Adaboost é um algoritmo, pertencente à classe dos algoritmos de *Boosting*, que combina vários classificadores simples (chamados de classificadores fracos) para criar um classificador forte e mais preciso (Hornyák e Iantovics 2023). Os classificadores fracos são treinados de forma repetida em diferentes subconjuntos dos dados, sendo que a cada repetição, é dada mais atenção aos exemplos que foram mal classificados anteriormente, ajustando os seus pesos, de forma a que os classificadores seguintes se concentrem nesses casos mais difíceis (Medium 2023a). O funcionamento do algoritmo, passo a passo, é o seguinte (Medium 2023a):

1. O algoritmo recebe os dados e atribui pesos iguais a todos os exemplos de treino no conjunto de dados. Estes pesos representam a importância de cada exemplo durante o processo de treino.
2. O algoritmo itera com alguns algoritmos durante um número especificado de iterações (ou até que um critério de paragem seja alcançado). O algoritmo treina um classificador fraco nos dados de treino. Este classificador fraco pode ser considerado um modelo que tem um desempenho ligeiramente melhor do que o palpite aleatório.
3. Durante cada iteração, o algoritmo treina o classificador fraco nos dados de treino com os pesos dos exemplos atuais. O objetivo do classificador fraco é minimizar o erro de classificação, ponderado pelos pesos dos exemplos.
4. Após o treino do classificador fraco, o algoritmo calcula o peso do classificador com base nos erros cometidos, sendo que um classificador fraco com um erro menor recebe um peso maior.
5. Uma vez completado o cálculo dos pesos, o algoritmo atualiza os pesos dos exemplos e atribui pesos mais altos aos exemplos mal classificados para que recebam mais atenção nas iterações seguintes.
6. Após a atualização dos pesos dos exemplos, estes são normalizados para que a soma seja igual a 1. As previsões de todos os classificadores fracos são combinadas usando um voto ponderado. Note-se, ainda, que os pesos dos classificadores fracos são considerados na previsão final.

7. Por fim, os passos 2 a 6 são repetidos pelo número especificado de iterações (ou até que o critério de paragem seja alcançado), atualizando os pesos dos exemplos a cada iteração. A previsão final é obtida agregando as previsões de todos os classificadores fracos com base nos seus pesos

O comportamento deste algoritmo está ilustrado na Figura 2.12

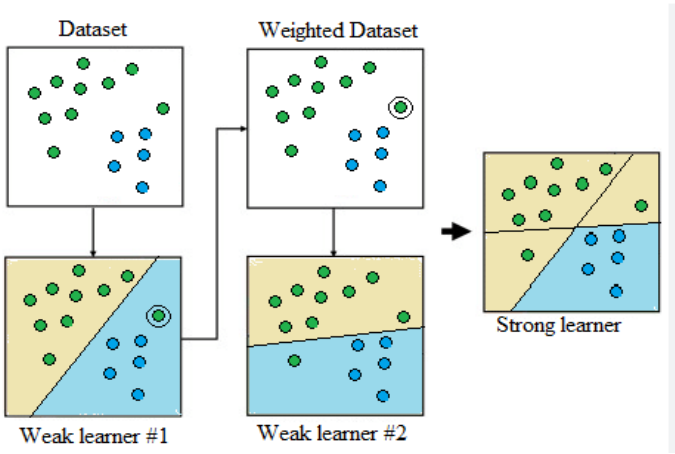

Figura 2.12: Exemplo do funcionamento de *AdaBoost* (Imtiaz Khan et al. 2020)

2.4.14 Light Gradient-Boosting Machine

Light Gradient-Boosting Machine (LGBM) é um poderoso algoritmo de ML utilizado não só em problemas de regressão, como também, em problemas de classificação (Aziz et al. 2020). Este algoritmo é altamente eficiente, pois faz uso de GB, de forma a ajustar gradualmente os pesos dos dados para reduzir os erros (Ramalingam et al. 2024).

Uma das suas principais características é a construção eficiente de árvores de decisão, o que leva a melhor gestão de memória e velocidades de treino mais rápidas (Ramalingam et al. 2024) (LightGBM s.d.). Neste algoritmo, os ramos são adicionados às árvores de decisão apenas onde são mais necessários, utilizando uma técnica conhecida como *leaf-wise growth* (Figura 2.13).

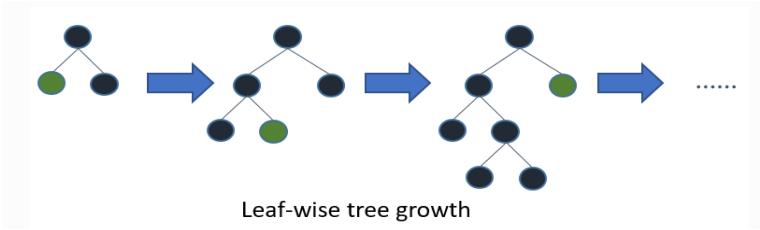

Figura 2.13: Exemplo do funcionamento de LGBM (LightGBM s.d.)

Em vez de expandir todas as folhas de forma uniforme em cada nível da árvore, como ocorre na técnica tradicional de *level-wise growth*, o LGBM seleciona, para ser expandida, a folha que resulta na maior redução de erro (Ramalingam et al. 2024) (LightGBM s.d.). Esta

abordagem permite que o modelo concentre os seus recursos nas áreas mais importantes, resultando em árvores mais profundas em partes específicas e, consequentemente, em modelos mais precisos e eficientes.

Capítulo 3

Preparação de Dados

3.1 Conjunto de dados

O conjunto de dados utilizado nesta tese é o *NBA games data* (Lauga s.d.) e é composto por cinco ficheiros distintos, cada um contendo informações detalhadas sobre diferentes aspetos dos jogos da NBA:

- **Dados de Jogos da NBA:** Este ficheiro contém informações detalhadas sobre os jogos da NBA realizados desde a temporada 2003/2004 até à temporada 2022/2023. Os dados incluem detalhes como a data do jogo, as equipas envolvidas, as estatísticas coletivas das equipas e se a equipa da casa foi a vencedora.
- **Informações dos Jogadores:** Este ficheiro contém os nomes dos jogadores, a equipa onde jogam e a época à qual o registo pertence. Este ficheiro não foi utilizado, pois não continha qualquer informação relevante
- **Ranking das Equipas por Dia:** Este ficheiro apresenta as informações do *ranking* de cada equipa, em cada dia, ao longo da temporada. Os *rankings* são atualizados diariamente, refletindo o desempenho contínuo das equipas ao longo da temporada. Contém dados como o número de jogos, o número de vitórias e o número de derrotas da equipa na presente temporada, não só no geral, como, também, apenas para a qualidade de visitante e de visitado. Estes dados são importantes, pois permitem inferir a força das equipas, pois equipas mais fortes possuem um rácio entre vitórias e derrotas maior enquanto que, nas equipas mais fracas, o oposto se verifica.
- **Informações das Equipas:** Este ficheiro contém informações gerais sobre as equipas, incluindo o nome da equipa, cidade sede, ano de fundação, estádio e respetiva capacidade, o dono da equipa, o *General Manager* (GM), o atual treinador e o nome da equipa B.
- **Estatísticas dos Jogadores por Jogo:** Este ficheiro inclui as estatísticas individuais de cada jogador para cada jogo realizado. As estatísticas abrangem diversos parâmetros de desempenho, tais como pontos marcados, assistências, ressaltos, minutos jogados, eficiência de lançamento, entre outros. Estes dados são essenciais, pois permitem a inclusão do fator de *performance* individual no modelo.

O conjunto de dados cobre um total de 20 épocas. Inclui informações sobre 26622 jogos realizados ao longo deste período e abrange estatísticas de 2686 jogadores diferentes, além das 30 equipas da NBA. Estes números fornecem uma base sólida para análises estatísticas e modelagem preditiva, permitindo um visão detalhada das dinâmicas e desempenhos na NBA ao longo de duas décadas.

Estes cinco ficheiros, quando combinados, fornecem uma visão abrangente e detalhada da dinâmica da NBA, ao longo de duas décadas, permitindo uma análise aprofundada dos dados, tanto a nível individual, quanto coletivo. A constituição destes ficheiros pode ser consultada no anexo A (Tabela A.1, Tabela A.2, Tabela A.3, Tabela A.4 e Tabela A.5)

Para além disso, o Kaggle (Kaggle s.d.) possui uma classificação dos seus *datasets*, sendo que o *dataset* escolhido possui uma classificação de 8.82 pontos em 10 possíveis. Nesta classificação, destaca-se a pontuação máxima para o critério "Credibilidade".

A escolha da utilização deste *dataset*, nesta tese, recaiu sobre os pontos acima mencionados.

Nos subcapítulos seguintes, é explicada a metodologia seguida para a preparação dos dados e treino dos modelos de previsão.

3.2 Tratamento dos dados

Inicialmente, foi feito o tratamento dos dados de cada ficheiro de forma individual. A constituição destes ficheiros, após a limpeza e tratamento, pode ser consultada no anexo A (Tabela A.6 , Tabela A.7, Tabela A.8 e Tabela A.9)

3.2.1 Jogos da NBA

A coluna *GAME_STATUS_TEXT* representa o estado do jogo. Visto que continha sempre o mesmo valor (Tabela 3.1), foi removida.

Tabela 3.1: Distribuição da coluna *GAME_STATUS_TEXT*

Valor	Frequência absoluta	Percentagem
Final	26651	100%

As colunas *HOME_TEAM_ID* e *TEAM_ID_home*, apesar de terem nomes diferentes, pareciam conter exatamente a mesma informação, isto é, o identificador da equipa da casa. Deste modo, foi feita uma verificação, de modo a averiguar se tal se sucedia, o que acabou por se confirmar. Com isto em mente, foi feita a remoção da coluna *TEAM_ID_home*. O mesmo se verificou para mais duas colunas, *VISITOR_TEAM_ID* e *TEAM_ID_away*, que representam o identificador da equipa visitante, pelo que a coluna *TEAM_ID_away* foi, também, removida.

O próximo passo foi verificar valores em falta. Foram encontradas 99 linhas onde se verificava que as estatísticas das equipas não estavam preenchidas. Por se tratar de uma pequena amostra relativamente ao *dataset* completo e por estas linhas serem relativas a jogos de pré-epoca, prosseguiu-se com a remoção delas.

Continuou-se com a verificação de chaves repetidas, ou seja, se havia duas ou mais linhas com o mesmo identificador de jogo. Como se verificou que existiam linhas com o mesmo identificador, porém, não exatamente duplicadas devido a algumas colunas que, apesar de terem o mesmo valor, apresentavam um número diferente de casas decimais, realizou-se o arredondamento para 3 casas decimais dessas mesmas colunas. Para além disso, foram, também, encontradas linhas com identificador repetido, porém com valores distintos em algumas estatísticas. Recorrendo ao *website* oficial da NBA (NBA s.d.) foi possível identificar

quais as linhas corretas e proceder à remoção das linhas com valores errados. Após isto, foram removidas as restantes linhas duplicadas.

De seguida, foi verificado se haveria valores com casas decimais em colunas de estatísticas que apenas poderiam ser um número inteiro, como por exemplo pontos marcados ou ressaltos. O mesmo não se verificou.

Foram feitos os gráficos *boxplot* para as colunas representativas das estatísticas dos jogos e, através da análise destes gráficos, foram descobertos alguns jogos onde algumas estatísticas eram extremamente baixas ou extremamente altas, nomeadamente os pontos marcados pela equipa da casa e pela equipa visitante (Figura 3.1 e Figura 3.2).

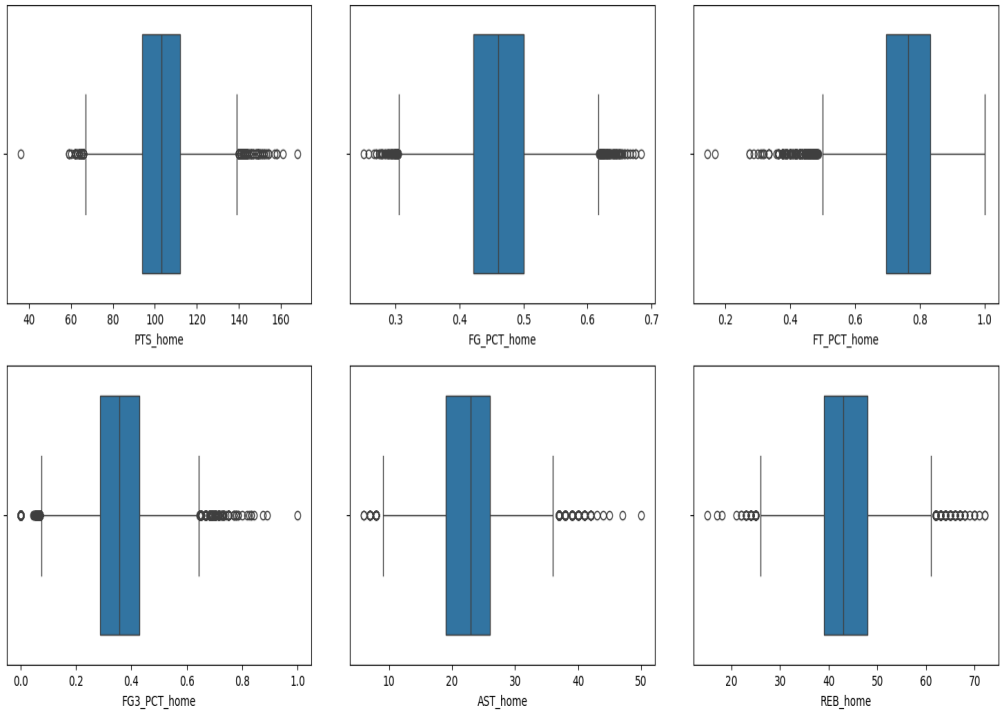

Figura 3.1: Boxplots para as estatísticas da equipa da casa

Recorrendo ao *website* oficial da NBA (NBA s.d.), foi possível perceber que, no caso das estatísticas extremamente altas, se tratavam de jogos onde existiram um ou mais prolongamentos, enquanto que, no caso das estatísticas extremamente baixas, se tratavam de jogos de pré-época onde a duração tinha sido de apenas 30 minutos, ao invés dos 48 minutos habituais. De forma a distinguir os jogos de pré-epoca dos jogos a valer, foi criada a coluna *IS_PRESEASON_GAME* com auxílio dos *websites RealGM* (RealGM s.d.) e *Basketball Reference* (Reference s.d.[a]), onde estão disponibilizados os calendários das épocas da NBA. Os jogos de pré-época servem, maioritariamente, para testar novas táticas e avaliar o potencial de jogadores recentemente adquiridos, num menor ritmo e com uma carga de trabalho mais baixa. Para além disso, os melhores jogadores têm por hábito limitar a sua participação nestes jogos, de forma a não arriscarem lesões que comprometam o início da época desportiva (AS 2023). Como se verificou que os jogos de pré-época eram uma amostra muito pequena, como é possível verificar na Tabela 3.2, e aliado ao facto de a dinâmica destes jogos ser bastante diferente, tomou-se a decisão de remover estes jogos e, consequentemente, a coluna *IS_PRESEASON_GAME*.

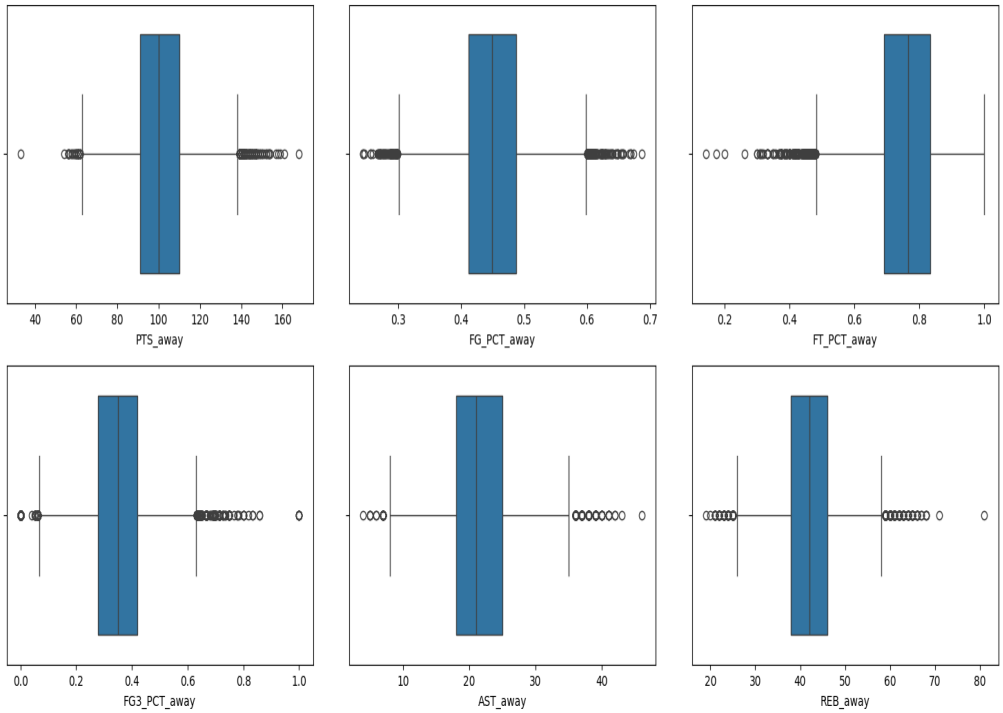

Figura 3.2: Boxplots para as estatísticas da equipa visitante

Tabela 3.2: Distribuição da coluna *IS_PRE_SEASON_GAME*

Valor	Frequência absoluta	Percentagem
Não	24881	93,8%
Sim	1642	6,2%

Foi verificado se existia desbalanceamento na quantidade de jogos ganhos pela equipa da casa e o número de jogos ganhos pela equipa visitante. Foi possível concluir que, no conjunto de dados, a equipa da casa havia ganho 58,7% das vezes, enquanto que a equipa visitante tinha sido a vencedora em 41,3% das vezes (Tabela 3.3).

Tabela 3.3: Distribuição da coluna *HOME_TEAM_WINS*

Valor	Frequência absoluta	Percentagem
Sim	14700	58,7%
Não	10181	41,3%

O fator casa na NBA é um elemento crucial que influencia significativamente o desempenho das equipas. A vantagem de jogar em casa é acentuada pela energia e entusiasmo do público local, que pode motivar a equipa da casa em momentos menos bons do jogo e intimidar os jogadores da equipa adversária. Esta vantagem é reforçada pelo fator familiaridade com o campo de jogo, onde a equipa da casa está mais acostumada às especificidades do seu terreno (News 2023). Adicionalmente, acredita-se que o impacto psicológico de jogar em casa cria um efeito placebo, onde os jogadores acreditam que essa vantagem contribui para o seu sucesso. Estatísticas mostram que, na NBA, a vantagem de jogar em casa é a mais significativa entre os quatro grandes desportos americanos, com as equipas a vencerem cerca

de 60% dos jogos da época regular e 64,9% dos jogos dos *playoffs* que realizam na condição de equipa da casa. Além disso, é colocada a hipótese de viés dos árbitros em relação à equipa da casa, tendendo a favorecê-la inconscientemente, devido à influência dos adeptos presentes na arena (Report 2013a). Visto que a distribuição entre a quantidade de vitórias entre a equipa da casa e a equipa visitante estava em linha com os dados apresentados e, de forma a replicar o fator casa, tomou-se a decisão de não balancear o conjunto de dados.

Para cada linha deste ficheiro, isto é, cada jogo realizado, estão representadas as estatísticas de ambas as equipas (pontos marcados, percentagem de acerto de lançamentos, percentagem de acerto de lançamentos livres, percentagem de acerto de lançamentos de 3 pontos, assistências e ressaltos). Visto que ao prever jogos que ainda não aconteceram, será impossível ter acesso a estes dados, foi necessário realizar uma transformação nestas colunas, de forma a não só representarem o atual momento de forma das equipas, como, também, serem informações que estejam disponíveis antes do jogo ocorrer. Deste modo, para cada uma destas estatísticas, juntando, também, os pontos sofridos, foi calculada a sua média nos últimos 4 jogos, dando origem às colunas :

- **AVG_POINTS_LAST_4_GAMES_HOME_TEAM:** Média de pontos marcados pela equipa da casa nos últimos 4 jogos.
- **AVG_POINTS_LAST_4_GAMES_VISITOR_TEAM:** Média de pontos marcados pela equipa visitante nos últimos 4 jogos.
- **AVG_ASSISTS_LAST_4_GAMES_HOME_TEAM:**Média de assistências feitas pela equipa da casa nos últimos 4 jogos.
- **AVG_ASSISTS_LAST_4_GAMES_VISITOR_TEAM:** Média de assistências feitas pela equipa visitante nos últimos 4 jogos.
- **AVG_FGPCT_LAST_4_GAMES_HOME_TEAM:** Média de percentagem de acertos de lançamentos da equipa da casa nos últimos 4 jogos.
- **AVG_FGPCT_LAST_4_GAMES_VISITOR_TEAM:** Média de percentagem de acertos de lançamentos da equipa visitante nos últimos 4 jogos.
- **AVG_FTPCT_LAST_4_GAMES_HOME_TEAM:** Média de percentagem de acertos de lançamentos livres da equipa da casa nos últimos 4 jogos.
- **AVG_FTPCT_LAST_4_GAMES_VISITOR_TEAM:** Média de percentagem de acertos de lançamentos livres da equipa visitante nos últimos 4 jogos.
- **AVG_FG3PCT_LAST_4_GAMES_HOME_TEAM:** Média de percentagem de acertos de lançamentosde 3 pontos da equipa da casa nos últimos 4 jogos.
- **AVG_FG3PCT_LAST_4_GAMES_VISITOR_TEAM:** Média de percentagem de acertos de lançamentos livres da equipa visitante nos últimos 4 jogos.
- **AVG_REB_LAST_4_GAMES_HOME_TEAM:** Média de ressaltos da equipa da casa nos últimos 4 jogos.
- **AVG_REB_LAST_4_GAMES_VISITOR_TEAM:** Média de ressaltos da equipa visitante nos últimos 4 jogos.
- **AVG_POINTS_CONCEDED_LAST_4_GAMES_HOME_TEAM:** Média de pontos sofridos pela equipa da casa nos últimos 4 jogos.

- **AVG_POINTS_CONCEDED_LAST_4_GAMES_VISITOR_TEAM:** Média de pontos sofridos pela equipa visitante nos últimos 4 jogos.

Para além disso, foram também criadas colunas onde estas médias foram calculadas apenas contabilizando os últimos 4 jogos em casa para a equipa da casa e os últimos 4 jogos fora para a equipa visitante.

A escolha relativamente ao número de jogos a utilizar para o cálculo das médias baseou-se no trabalho de (Chen et al. 2021), que concluiu que o número de jogos que apresentava melhores resultados era 4.

Apesar de terem sido encontrados *outliers* nos dados durante a visualização dos *boxplots* e de estes serem dados significativamente diferentes que podem distorcer a análise e comprometer suposições (obviously.ai 2022), optou-se por não os remover. A remoção dos jogos onde esses *outliers* ocorrem afetaria o cálculo da média das estatísticas nos últimos quatro jogos, comprometendo a integridade dos dados. Assim, manter estes *outliers* permite uma análise mais fiel à realidade.

3.2.2 Informações das Equipas

Após a conclusão da limpeza e tratamento dos dados do ficheiro que contem as informações sobre os jogos, procedeu-se, então, à mesma operação para o ficheiro relativo às informações sobre as equipas.

As colunas *LEAGUE_ID* e *MAX_YEAR* foram removidas, pois continham sempre o mesmo valor.

As colunas *ABBREVIATION*, *NICKNAME*, *ARENA* e *CITY* que representam a abreviatura da equipa, a sua alcunha, o nome da arena onde joga e a cidade a que pertence, respetivamente, foram removidas pelo facto de não possuírem qualquer valor semântico.

As colunas *MIN_YEAR* e *YEARFOUNDED* aparentavam ambas representar o ano de fundação da equipa. Uma verificação revelou que o seu conteúdo era exatamente igual, levando à remoção da coluna *MIN_YEAR*.

Uma característica presente neste ficheiro era a capacidade das arenas onde as equipas atuam. Equipas que jogam numa arena com maior capacidade poderiam beneficiar de um maior fator casa. Porém, verificou-se que as capacidades das arenas eram bastante similares, sendo a diferença entre a arena de menor capacidade e a arena de maior capacidade de apenas cerca de 4000 lugares. Para além disso, foram também encontrados valores ausentes nesta coluna e, para uma das equipas, à capacidade da sua arena estava incorretamente atribuído o valor 0. A distribuição dos valores desta coluna pode ser visualizada no gráfico de dispersão ilustrado na Figura 3.3. Pelos motivos apresentados a coluna foi eliminada.

Foi verificada a presença de linhas duplicadas e de valores em falta. Nenhum dos cenários se verificou.

A coluna *DLEAGUEAFFILIATION* contem o nome da equipa B, caso exista. A partir desta informação, foi criada uma nova coluna *booleana*, *'HAS_DLEAGUE_TEAM'*, cujo objetivo era representar se uma equipa possuía equipa B. A distribuição desta nova coluna foi analisada, estando a sua distribuição demonstrada na Tabela 3.4. Como apenas 3 das 30 equipas não possuíam uma equipa B, a coluna foi removida por conter predominantemente o mesmo valor. A coluna original também foi removida devido à sua falta de valor semântico.

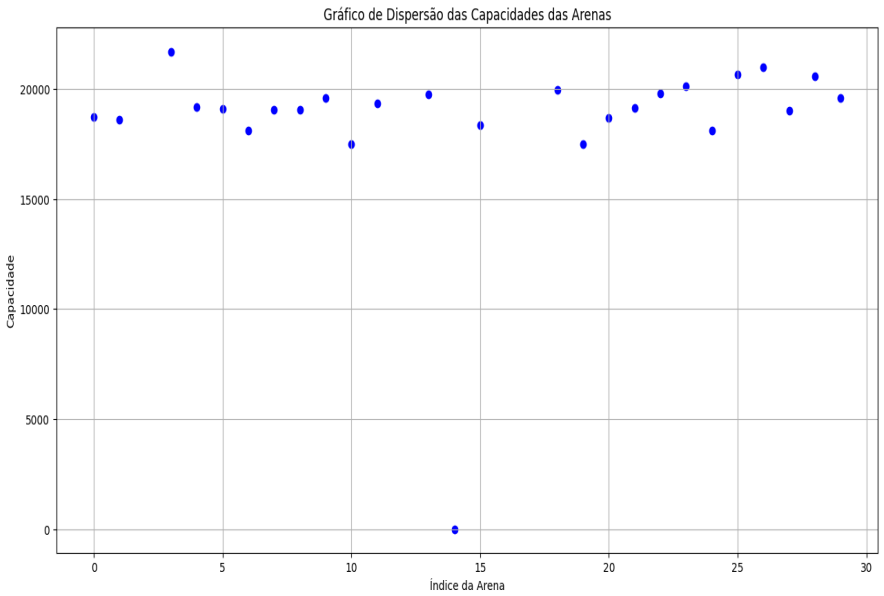

Figura 3.3: Gráfico de disperção para a coluna ARENACAPACITY

Tabela 3.4: Distribuição da coluna *HAS_DLEAGUE_TEAM*

Valor	**Frequência absoluta**	**Percentagem**
Sim	27	90%
Não	3	10%

Na NBA, os treinadores desempenham um papel crucial no desempenho das equipas. A capacidade de um treinador desenvolver estratégias eficazes, ajustar táticas durante os jogos, desenvolver jogadores e criar química de equipa pode determinar a diferença entre uma equipa medíocre e uma equipa de sucesso (Sportskeeda 2023) (HoopSocial 2023). Deste modo, a coluna *HEADCOACH*, que representa o treinador atual da equipa, aparentava ser valiosa. Contudo, como a coluna representava apenas o treinador atual, não poderia ser usada de forma precisa para analisar toda a abrangência temporal do *dataset*, que vai desde a temporada 2003-2004 até à mais recente, 2022-2023. Durante esse período, houve várias trocas de treinadores nas equipas. Para obter informações precisas sobre estas trocas entre as temporadas e até mesmo com as temporadas em curso, foram realizadas pesquisas em vários *websites*. Apesar de ter sido possível encontrar informações sobre as trocas de treinadores, as datas específicas em que essas trocas ocorreram não estavam disponíveis, o que seria um obstáculo na introdução de informação precisa sobre os treinadores no *dataset*. Por este motivo, esta coluna acabou por ser removida.

O GM desempenha um papel crucial na *performance* das equipes da NBA, sendo responsável por operações fundamentais como a seleção de treinadores, a observação de talentos e as negociações contratuais (Wong e Deubert 2010). Estas decisões impactam diretamente a composição da equipa, o que, por sua vez, afeta o desempenho da mesma. A capacidade do GM de montar uma equipa equilibrada e competitiva é essencial para o sucesso dentro e fora da quadra. Tal como o treinador, o GM aparentava ser uma característica importante. No entanto, foi encontrado o mesmo desafio encontrado na análise do treinador, que era

o facto de apenas ter a informação do GM atual. A metodologia seguida foi a mesma e o resultado, também. Deste modo, a coluna com esta informação, *GENERALMANAGER*, foi eliminada.

A *performance* das equipas na NBA é fortemente influenciada pela qualidade da gestão dos seus donos. Uma liderança sólida e visionária pode transformar uma equipa, promovendo uma cultura positiva e fazendo mudanças estratégicas que levam ao sucesso. Em contraste, uma má gestão, caracterizada por decisões impulsivas e falta de delegação eficaz, pode resultar em mau desempenho e fracasso a longo prazo (Report 2013b). Tal como o treinador e o GM, o dono da equipa aparentava ser uma característica importante, porém foi encontrado o mesmo obstáculo. Deste modo, a coluna *OWNER* foi removida.

3.2.3 Ranking das equipas

Após a conclusão da limpeza e tratamento dos dados do ficheiro que contem as informações sobre as equipas, procedeu-se, então, à mesma operação para o ficheiro relativo às informações sobre o *ranking*.

A coluna que representa o identificador da liga, *LEAGUE_ID*, tal como no ficheiro anterior, tinha um valor constante, pelo que a coluna foi removida. A coluna *RETURNTOPLAY*, que indica se a data era uma data de regresso à competição após paragem, foi também removida, pois era maioritariamente constituída pelo mesmo valor, como é possível verificar na Tabela 3.5.

Tabela 3.5: Distribuição da coluna *RETURNTOPLAY*

Valor	Frequência absoluta	Percentagem
Valor em falta	206352	98,1%
Sim	2394	1,1%
Não	1596	0.8%

De seguida, procedeu-se com a remoção da coluna *TEAM*, visto que o nome da equipa não tem qualquer correlação com o desempenho ao longo da temporada, isto é, o nome não tem qualquer valor semântico.

Procedeu-se com a verificação de valores em falta, linhas duplicadas, linhas com a mesma chave (ou seja, dois *rankings* para uma mesma equipa, num mesmo dia) e se existiam valores com casas decimais em colunas de estatísticas que apenas poderiam ser um número inteiro, como o número de jogos, número de vitórias e número de derrotas. Apenas se verificou o cenário das linhas duplicadas, linhas essas que foram removidas.

Visto que, as colunas *HOME_RECORD* e *ROAD_RECORD* estavam no formato Vitórias-Derrotas, foi necessário separar cada uma destas colunas em duas. Deste modo, foram originadas as colunas *HOME_WINS*, *HOME_LOSSES*, *ROAD_WINS* e *ROAD_LOSSES*, que representam o número de vitórias em casa, o número de derrotas em casa, o número de vitórias fora de casa e o número de derrotas fora de casa, respetivamente. As duas colunas originais foram apagadas. Após esta transformação, foi verificado, através de *box-plots* (Figura 3.4) se existiam valores incongruentes nas colunas referentes ao registo das equipas, isto é, se existiam valores negativos para as colunas referentes ao número de jogos, percentagem de vitórias, vitórias e derrotas. O mesmo não se verificou.

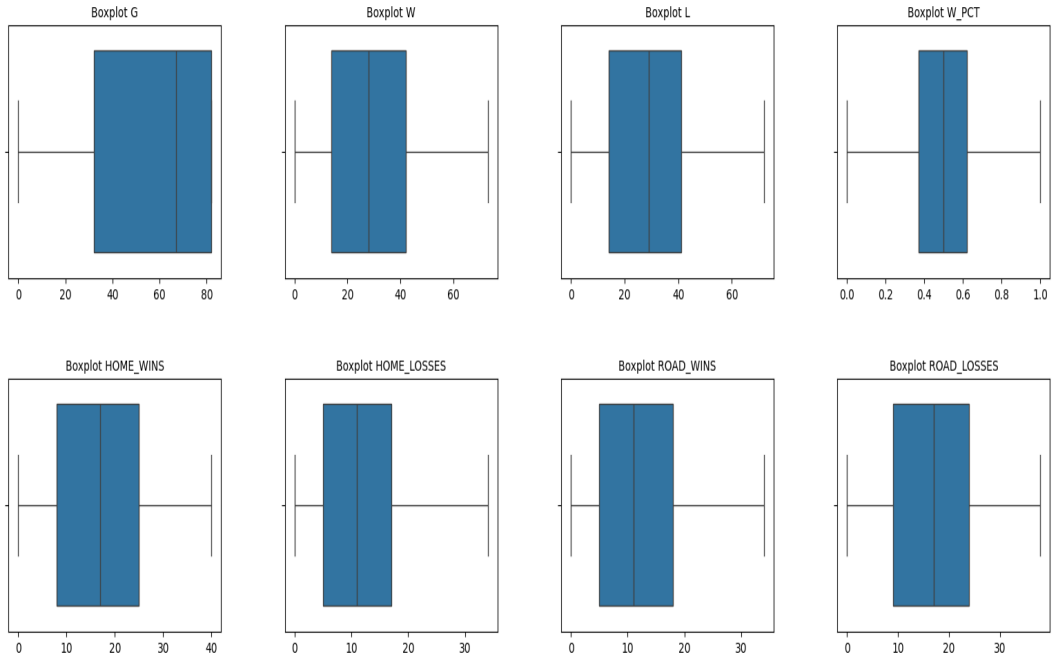

Figura 3.4: *Boxplots* para o registo das equipas

A coluna *CONFERENCE* representa a conferência onde a equipa joga, "Este"ou "Oeste". Visto ser uma coluna com texto, foi necessário obter *dummies*, dando origem a duas novas colunas *IS_EAST_CONFERENCE* e *IS_WEST_CONFERENCE*. A coluna original foi eliminada.

O *ranking* de um dia já inclui os jogos realizados naquele dia. Portanto, o *ranking* de uma equipa ao entrar em campo deve ser o *ranking* do dia anterior ao jogo. No primeiro dia de cada temporada, como o *ranking* do dia anterior era baseado nas informações da pré-temporada, foi necessário colocar a 0 todos os valores referentes ao número de jogos, vitórias e derrotas.

3.2.4 Estatísticas dos Jogadores

Inicialmente, foram eliminadas as colunas relacionadas com o nome da equipa e com o nome do jogador (*TEAM_ABBREVIATION*, *TEAM_CITY*, *PLAYER_NAME* e *NICKNAME*) por não terem qualquer valor semântico.

A coluna *COMMENT* contem uma explicação para a ausência de um jogador do jogo, caso, de facto, ele não tenha jogado. Deste modo, de forma a eliminar registos de jogadores que não tenham participado da partida, todas as linhas onde esta coluna estivesse preenchida foram eliminadas. Visto que, com esta operação, esta coluna passou a ser constituída apenas por valores em falta, foi eliminada.

Seguidamente, as colunas que correspondem a percentagens de acerto de lançamentos, (*FG_PCT*, *FG3_PCT* e *FT_PCT*) foram arredondadas para três casas decimais.

Procedeu-se com a verificação de colunas duplicadas e consequente remoção e com a verificação de linhas com chave repetida, isto é, linhas com os mesmos identificadores de jogo e de jogador. Foram encontradas algumas linhas com chave duplicadas, porém, com diferentes valores. Com auxílio dos *websites* oficiais da NBA (NBA s.d.) e da ESPN (ESPN s.d.) foi possível identificar as linhas corretas e proceder com a eliminação das incorretas.

Cada equipa é composta por dois bases, dois extremos e um poste. Portanto, num jogo da NBA, devem começar quatro bases, quatro extremos e dois postes. Tendo isto em conta, foi realizada uma verificação para garantir que todos os jogos cumprissem estes requisitos. Foram identificados alguns jogos que não atendiam a estas condições e, após análise, concluiu-se que eram jogos de pré-época. Como os jogos desta tipologia já haviam sido removidos do ficheiro de dados dos jogos, procedeu-se, também, à remoção destes jogos neste ficheiro.

A operação anterior foi validada, analisando a distribuição da coluna *START_POSITION*. O valor "C"(poste) deveria ter metade das ocorrências do valor "G"(base) e "F"(extremo), e as ocorrências destes dois últimos deveriam corresponder a um quarto do número total de jogos contidos no ficheiro. Esta distribuição foi confirmada.

Os valores NaN nesta coluna representam jogadores que participaram no jogo, mas que não foram titulares. Para preencher estes valores em falta e atribuir-lhes um significado, procedeu-se ao preenchimento com o valor "DIDN'T START"(Não começou). A distribuição desta coluna após este tramento pode ser verificada na Tabela 3.6.

Tabela 3.6: Distribuição da coluna *START_POSITION* após tratamento

Valor	**Frequência absoluta**	**Percentagem**
DIDN'T START	279367	52,2%
F	102260	19,1%
G	102260	19,1%
C	51130	9,6%

Foi validado se existiam mais valores em falta no ficheiro e foi encontrada uma linha com este cenário. Foram consultados vários *websites*, mas todos apresentavam valores diferentes para as estatísticas desse jogador nesse jogo. Além disso, o jogo ao qual esse registo pertencia era um jogo de pré-época, pelo que, mantendo o critério anteriormente utilizado, foi eliminado.

De seguida, foi verificado se existiam valores com casas decimais em colunas de estatísticas que apenas poderiam ser um número inteiro, como por exemplo pontos marcados ou assistências. O mesmo não se verificou.

Visto que a coluna representativa dos minutos jogados pelo jogador, *MIN* se encontrava no formato MINUTOS:SEGUNDOS, sendo este um formato de texto, foi necessário realizar uma transformação na coluna para conter apenas números inteiros. Assim, manteve-se apenas o valor dos minutos.

Por fim, foi verificado, através de *boxplots* (Figura 3.5 e Figura 3.6) se existiam valores incongruentes nas colunas referentes às estatísticas dos jogadores, isto é, se existiam valores negativos, visto que, à exceção da estatística *PLUS_MINUS*, não é possível um jogador atingir valores abaixo de 0. O mesmo não se verificou.

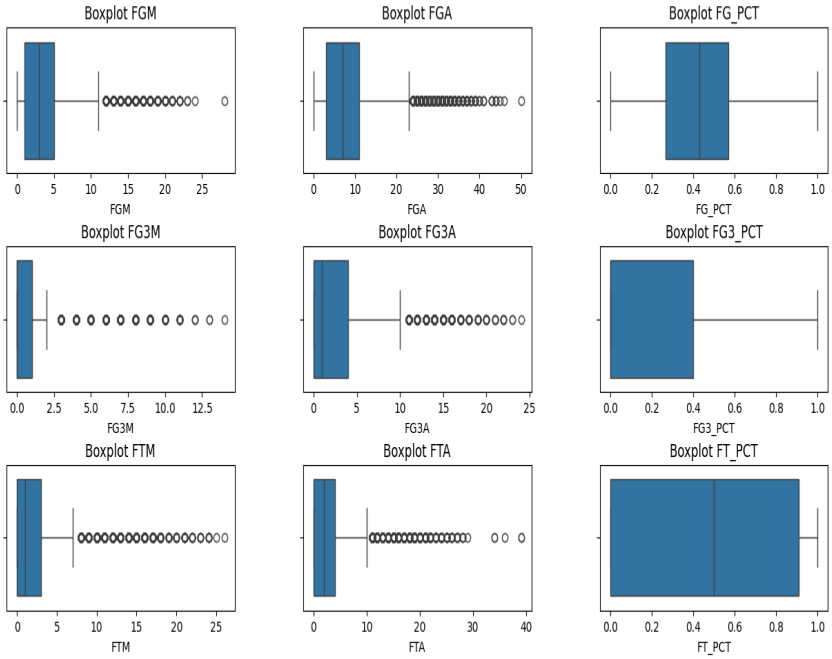

Figura 3.5: Boxplots para as estatísticas referentes a lançamentos

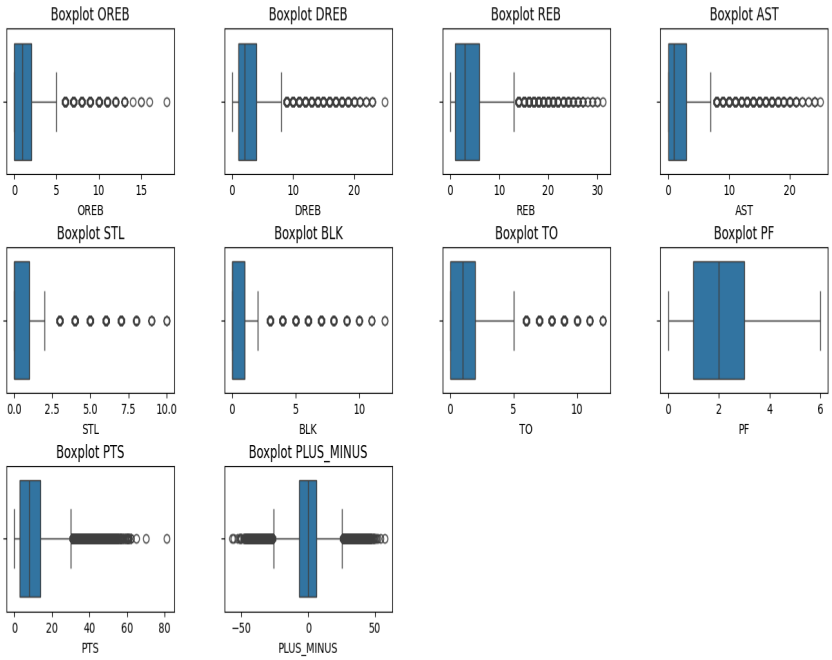

Figura 3.6: Boxplots para as restantes estatísticas

Após a limpeza e tratamento individual dos ficheiros, procedeu-se com a junção do ficheiro dos jogos da NBA com o ficheiro das informações da equipa. A constituição do *dataset*

resultante pode ser consultada na Tabela A.10 presente no Anexo A.

3.2.5 Junção do ficheiro dos rankings

O próximo passo foi unir o ficheiro relacionado com os *rankings* das equipas ao *dataset*.

Ao realizar esta junção, foi notado que o *ranking* das equipas apenas era atualizado até ao fim da época regular. A partir do momento em que se entrava na fase de *playoffs*, o número de jogos, vitórias e derrotas das equipas deixava de sofrer alterações. Deste modo, foi necessário calcular, para cada equipa, o número de jogos, o número de vitórias, o número de derrotas e a percentagem de vitórias, tanto no geral como, especificamente, para os jogos em casa no caso da equipa da casa e para os jogos fora no caso da equipa visitante. Além disso, foram, também, calculadas a atual sequência de vitórias/derrotas de cada equipa no âmbito geral, a sequência de vitórias/derrotas em jogos em casa da equipa da casa e a sequência de vitórias/derrotas fora de casa da equipa visitante. Estas operações foram essenciais para garantir que o *dataset* refletisse, com precisão, o desempenho de cada equipa durante toda a época, e não apenas até ao fim da época regular. A inclusão da sequência de vitórias/derrotas é particularmente importante, pois permite uma análise mais detalhada do momento atual das equipas. Sequências positivas ou negativas podem influenciar significativamente o moral e a confiança das equipas, afetando o seu desempenho.

Neste ponto, foram descobertos alguns dados com erros, através da comparação das colunas originais com as novas colunas criadas. Foi possível perceber que existiam jogos em falta na época 2022/2023. Como apenas havia dados desta época até à data de 22 de dezembro de 2022, e alguns jogos estavam em falta, procedeu-se à remoção dos jogos desta época do *dataset*.

Além disso, devido a uma interrupção na época de 2019/2020 causada pela COVID-19, foi realizada uma pequena pré-época antes do regresso à competição, o que levou a que os registos de *ranking* do dia do regresso à competição incluíssem as informações dessa pré-época. Desse modo, foi necessário corrigir estes registos para que estivessem alinhados com as informações prévias à paragem devido à COVID-19. Esta correção foi crucial para assegurar que os dados de *ranking* refletissem corretamente o desempenho das equipas antes da interrupção, evitando assim qualquer viés introduzido pela pré-época adicional. Para além disso, os jogos relativos a essa pequena pré-epoca foram removidos. A distribuição entre a quantidade de jogos ganhos pela equipa da casa e pela equipa visitante, após estas correções, pode ser visualizada na Tabela 3.7.

Tabela 3.7: Distribuição da coluna *HOME_TEAM_WINS* após junção do ficheiro dos *rankings*

Valor	Frequência absoluta	Percentagem
Sim	14394	59,1%
Não	9977	40,9%

A constituição do *dataset* resultante após esta junção pode ser consultada na Tabela A.11 presente no anexo A.

3.2.6 Cansaço das equipas

(Zheng 2022) demonstrou a importância do cansaço dos jogadores no desempenho das equipas.

Deste modo, foi introduzido o fator de cansaço no *dataset*. Este fator foi incluído através do cálculo das seguintes características:

- **Número de jogos nos últimos 7, 10 e 14 dias**: Esta métrica é fundamental para avaliar a intensidade do calendário de uma equipa e quantificar o desgaste físico acumulado num curto espaço de tempo. Equipas com um maior volume de jogos num curto espaço de tempo podem apresentar sinais de cansaço que afetam o desempenho.
- **Número de jogos fora nos últimos 7, 10 e 14 dias:** Viagens frequentes para jogos fora de casa podem aumentar o cansaço das equipas devido ao tempo gasto em deslocações e à falta de descanso adequado. Esta métrica ajuda a identificar equipas que podem estar em desvantagem devido a um calendário exigente de jogos fora.
- **Atual sequência de jogos fora de casa seguidos:** Jogar consecutivamente fora de casa pode ser particularmente desgastante, pois as equipas não têm a oportunidade de recuperar no seu ambiente habitual e precisam de viajar de forma consecutiva. Esta característica permite identificar períodos críticos onde o desempenho da equipa pode ser afetado pelo cansaço acumulado e pelas constantes deslocações.
- **Média de jogadores utilizados nos últimos 4 jogos:** A rotação de jogadores é uma estratégia usada para mitigar o cansaço. Equipas que utilizam mais jogadores podem gerir melhor a carga de trabalho e manter um desempenho mais consistente. Esta métrica avalia a profundidade do plantel e a capacidade de gerir o cansaço.
- **Média de minutos por jogador nos últimos 4 jogos:** Jogadores que acumulam muitos minutos de jogo consecutivos tendem a ficar mais fatigados, o que pode prejudicar o seu desempenho. Esta característica ajuda a identificar equipas que estão sob maior pressão física.
- **Média de minutos jogados pelos titulares nos últimos 4 jogos:** Os titulares normalmente jogam mais minutos e têm um impacto maior no desempenho da equipa. Monitorar os minutos jogados pelos titulares é crucial para avaliar o nível de cansaço dos principais jogadores da equipa.

Com a inclusão destas características no *dataset*, é possível uma análise mais completa e precisa do impacto do cansaço no desempenho das equipas, ajudando a identificar padrões e prever possíveis quedas de rendimento. Na tabela A.12, presente no anexo A está representada a constituição do *dataset* resultante.

3.2.7 Inclusão de estatísticas de equipa que não estavam presentes no ficheiro dos jogos

Apesar do ficheiro de informações dos jogos incluir as estatísticas das equipas, algumas estatísticas importantes estavam em falta. No entanto, foi possível calcular essas estatísticas somando os valores correspondentes de cada jogador que participou no jogo. Esta soma foi realizada utilizando o ficheiro de estatísticas dos jogadores, que continha, não só as estatísticas presentes no ficheiro dos jogos, como também, outras estatísticas adicionais. Através deste ficheiro foram calculadas as seguintes estatísticas:

- **Ressaltos ofensivos e ressaltos defensivos:** embora o ficheiro com as informações dos jogos já incluísse o total de ressaltos das equipas, optou-se por separar os ressaltos em duas características distintas. Os ressaltos ofensivos são cruciais, pois proporcionam uma segunda oportunidade de pontuar, mesmo após um lançamento falhado, aumentando, assim, as probabilidades de sucesso (Koutsouridis et al. 2020). Por outro lado, os ressaltos defensivos permitem que uma equipa recupere a posse da bola, reduzindo as tentativas de pontuar da equipa adversária e abrindo caminho para um contra-ataque (Koutsouridis et al. 2020). Em suma, os ressaltos ofensivos e defensivos, são indicadores da capacidade ofensiva e defensiva, respetivamente, das equipas.
- **Roubos de bola:** os roubos de bola, não só proporcionam mais posses ofensivas, o que pode resultar em mais pontos, como, também, interrompem a posse de bola do adversário, permitindo contra-ataques (FiveThirtyEight 2014) (Student s.d.). Estes são indicadores da capacidade defensiva de uma equipa.
- **Bloqueios:** interrompem a tentativa de lançamento do adversário. Permitem averiguar a capacidade de uma equipa em defender o seu cesto.
- **Perdas de posse de bola:** quando uma equipa perde a posse de bola, desperdiça uma oportunidade de pontuar. Para além disso, oferece ao adversário a possibilidade de pontuar, através de transições rápidas. Equipas com um número maior de perdas de posse da bola são, geralmente, equipas em que o processo ofensivo não está tão bem delineado. (Hoop 2024).
- **Faltas cometidas:** o número de faltas pode ser um indicador tanto de agressividade defensiva quanto de disciplina. Se uma equipa fizer um número elevado de faltas, irá beneficiar frequentemente o adversário com lançamentos livres, sendo um ponto negativo. No entanto, uma quantidade moderada de faltas pode ser indicativa de uma defesa agressiva e física.
- **Tentativas de lançamento e lançamentos bem sucedidos:** geralmente, equipas que realizam mais tentativas de lançamento jogam a um nível mais acelerado, o que leva a uma maior carga de trabalho da defesa adversária, causando, assim, cansaço e possíveis erros. O número de tentativas de lançamento torna-se ainda mais importante caso a equipa consiga converter com sucesso grande parte destas tentativas, pois equipas que tenham esta capacidade possuem uma grande vantagem competitiva (Basketball s.d.).
- **Tentativas de lançamentos de 3 pontos e lançamentos de 3 pontos bem-sucedidos**: Um número alto de tentativas de lançamento de 3 pontos pode indicar que uma equipa é capaz de abrir espaços para que os seus melhores lançadores não tenham oposição quando realizam o lançamento. Por outro lado, poderá, também, ser indicador de que a equipa tem dificuldades em penetrar na defesa adversária e chegar perto do cesto, principalmente se a taxa de acerto destes lançamentos for baixa. Lançamentos de 3 pontos oferecem uma vantagem comparativamente aos lançamentos de 2 pontos, pois permitem pontuar mais com menos lançamentos. Mesmo com uma eficácia menor nos lançamentos de 3 pontos, uma equipa pode pontuar o mesmo, ou até mais, do que com lançamentos de dois pontos (Coach&A.D. s.d.).
- **Tentativas de lançamentos livres e lançamentos livres bem sucedidos:** Um número elevado de tentativas de lançamentos livres significa que uma equipa sofreu bastantes faltas. Estes lançamentos permitem às equipas ter uma oportunidade de marcar pontos

sem a pressão da defesa adversária e são cruciais, pois podem fazer a grande diferença no resultado final, especialmente em jogos equilibrados e com margens estreitas.

É crucial salientar que, antes do cálculo das estatísticas mencionadas, foi realizada uma verificação, de forma a identificar possíveis dados incorretos. Este processo envolveu a comparação das estatísticas listadas no ficheiro dos jogos com a soma dos valores individuais de cada jogador pertencente à equipa para o mesmo jogo (ficheiro com as estatísticas dos jogadores). Durante esta análise, foram identificados valores discrepantes, especialmente nas épocas 2003/2004 e 2004/2005, resultando na remoção dos jogos correspondentes a essas temporadas. Quanto aos jogos restantes, procedeu-se com as correções necessárias, com consulta do *website* oficial da NBA (NBA s.d.) para validar e atualizar os dados necessários. A distribuição entre a quantidade de jogos ganhos pela equipa da casa e pela equipa visitante, após esta correção, pode ser visualizada na Tabela 3.8

Tabela 3.8: Distribuição da coluna *HOME_TEAM_WINS* após remoção das épocas 2003/2004 e 2004/2005 *rankings*

Valor	**Frequência absoluta**	**Percentagem**
Sim	12814	58,8%
Não	8972	41,2%

Visto que seria impossível obter os valores destas estatísticas antes do jogo se realizar, foi calculada a sua média nos últimos 4 jogos, tanto para a equipa da casa como para a equipa visitante. Para além disso, foi também calculada a sua média nos últimos 4 jogos em casa para a equipa da casa e nos últimos 4 jogos fora para a equipa visitante.

Na tabela A.13, presente no anexo A está representada a constituição do *dataset* resultante destas operações. O *dataset* é composto por 74 colunas, excluindo a variável alvo, das quais 20 colunas são relacionadas com o cansaço das equipas, 16 com os *rankings*, 32 com o momento de forma das equipas (média das estatísticas dos últimos jogos e sequências de vitórias/derrotas) e 6 com informações das equipas (data de fundação e conferência a que pertencem). No total, o *dataset* contém 21786 jogos (linhas).

Capítulo 4

Análise exploratória dos dados

Os dados foram explorados através de análises que investigam diversos aspectos quantitativos e qualitativos da NBA, com o objetivo de compreender as dinâmicas e transformações da liga ao longo do tempo. Ao todo, foram realizadas quatro análises, cada uma abordando uma faceta distinta do jogo e o seu impacto nos jogadores e equipas.

É importante destacar que, para estas análises, o conjunto de dados foi complementado com informações provenientes de duas fontes adicionais. De *Basketball Reference* (Reference s.d.[a]), foram recolhidas estatísticas detalhadas dos jogadores por temporada, incluindo a idade, jogos realizados, média de minutos por jogo, média de pontos por jogo, etc, para apoiar as análises sobre o impacto das trocas de equipa e a evolução do desempenho dos jogadores com a idade. De *StatsMuse* (StatsMuse s.d.), foram obtidas as médias por jogo de várias estatísticas das temporadas 2003/2004 e 2004/2005, que haviam sido removidas do conjunto de dados por conterem um elevado número de jogos com estatísticas com valores incorretos, assim como da temporada 2022/2023, que havia sido excluída por estar incompleta. Foram também obtidas as médias das estatísticas por posição em campo para cada temporada do período em estudo. Estas informação foram essenciais para as análises da evolução da NBA e das posições em campo ao longo do tempo.

Inicialmente, foi examinada a evolução da NBA ao longo das temporadas, destacando as principais mudanças e tendências que moldaram o estilo de jogo. De seguida, foi investigado o impacto da mudança de equipa no desempenho dos jogadores, analisando como transferências e trocas influenciam o desempenho individual e coletivo. Na terceira análise, foi abordada a transformação das posições na NBA ao longo dos anos, observando como o papel das posições evoluiu e se adaptou às novas exigências do jogo. Por fim, foi discutida a evolução do desempenho dos jogadores com a idade, avaliando como fatores relacionados à idade influenciam o rendimento e a longevidade na carreira dos atletas.

4.1 Evolução da NBA ao longo das temporadas

Foi realizada uma análise detalhada das mudanças ocorridas na dinâmica de jogo da NBA nas últimas 20 épocas, com foco nos aspectos ofensivos e defensivos. Através da avaliação das transformações nas estratégias das equipas, a análise oferece uma visão crítica sobre como o *basketball* evoluiu e como estas mudanças têm impactado o estilo de jogo na NBA. O objetivo é entender as dinâmicas que impulsionaram essas alterações e avaliar as suas implicações para o *basketball* contemporâneo.

Inicialmente, foi analisada a evolução ofensiva das equipas, estando os dados que sustentam esta análise representados na Figura 4.1.

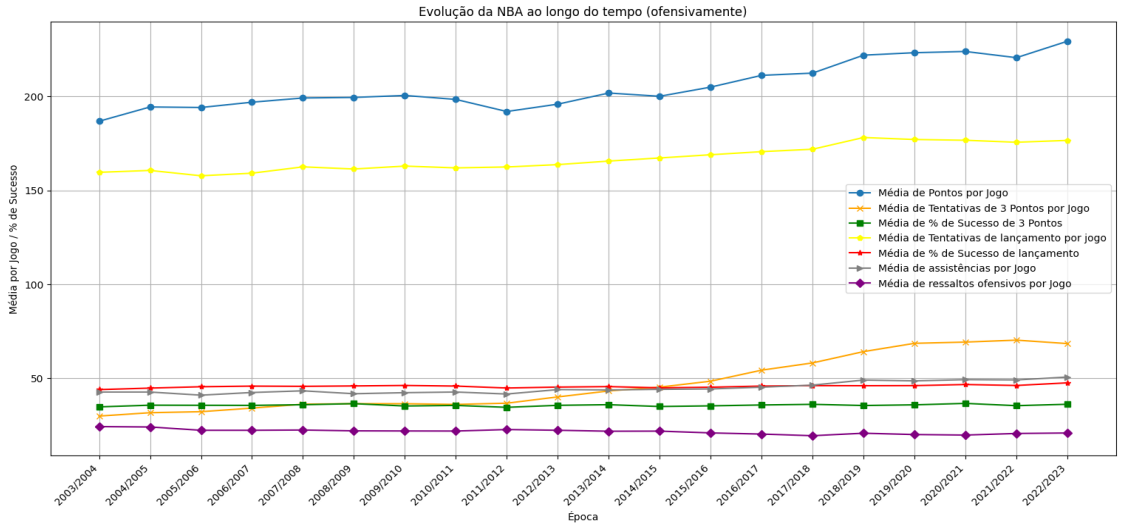

Figura 4.1: Evolução da NBA a nível ofensivo

A análise da evolução dos pontos por jogo na NBA durante o período em estudo revela tendências e variações significativas. Entre as épocas 2003/2004 e 2009/2010, observou-se um aumento gradual na média de pontos por jogo, que passou de aproximadamente 187 para cerca de 200. Nas duas épocas seguintes, a média caiu. Posteriormente, a média de pontos por jogo aumentou de forma progressiva até à época 2020/2021, sendo que na temporada seguinte a média desceu, mas sem se observar grande diferença. Na última temporada em análise, a média voltou a aumentar, atingindo o valor máximo observado, com cerca de 229 pontos por jogo.

Entre as temporadas de 2005/2006 e 2018/2019, a média de tentativas de lançamento por jogo na NBA apresentou uma tendência de aumento contínuo, refletindo mudanças significativas nas estratégias e no estilo de jogo das equipas, indicando uma transição para um jogo mais rápido e com maior frequência de lançamentos. Na temporada 2018/2019, foi registado o pico no período em estudo, com cerca de 178 tentativas por jogo, seguido de uma ligeira diminuição e subsequente estabilização nas temporadas seguintes, com valores situados entre 175 e 177 tentativas por jogo. Esta evolução evidencia um ênfase crescente no ritmo acelerado de jogo e nos lançamentos efetuados, características distintivas do basquetebol contemporâneo.

Paralelamente, a análise da evolução das tentativas de lançamentos de três pontos por jogo revela uma mudança significativa na estratégia ofensiva das equipas. Entre as épocas 2003/2004 e 2011/2012, a média de tentativas de três pontos por jogo aumentou gradualmente de aproximadamente 30 para cerca de 40. A partir da época 2012/2013, verificou-se um aumento acentuado, atingindo o valor máximo de cerca de 70 tentativas por jogo na época 2021/2022. Na época seguinte foi observado uma diminuiçao, mas sem diferenças significativas. Esta transformação reflete a crescente tendência das equipas em maximizar os lançamentos de longa distância, evidenciando uma estratégia ofensiva orientada para fazer o maior número de pontos possível. Em ambas as análises, da percentagem média de lançamentos bem sucedidos por jogo, não se observam mudanças significativas.

A análise da média de assistências por jogo na NBA durante o período em estudo revela uma tendência de crescimento substancial. A média inicial de cerca de 42 assistências por jogo em 2003/2004 aumentou de forma consistente, alcançando aproximadamente 50 em

2022/2023. Este aumento contínuo sugere uma evolução no estilo de jogo das equipas, refletindo um crescente ênfase na circulação da bola e na maximização das oportunidades de tiro, que coincide com o aumento das tentativas de lançamento, especialmente as de três pontos. Estes dados evidenciam um jogo cada vez mais coletivo e estrategicamente evoluído, permitindo às equipas abrir mais espaços e criar melhores ocasiões de lançamento.

A análise da média de ressaltos ofensivos por jogo no período em questão revela uma tendência geral de estabilidade. Embora tenham ocorrido algumas flutuações, a média não apresenta uma evolução marcante, apesar das mudanças no estilo de jogo e nas estratégias das equipas.

Em suma, através da análise comportamental das equipas a nível ofensivo, é possível concluir que o aumento do número de pontos por jogo na NBA está diretamente relacionado com a crescente frequência de tentativas de lançamento, em particular de lançamentos de três pontos. Apesar de as taxas de sucesso de lançamentos se manterem relativamente constantes, a maior frequência de tentativas leva a um maior número de pontos marcados. Um fator crucial neste cenário é a evolução do jogo coletivo das equipas, que apresentam uma melhor circulação de bola e maior capacidade de abrir espaços. Esta melhoria permite criar mais oportunidades para lançamentos, aumentando assim o volume de tentativas, especialmente as de três pontos. Este foco nos lançamentos de longa distância sublinha uma evolução no *basketball* contemporâneo, onde a eficiência e o volume de lançamentos se tornaram fundamentais para o sucesso ofensivo das equipas.

Em seguida, foi analisada a evolução defensiva das equipas, estando os dados que sustentam esta análise representados na Figura 4.2.

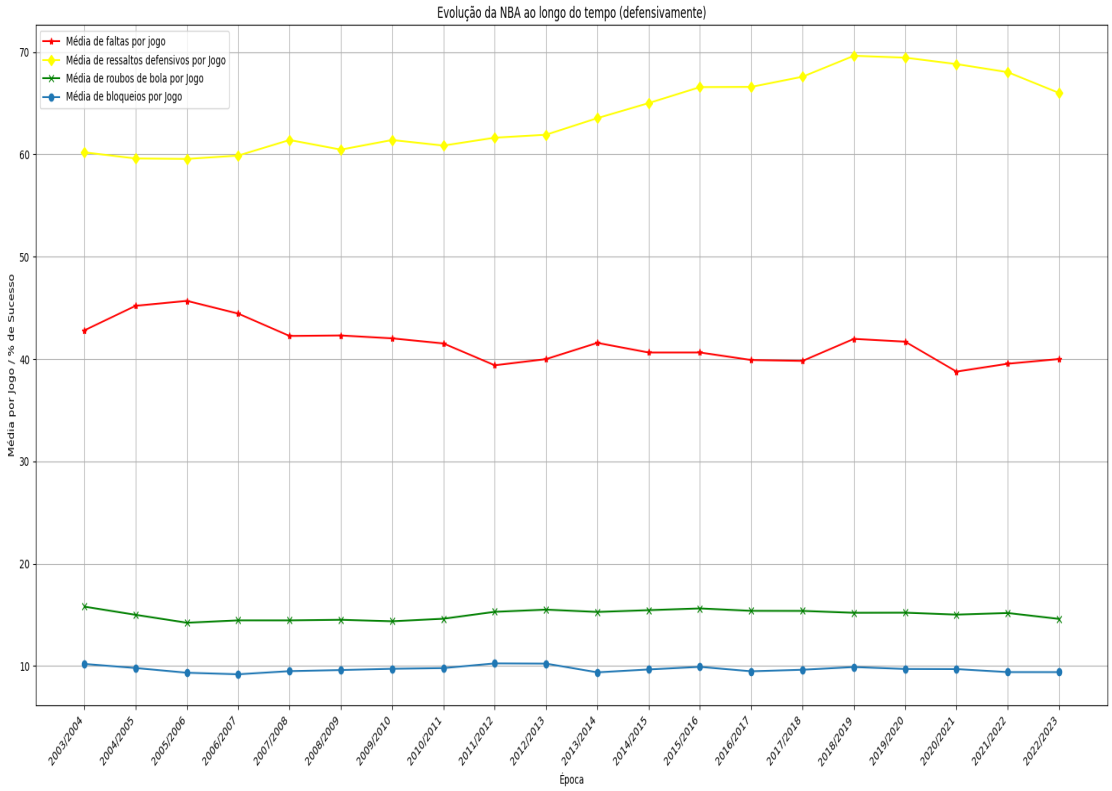

Figura 4.2: Evolução da NBA a nível defensivo

Durante o período em estudo, observou-se uma tendência crescente no número médio de ressaltos defensivos por jogo. A média iniciou-se em cerca de 60 ressaltos defensivos por jogo na primeira época estudada e finalizou com aproximadamente 66 na última, refletindo um crescimento total de cerca de 6 ressaltos ao longo do período analisado. A partir da temporada 2007/2008, verificou-se um crescimento contínuo nesta métrica, com exceção de algumas flutuações menores. O valor máximo foi registado em 2018/2019, com uma média de aproximadamente 70 ressaltos, seguido por uma ligeira redução nos anos subsequentes. Este aumento sustentado pode ser atribuído a várias alterações estratégicas no jogo, incluindo um maior ênfase na defesa e nos ressaltos defensivos, bem como à evolução das competências individuais dos jogadores e das estratégias das equipas.

A análise da evolução das médias de faltas, roubos de bola e bloqueios no período em estudo revela aspetos significativos sobre a evolução das estratégias defensivas das equipas. Observa-se uma redução consistente na média de faltas, que caiu de cerca de 46, valor observado na época 2005/2006, para 40, na última época analisada. No que diz respeito às outras métricas, estas mantiveram-se praticamente constantes ao longo do período em análise. Estes dados indicam que, apesar da diminuição no número de faltas, as equipas conseguiram manter a sua capacidade defensiva. Esta tendência sugere que as equipas foram capazes de refinar as suas estratégias defensivas, conseguindo equilibrar a consistência defensiva com a agressividade.

Em suma, ao longo do período em estudo, as equipas apresentaram uma evolução significativa tanto na sua capacidade ofensiva quanto na defensiva. Atualmente, a NBA tem um volume ofensivo maior, com um ritmo mais acelerado e jogo coletivo mais proeminente, sendo o principal foco fazer lançamentos de três pontos. Para além disso, o processo defensivo das equipas é melhor, com as equipas a terem uma maior capacidade de controlo do jogo na defesa, evidenciado pelo aumento do número médio de ressaltos defensivos, e um maior equilíbrio entre a agressividade e a consistência defensiva, evidenciado pela redução no número de faltas e pela manutenção dos roubos de bola e bloqueios.

4.2 Impacto das trocas de equipa no desempenho dos jogadores

Foi analisado o impacto das trocas de equipa no desempenho dos jogadores da NBA ao longo dos últimos 20 anos, considerando dois cenários distintos: mudanças de equipa realizadas durante a mesma temporada, ou seja, com a temporada desportiva a decorrer, e mudanças realizadas entre o final de uma temporada e o início da seguinte. A análise abrangeu trocas ocorridas ao longo desse período, com base em cinco métricas principais: Contribuição Ofensiva (OFC), Contribuição Defensiva (DFC), minutos jogados, jogos jogados e jogos como titular. No total, foram analisadas 1126 trocas realizadas com a temporada a decorrer e 800 trocas realizadas entre temporadas ao longo dos 20 anos estudados.

As métricas de OFC e DFC foram calculadas através das seguintes fórmulas personalizadas, criadas especificamente para esta análise.

$$\text{OFC} = \frac{\text{PTS} + 0.4\text{FG} - 0.7(\text{FGA} - \text{FG}) - 0.4(\text{FTA} - \text{FT}) + 0.7\text{ORB} + 0.3\text{AST} - 0.4\text{TO}}{\text{MP}} \tag{4.1}$$

$$\text{OFC} = \frac{0.7\text{DRB} + 0.4\text{STL} + 0.4\text{BLK}}{\text{MP}} \tag{4.2}$$

A fórmula para OFC (Fórmula 4.1) leva em consideração os pontos marcados (PTS), lançamentos bem sucedidos (FG), lançamentos tentados (FGA), ressaltos ofensivos (ORB), assistências (AST), lances livres tentados (FTA), lances livres convertidos (FT) e perdas de bola (TO), ajustado pelo tempo jogado pelo jogador (MP).

A fórmula para DFC (Fórmula 4.2) tem em conta os ressaltos defensivos (DRB), roubos de bola (STL) e bloqueios (BLK), ajustado pelo tempo jogado pelo jogador (MP).

Estas fórmulas foram baseadas no *Player Efficiency Rating* (PER) (Reference s.d.[b]), na medida em que somam as ações positivas e substraem as negativas e atribuem pesos às estatísticas, considerando os minutos jogados. Optou-se por utilizar estas duas fórmulas em vez do PER devido à complexidade deste último, que envolve muitas variáveis, algumas das quais não estavam disponíveis, como, por exemplo, o factor de ritmo da equipa.

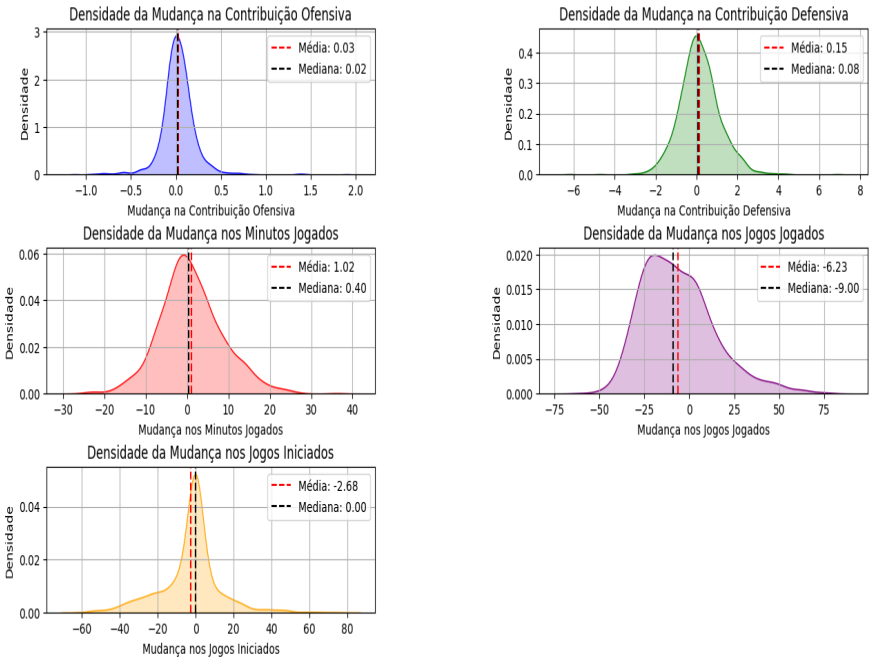

Figura 4.3: Impacto das trocas de equipa com a temporada a decorrer

Nas trocas que ocorrem com a temporada a decorrer (Figura 4.3) , a variação na contribuição ofensiva é mínima, com uma média de 0.03 e uma mediana de 0.02, o que caracteriza uma variação baixa. Estes valores indicam que, em geral, os jogadores mantêm um desempenho ofensivo semelhante após a troca, sem grandes oscilações. A contribuição defensiva também apresenta uma variação baixa, com uma média de 0.15 e uma mediana de 0.08. Apesar de uma leve tendência de melhoria na performance defensiva, essa variação é marginal, refletindo estabilidade. A variação nos minutos jogados é igualmente baixa, com uma média de 1.02 e uma mediana de 0.40, sugerindo que o tempo de jogo dos jogadores não sofre alterações significativas após a troca. Por outro lado, a métrica de jogos jogados revela um impacto mais expressivo, com uma variação alta: a média de -6.23 e a mediana de -9.00 indicam que muitos jogadores veem uma redução significativa no número de partidas disputadas após a troca. Nos jogos como titular, a variação é baixa, com uma média de

-2.68 e uma mediana de 0.00. Isto sugere que os jogadores costumam manter o seu estatuto de titular ou jogador de banco.

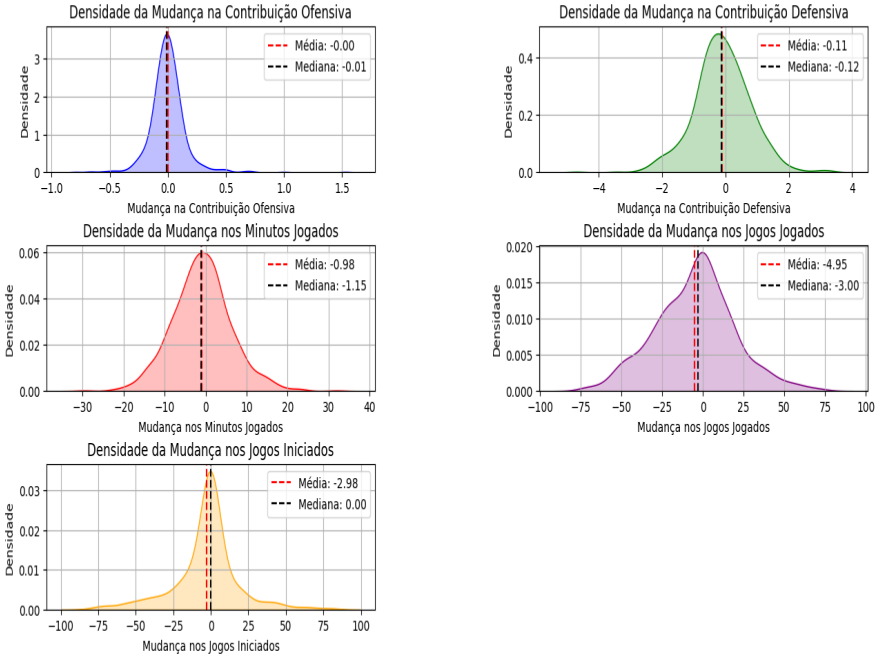

Figura 4.4: Impacto das trocas de equipa entre temporadas

Em relação às trocas que ocorrem de temporada para temporada (Figura 4.4), a contribuição ofensiva mantém uma variação baixa, com uma média de -0.00 e uma mediana de -0.01. Isto reforça a ideia de que, mesmo com a mudança de equipa, o desempenho ofensivo dos jogadores se mantém relativamente estável. A contribuição defensiva apresenta uma leve queda, com uma média de -0.11 e uma mediana de -0.12, também caracterizando uma variação baixa. Embora a mudança de equipa possa ter um leve impacto negativo na defesa, a maioria dos jogadores consegue preservar o seu nível de desempenho. A variação nos minutos jogados após trocas entre épocas é baixa, com uma média de -0.98 e uma mediana de -1.15, sugerindo que o tempo em campo não sofre alterações drásticas. Já a métrica de jogos jogados revela uma variação moderada, com uma média de -4.95 e uma mediana de -3.00, indicando uma redução perceptível no número de partidas disputadas. Nos jogos como titular, a variação é baixa, com uma média de -2.98 e uma mediana de 0.00, sugerindo que os jogadores costumam manter o seu estatuo.

Em resumo, as trocas com a temporada a decorrer tendem a causar um impacto mais significativo no número de jogos jogados. Quanto às restantes métricas analisadas, não se notaram grandes diferenças entre as duas tipologias de troca analisadas. Estes resultados sugerem que, embora as trocas possam alterar a participação dos jogadores, o desempenho individual em termos de contribuição ao jogo mantém-se relativamente estável, o que reflete uma capacidade de adaptação por parte dos atletas.

4.3 Evolução das posições na NBA ao longo das temporadas

Foi feita uma análise da evolução das cinco posições na NBA ao longo do tempo. Esta análise focou-se em métricas de desempenho dos jogadores, como pontos por jogo (PPG), assistências por jogo (APG), tentativas de lançamento por jogo (FGA), percentagem de acerto de lançamentos (FG%), tentativas e percentagem de acerto de três pontos (3PA e 3P%), ressaltos ofensivos (OREB), ressaltos defensivos (DREB), bloqueios (BPG) e roubos de bola (SPG). É importante referir que optou-se por contabilizar apenas os jogadores titulares, de forma a garantir maior precisão na análise, uma vez que jogadores com poucos minutos de jogo e desempenhos muito inferiores poderiam influenciar negativamente as médias e distorcer a análise.

4.3.1 Point Guard

A posição de Point Guard (PG) na NBA tem experienciado uma evolução significativa ao longo das últimas duas décadas, refletindo mudanças nas estratégias de jogo e nas demandas das equipas. Analisando as estatísticas de desempenho para Point Guards entre as temporadas de 2003/2004 a 2022/2023 (Figura 4.5), observa-se uma transformação notável em várias dimensões do jogo.

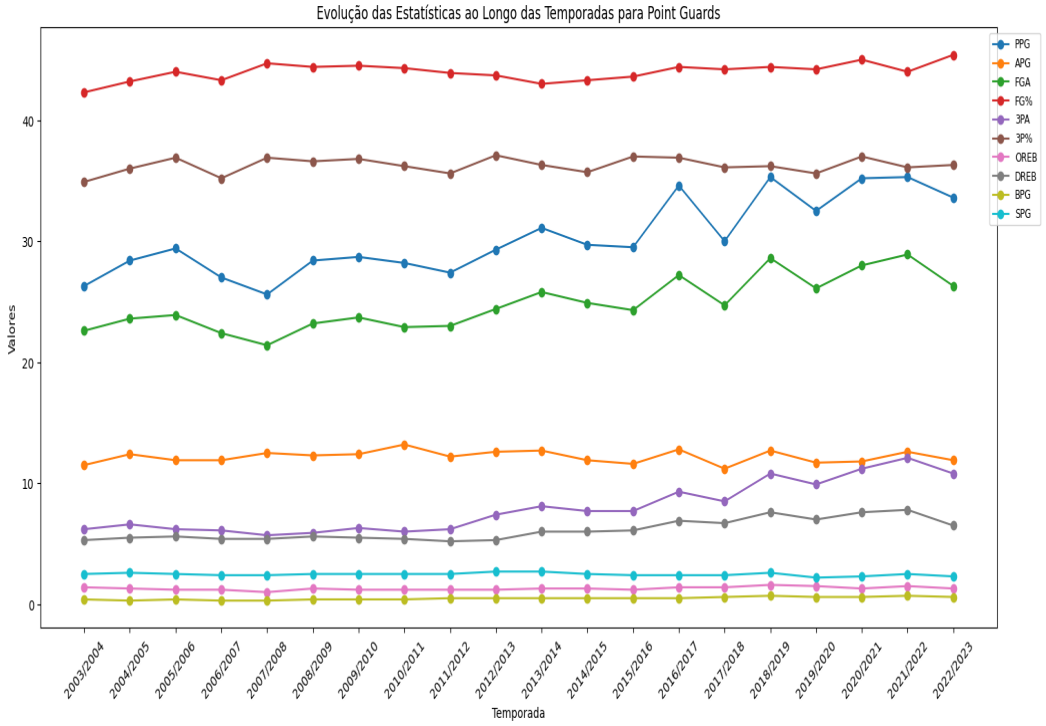

Figura 4.5: Evolução da posição *Point Guard*

Em termos de produção ofensiva, os pontos por jogo (PPG) dos Point Guards apresentaram uma tendência de crescimento ao longo do período analisado. Em 2003/2004, a média era de cerca de 26 pontos por jogo, tendo esse valor crescido de forma gradual até aos anos mais recentes, atingindo picos em 2018/2019, 2020/2021 e 2021/2022 com cerca de 35 pontos. Além disso, a pontuação manteve-se alta entre 2018/2019 e 2022/2023, o que

sugere que os *Point Guards* modernos têm assumido papéis cada vez mais centrais no ataque das equipas.

O volume de lançamentos tentados por jogo (FGA) também aumentou ao longo do tempo, subindo de cerca de 22 em 2003/2004 para cerca de 29 em 2021/2022, onde atingiu o pico. Este aumento no volume de lançamentos é indicativo de um papel mais fundamental dos *Point Guards* no processo ofensivo das equipas, principalmente na marcação de pontos. Apesar deste crescimento no número de lançamentos, a percentagem de acerto (FG%) mostrou uma leve tendência de aumento, começando em cerca de 42,3% e atingindo cerca de 46% em 2022/2023. Este incremento na eficiência sugere que os *Point Guards* têm se adaptado às exigências modernas do jogo, melhorando sua capacidade de conversão dos lançamentos tentados. O aumento das tentativas de lançamentos de três pontos (3PA) é particularmente notável, refletindo a ênfase crescente nos lançamentos de longa distância. Apesar do aumento no volume, a percentagem de acerto de lançamentos de três pontos (3P%) permaneceu relativamente estável, variando entre cerca de 35% e cerca de 37%. Isto sugere que, embora o número de tentativas tenha aumentado, a eficiência dos lançamentos de três pontos manteve-se constante, refletindo uma adaptação ao estilo de jogo mais orientado para o lançamento de longa distância.

Os ressaltos ofensivos (OREB) e as assistências (APG) permaneceram relativamente estáveis ao longo do período analisado. No caso das assistências, variaram entre cerca de 11 e 13. Esta estabilidade indica que, apesar do crescente papel dos *Point Guards* como marcadores de pontos, estes continuam a desempenhar o papel principal na facilitação e criação de jogadas para as suas equipas.

Em termos de contribuição defensiva, os roubos de bola (SPG) e bloqueios (BPG) por jogo mostraram uma tendência de estabilidade. Os ressaltos defensivos (DREB) algum crescimento, o que pode refletir uma maior responsabilidade dos *Point Guards* no contexto defensivo.

Em resumo, a análise das estatísticas de desempenho dos *Point Guards* revela uma posição em transformação. Os *Point Guards* têm se tornado cada vez mais importantes como marcadores de pontos, adaptando-se às mudanças no estilo de jogo com um aumento nas tentativas de lançamentos de três pontos e uma eficiência ofensiva aprimorada. Simultaneamente, eles têm aumentado sua contribuição defensiva, com um crescimento nos ressaltos defensivos. Apesar destas mudanças, a capacidade de facilitar o jogo e criar oportunidades continua a ser uma constante, evidenciando o papel multifacetado dos *Point Guards* na NBA moderna.

4.3.2 Shooting Guard

A análise da evolução da posição de *Shooting Guard* na NBA (Figura 4.6) revela transformações significativas no papel e nas estatísticas desses jogadores ao longo das últimas duas décadas. Desde 2003/2004 até 2022/2023, a posição passou por mudanças substanciais, refletindo a evolução do *basketball* e as novas exigências do jogo.

Em termos de produção ofensiva, os *Shooting Guards* exibiram um aumento notável na média de pontos por jogo (PPG), que subiu de cerca de 32 em 2003/2004 para cerca de 46 em 2022/2023. Este crescimento sublinha uma tendência crescente de valorização da capacidade de marcar pontos por parte dos jogadores desta posição. Estes dados refletem o papel da posição como principal pontuadora das equipas.

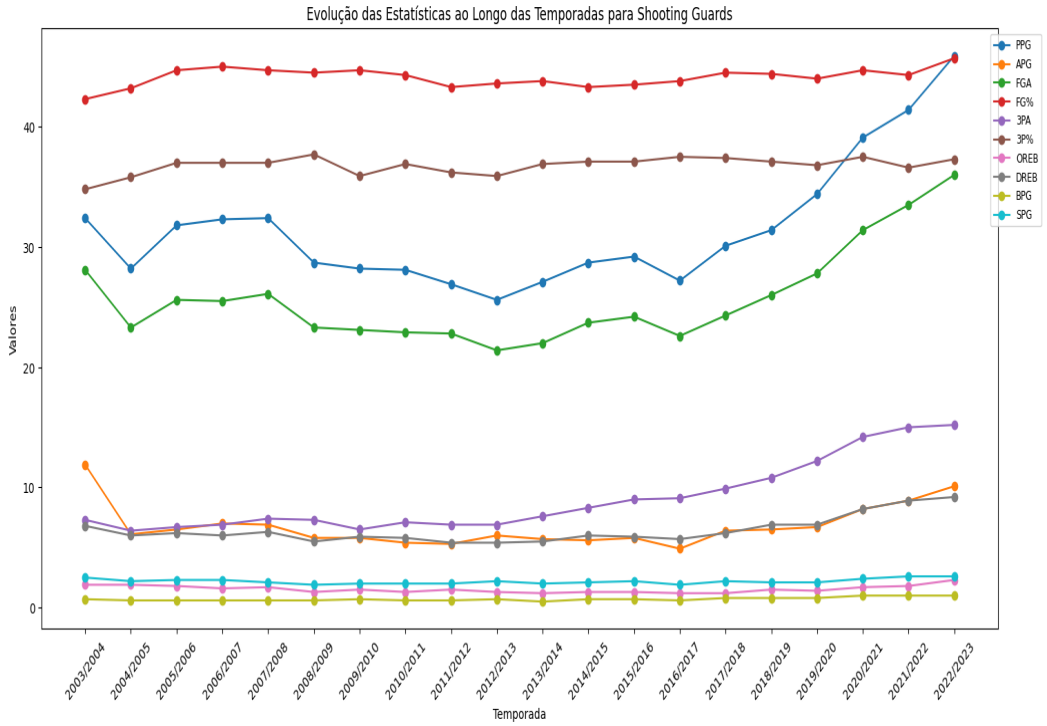

Figura 4.6: Evolução da posição *Shooting Guard*

O volume de lançamentos tentados (FGA) também demonstrou uma evolução significativa, passando de cerca de 28 lançamentos por jogo em 2003/2004 para cerca de 36 em 2022/2023. Este aumento substancial indica uma abordagem mais agressiva e uma maior responsabilidade atribuída aos *Shooting Guards* na criação e conversão de oportunidades ofensivas. Com o estilo de jogo da NBA a enfatizar a velocidade e a eficiência ofensiva, os *Shooting Guards* adaptaram-se a essas mudanças, assumindo um papel mais proativo. A importância do lançamento de três pontos na NBA aumentou consideravelmente ao longo dos anos, e os *Shooting Guards* são um reflexo claro dessa mudança. As tentativas de lançamentos de três pontos (3PA) subiram de cerca de 7 por jogo em 2003/2004 para cerca de 15 em 2022/2023. Este crescimento evidência a crescente importância dos lançamentos de longa distância, que se tornaram fundamentais. A percentagem de acerto em lançamentos de três pontos (3P%) manteve-se relativamente estável, indicando que, apesar do aumento na frequência, os jogadores mantiveram uma consistência na sua capacidade de conversão. Em termos de eficiência de lançamento geral, a percentagem de acerto (FG%) melhorou, passando de cerca de 42% em 2003/2004 para cerca de 46% em 2022/2023. Este aumento na eficiência reflete melhorias na precisão dos lançamentos e na capacidade dos jogadores em aproveitar as oportunidades ofensivas.

Os ressaltos ofensivos (OREB) mostraram uma tendencia de estabilidade. Por outro lado, os ressaltos defensivos (DREB) mostraram um pequeno crescimento, subindo de cerca de 7 em 2003/2004 para cerca de 9 em 2022/2023. Este aumento sugere uma maior participação no contexto defensivo.

Em relação às assistências (APG), houve uma variação significativa. As assistências começaram em cerca de 12 em 2003/2004, existindo uma diminuição gradual até atingir o valor mais baixo em 2016/2017. Esta diminuição reflete uma mudança na função dos

Shooting Guards, com uma ênfase crescente na pontuação e menos na criação de jogadas. No entanto, observou-se uma recuperação recente na capacidade de distribuir o jogo, verificando-se sempre um aumento nos anos seguintes.

O número de roubos de bola (SPG) e bloqueios (BPG) revelou estabilidade.

Em resumo, a posição de *Shooting Guard* na NBA evoluiu significativamente, com uma clara ênfase no aumento da produção ofensiva e na adaptação às novas exigências do jogo. O crescimento na média de pontos, no volume de lançamentos e na importância dos lançamentos de três pontos, combinado com melhorias na eficiência e na contribuição defensiva, evidencia uma transformação multifacetada e uma adaptação contínua aos estilos de jogo modernos e às exigências da NBA.

4.3.3 Small Forward

A posição de *Small Forward* na NBA tem mostrado uma evolução marcante ao longo das últimas duas décadas, refletindo mudanças significativas nas funções e no estilo de jogo desta posição. O desempenho dos *Small Forwards* em várias métricas estatísticas ilustra como este papel se transformou e se adaptou ao jogo moderno (Figura 4.7).

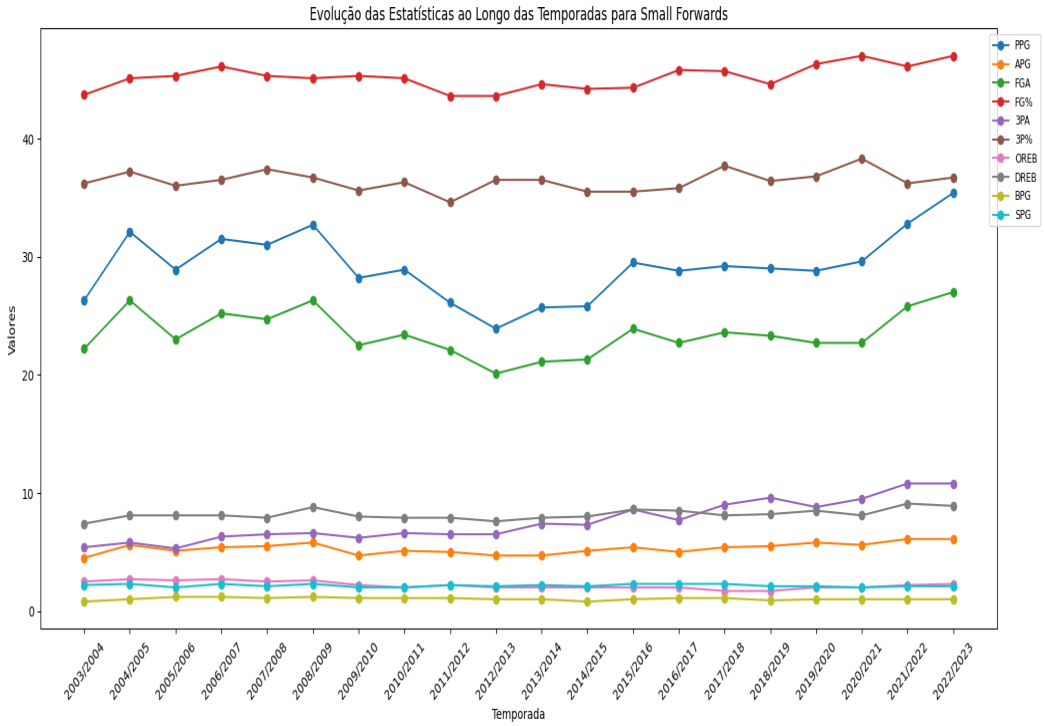

Figura 4.7: Evolução da posição *Small Forward*

Em termos de pontos por jogo (PPG), os *Small Forwards* demonstraram um aumento notável na sua produção ofensiva. Em 2003/2004, a média de pontos por jogo era de cerca de 26, sendo que essa média subiu para cerca de 36 em 2022/2023, valor mais alto do perído em análise. Este crescimento é indicativo da importância crescente desta posição como uma fonte primária de pontos para as equipas.

A eficiência dos *Small Forwards* no lançamento (FG%), também melhorou ao longo do tempo. A percentagem de acerto em lançamentos subiu de cerca de 44% em 2003/2004

para cerca de 47% em 2022/2023. Esta evolução reflete uma maior precisão, com os *Small Forwards* a serem cada vez mais eficazes nas suas tentativas. É também notado um aumento no número de tentativas de lançamento (FGA) entre o início e o fim do período em análise. Esta métrica está diretamente ligada aos pontos marcados por jogo, visto que se observa que quando esta métrica sofreu aumentos, existiu também um aumento dos pontos e que, quando esta métrica diminiu, os pontos diminuiram também. Ambas as métricas sofreram um declínio entre as épocas 2009/2010 e 2012/2013, porém após este período, observou-se um aumento gradual de ambas até ao final do período em análise. O número de tentativas de lançamento de três pontos (3PA) dos *Small Forwards* aumentou significativamente, passando de cerca de 5 em 2003/2004 para cerca de 11 em 2022/2023. Este aumento ilustra a crescente ênfase na capacidade de lançar de longa distância, uma característica crucial no estilo de jogo moderno da NBA. Observaram-se flutuações na percentagem de acerto em triplos (3P%), que variaram entre cerca de 36% e 38%.

A contribuição dos *Small Forwards* em termos de ressaltos é igualmente significativa. Os ressaltos defensivos (DREB) mostraram uma média estável, com valores que variaram entre cerca de 7 e 9 ao longo dos anos, indicando que os *Small Forwards* continuam a desempenhar um papel crucial na proteção do cesto e na recuperação da posse de bola. Os ressaltos ofensivos (OREB) mantiveram-se relativamente baixos, refletindo uma maior ênfase em outros aspectos do jogo.

As assistências (APG), bloqueios (BPG) e roubos de bola (STG) mantiveram uma tendência de estabilidade.

Em suma, a posição de *Small Forward* evoluiu para um papel mais completo e multifacetado na NBA moderna. O aumento na capacidade de pontuar, a melhoria na eficiência de lançamento e a ênfase no lançamento de três pontos refletem um papel mais dinâmico e influente dentro das equipas. Ao mesmo tempo, a consistência nas contribuições defensivas demonstra a importância contínua desta posição na proteção do cesto e na pressão defensiva. A combinação destas características faz dos *Small Forwards* jogadores fundamentais e versáteis na NBA contemporânea.

4.3.4 Power Forward

A posição de *Power Forward* na NBA tem sofrido mudanças ao longo das últimas duas décadas. O desempenho dos *Power Forwards* em várias métricas estatísticas ilustra como este papel se transformou e se adaptou ao jogo moderno (Figura 4.8).

O número de pontos por jogo (PPG) flutuou ao longo das temporadas. Começou elevado, com cerca de 29 pontos em 2003/2004, atingiu um pico de cerca de 33 pontos em 2020/2021 e, posteriormente, sofreu uma leve queda para aproximadamente 28 pontos em 2022/2023. Embora haja variações, a diferença entre o início e o fim do período analisado não é muito significativa, indicando que a capacidade de pontuar dos *Power Forwards* permaneceu relativamente constante. O mesmo se verificou nas assistências por jogo (APG).

As tentativas de lançamento (FGA) também sofreram flutuações, mas no geral não se observam grandes diferenças entre o início e o fim do período em estudo. Em relação às tentativas de três pontos (3PA), houve um aumento significativo ao longo dos anos. No início do período em estudo, a média era de cerca de duas tentativas por jogo, enquanto que, nas últimas épocas, a média variava entre 7 e 9 tentativas. Esta mudança reflete uma maior ênfase nos lançamentos de longa distância e uma adaptação ao estilo de jogo moderno que valoriza mais os tiros de três pontos. A percentagem de lançamentos bem-sucedidos

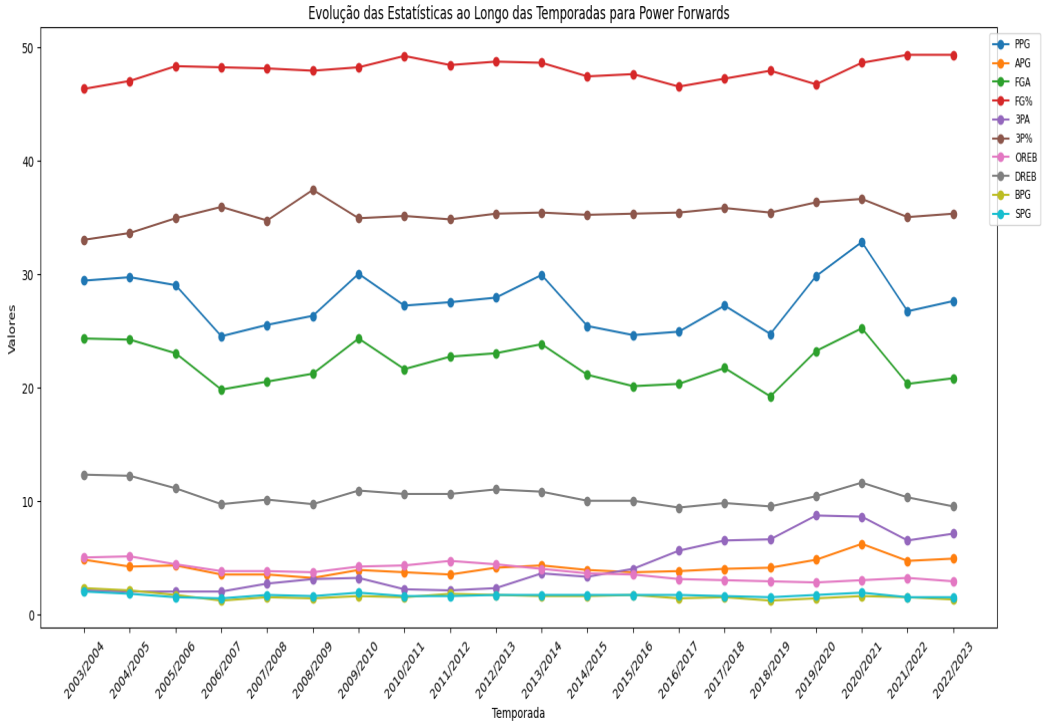

Figura 4.8: Evolução da posição *Power Forward*

(FG%) e a percentagem de lançamentos de três pontos bem-sucedidos (3P%) permaneceram relativamente estáveis, com uma leve melhoria ao longo dos anos. A percentagem de lançamentos (FG%) começou em cerca de 46% em 2003/2004 e terminou em cerca de 49% em 2022/2023. A percentagem de três pontos (3P%) variou de cerca de 33% para cerca de 35% ao longo do período, mostrando uma leve melhoria.

Os ressaltos ofensivos (OREB) e defensivos (DREB) mostraram uma ligeira tendência de queda ao longo do período analisado. Em 2003/2004, os ressaltos ofensivos eram cerca de 5 por jogo, enquanto que, em 2022/2023, caíram para cerca de 3. Da mesma forma, os ressaltos defensivos variaram de aproximadamente 12 em 2003/2004 para cerca de 10 em 2022/2023. Apesar desta tendência de queda, a variação não é muito significativa. Os bloqueios (BPG) e os roubos de bola (SPG) mantiveram-se relativamente estáveis

Em suma, a análise das estatísticas dos *Power Forwards* ao longo das temporadas revela uma evolução significativa em alguns aspectos, enquanto outros permaneceram relativamente estáveis. A capacidade de pontuar e de assistir apresentou flutuações ao longo do período, mas a variação não foi tão drástica. O aumento nas tentativas de três pontos e a leve melhoria nas percentagens de acerto evidenciam uma adaptação ao estilo moderno de jogo que valoriza o tiro de longa distância. Embora os ressaltos ofensivos e defensivos tenham mostrado uma ligeira tendência de queda, as diferenças não são substanciais. Por outro lado, os bloqueios e roubos de bola mantiveram-se relativamente constantes ao longo das temporadas, indicando uma estabilidade na capacidade defensiva.

4.3.5 Poste

A posição de Poste na NBA tem mostrado uma grande evolução ao longo das últimas duas décadas, refletindo mudanças bastante significativas nas funções e no estilo de jogo desta posição. O desempenho dos Postes em várias métricas estatísticas ilustra como este papel se transformou e se adaptou ao jogo moderno (Figura 4.9).

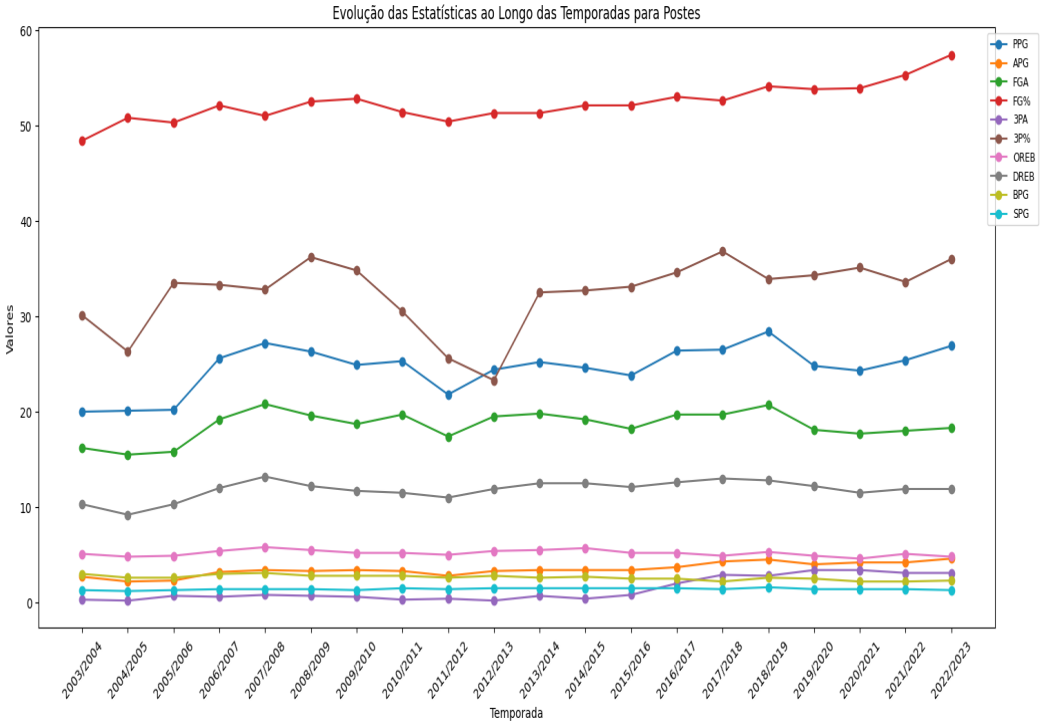

Figura 4.9: Evolução da posição Poste

O número de pontos por jogo (PPG) apresentou um crescimento notável ao longo das temporadas, começando em cerca de 20 pontos em 2003/2004 e atingindo um pico de aproximadamente 28 pontos em 2018/2019, antes de estabilizar em torno de 27 pontos em 2022/2023. Este aumento demonstra uma capacidade crescente dos postes para contribuir ofensivamente, com um desempenho robusto ao longo do período analisado. Nas assistências por jogo (APG) observou-se uma ligeira tendência de crescimento, iniciando-se em cerca de 3 em 2003/2004 e alcançando aproximadamente 5 assistências em 2022/2023. Este aumento reflete uma maior versatilidade dos postes, que passaram a desempenhar um papel mais ativo na criação de jogadas e na dinâmica ofensiva das suas equipas.

O número de tentativas de lançamento por jogo (FGA) sofreu flutuações, com aumentos e reduções, mas no geral não se notam grandes diferenças entre o início e o fim do período em análise. A percentagem de acerto de lançamentos (FG%) mostrou uma tendência de melhoria, iniciando-se em aproximadamente 48% e atingindo aproximadamente 57% no fim do período em estudo. Este aumento é bastante significativo, sugerindo uma melhoria na precisão dos Postes e também uma maior capacidade de se isolarem, de forma a criar oportunidades de lançamento sem oposição. Quanto às tentativas de três pontos por jogo (3PA), houve uma mudança notável ao longo das temporadas. Em 2003/2004, os postes tentavam aproximadamente 0 lançamentos de três pontos por jogo, um valor extremamente

baixo. Este número começou a aumentar a partir de 2016/2017, alcançando aproximadamente 3 tentativas por jogo em 2022/2023. Esta evolução sugere uma adaptação ao estilo de jogo moderno, que valoriza cada vez mais os lançamentos de longa distância. A percentagem de acerto de três pontos (3P%) também mostra uma tendência de melhoria, apesar de um período de queda significativa entre 2009/2019 e 2012/2013. Começou em aproximadamente 30% em 2003/2004 e atingiu cerca de 36% em 2022/2023. Embora o número de lançamentos de três pontos ainda não seja uma característica central do jogo dos Postes, o aumento no número de tentativas e na percentagem de acerto reflete uma maior eficácia nestes lançamentos.

Os ressaltos ofensivos (OREB) mostraram uma tendência de estabilidade. Quanto aos ressaltos defensivos (DREB) apresentaram uma evolução significativa, começando em aproximadamente 10 em 2003/2004 e aumentando até cerca de 13 em 2007/2008. A partir desse ponto, os valores mostraram uma leve flutuação, estabilizando-se em torno de 12 em 2021/2022 e 2022/2023. Esta tendência sugere que, embora os postes tenham continuado a ser eficazes na captura de ressaltos defensivos, houve uma leve diminuição nos valores mais recentes, possivelmente devido a mudanças na dinâmica do jogo ou nas funções específicas dos jogadores.

Os roubos de bola (SPG) mantiveram-se consistentes. O mesmo se verificou para os bloqueios (BPG), o que indica que estes continuam a ser uma parte importante do jogo defensivo dos postes, refletindo a sua capacidade de proteger o cesto e influenciar o jogo defensivo.

Em suma, os dados demonstram que a posição de Poste evoluiu significativamente ao longo dos anos, com melhorias notáveis na capacidade de pontuar e na eficiência dos lançamentos. O aumento nas assistências e nas tentativas de três pontos reflete uma maior versatilidade e adaptação ao estilo de jogo moderno. Embora haja uma leve diminuição em alguns indicadores defensivos, como os ressaltos defensivos, a posição de Poste continua a desempenhar um papel crucial tanto no ataque quanto na defesa, ajustando-se às exigências evolutivas do basquetebol contemporâneo.

Concluindo, no geral, as posições na NBA tornaram-se mais versáteis ao longo do período analisado, com um aumento na capacidade ofensiva e na adaptação ao jogo moderno. Os jogadores evoluíram para se focarem mais em lançamentos de três pontos e em habilidades multifacetadas, refletindo uma tendência em direção a um estilo de jogo mais dinâmico e flexível.

4.4 Desempenho dos jogadores com o evoluir da idade

Foi realizada uma análise do desempenho dos jogadores da NBA em função da idade, utilizando faixas etárias específicas para capturar com precisão as variações nas métricas de desempenho ao longo das suas carreiras. Os dados cobrem 20 épocas da NBA e incluem estatísticas para os mesmos jogadores em diferentes temporadas e idades, o que permite uma análise mais detalhada e precisa das dinâmicas de desempenho ao longo do tempo.

As faixas etárias selecionadas foram: '18-23 anos', '24-27 anos', '28-30 anos', '31-33 anos', '34-36 anos', '37-39 anos' e '40 ou mais anos'. Estas categorias foram definidas para refletir o desenvolvimento inicial dos jogadores, o pico da sua *performance* e as mudanças na fase final da carreira. A faixa mais ampla de '18-23 anos' engloba a fase de crescimento e

adaptação dos atletas, enquanto as subdivisões mais detalhadas para as idades mais avançadas permitem uma análise mais minuciosa das alterações no desempenho associadas ao envelhecimento. Esta abordagem proporciona uma visão clara e detalhada das diferentes fases da carreira dos jogadores, desde o início promissor até os estágios finais, oferecendo uma compreensão aprofundada das dinâmicas de *performance* ao longo do tempo.

Foram consideradas diversas métricas, tanto relacionadas com a participação dos jogadores em termos de jogos e minutos jogados, como também, estatísticas que medem a eficiência e a contribuição em várias dimensões do jogo, como pontos, assistências, ressaltos, etc. Para além disso, apenas foram contabilizados jogadores com pelo menos em média 5 minutos jogados por jogo, de forma a excluir jogadores com papel marginal ou tempo de jogo insuficiente, que poderiam distorcer a análise.

A Figura 4.10 descreve a evolução do desempenho dos jogadores da NBA com o passar da idade e a Tabela 4.1 contém a distribuição das faixas etárias.

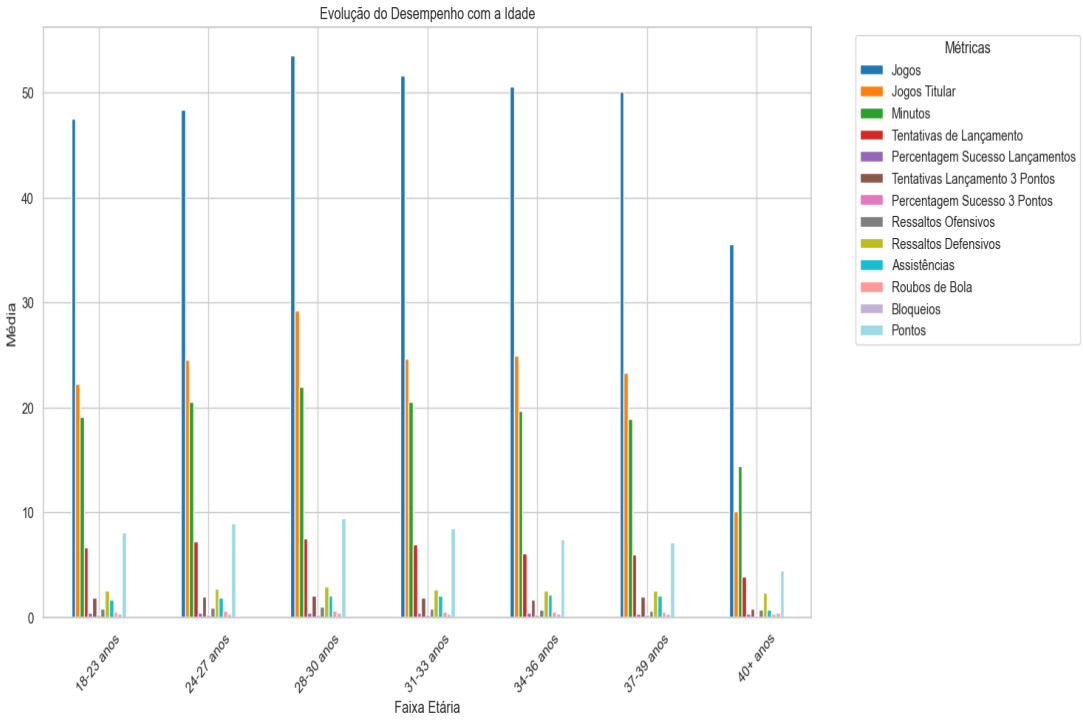

Figura 4.10: Evolução do desempenho com a idade

Tabela 4.1: Distribuição dos jogadores por faixa etária

Faixa etária	**Frequência**
18-23 anos	1338
24-27 anos	1264
28-30 anos	566
31-33 anos	356
34-36 anos	174
37-39 anos	66
40+ anos	13

Durante a fase inicial da carreira, que abrange dos 18 aos 23 anos, os jogadores demonstram uma presença em campo relativamente limitada. Nesta fase, observa-se um menor número de jogos jogados, jogos como titular e minutos em campo, refletindo o processo de adaptação ao nível competitivo da NBA. As estatísticas relacionadas com o desempenho em campo tendem a ser moderadas. Nesta faixa etária estão incluídos 1338 jogadores, refletindo a grande quantidade de principiantes e jogadores em início de carreira. Esse número elevado indica que muitos jogadores entram na NBA jovens.

Entre os 24 e os 27 anos, verifica-se um aumento em várias métricas de desempenho. Os jogadores atingem uma fase de maior maturidade e estabilidade, evidenciada por um número crescente de jogos jogados, jogos como titular e minutos em campo. As tentativas de lançamento e os pontos marcados apresentam um ligeiro incremento, enquanto as restantes estatísticas de jogo permanecem relativamente estáveis. Nesta faixa etária existem 1264 jogadores, o que sugere que, após os primeiros anos, muitos conseguem afirmar-se e permanecer na liga.

Na faixa etária de 28 a 30 anos, com 566 jogadores, observa-se o auge do desempenho no *basketball*. Esta redução do número de jogadores em relação às faixas anteriores pode indicar que apenas os jogadores que atingiram um nível elevado de desempenho permanecem ativos na liga até esta idade. Nesta fase, os jogadores combinam a experiência acumulada ao longo das temporadas com o auge físico e mental, o que lhes permite apresentar os melhores resultados em termos de jogos jogados, jogos como titular e minutos em campo. As estatísticas mostram um aumento nos pontos e nas tentativas de lançamento em comparação com a faixa etária anterior, atingindo o pico durante este período. As restantes estatísticas de jogo permanecem relativamente estáveis. Apesar disso, a necessidade das equipas de renovarem os seus plantéis com talentos mais jovens, que oferecem maior potencial de desenvolvimento e contratos mais acessíveis, pode explicar a redução no número de jogadores nesta faixa etária em comparação com as anteriores. Assim, embora os atletas que atingem um alto nível de rendimento se mantenham ativos, a transição gradual para jogadores mais jovens torna-se uma estratégia comum nas equipas.

Após os 30 anos, inicia-se um processo gradual de declínio no desempenho. Embora os jogadores continuem a desempenhar papéis importantes, tendem a apresentar uma redução nas métricas de minutos jogados, jogos e jogos como titular. As tentativas de lançamento e os pontos marcados começam a diminuir ligeiramente, refletindo as alterações na capacidade física dos jogadores. Para além disso, o número de jogadores começa a cair de forma mais acentuada, o que sugere que apenas os jogadores de topo se conseguem manter na liga após esta idade. Entre os 31 e 33 anos, há apenas 356 jogadores ativos, o que reflete a potencial consideração de aposentadoria ou de mudança para ligas menos exigentes por parte dos jogadores. Na faixa de 34 a 36 anos, a quantidade reduz para 174 jogadores, o que confirma que a maioria dos atletas se aposentou ou migrou para outra liga. Entre os 37 e 39 anos, há apenas 66 jogadores ativos, indicando que a continuidade na NBA após os 36 anos é rara e que poucos atletas conseguem manter-se na liga por tanto tempo.

A partir dos 40 anos ou mais, observam-se reduções significativas nas métricas de desempenho. O número de jogos jogados, jogos a titular e minutos em campo é consideravelmente menor. Para além disso, o número de tentativas de lançamento e de pontos marcados atinge o ponto mais baixo, o que sugere que a perda de capacidades físicas, como agilidade, velocidade e impulso afetam a capacidade de os jogadores criarem situações favoráveis a lançamentos e, consequentemente, reduz a sua capacidade de pontuar. As restantes estatísticas de jogo permanecem relativamente estáveis comparativamente às faixas etárias

anteriores, o que sugere que, nos restantes aspetos do jogo, os jogadores conseguem compensar a perda de capacidades físicas com a experiência obtida. Esta faixa etária inclui apenas 13 jogadores, sendo esta a menor quantidade observada. Esta informação reflete as dificuldades crescentes em manter o nível de desempenho necessário para competir em alto nível, visto que, apenas um número muito restrito de jogadores se conseguem manter na liga até uma idade tão avançada.

Em suma, os jogadores da NBA geralmente atingem o auge do seu desempenho entre os 28 e os 30 anos, evidenciando a combinação ideal de habilidades técnicas e capacidade física. Após esta fase, inicia-se um declínio gradual que se torna bastante acentuado a partir dos 40 anos, refletindo as mudanças naturais associadas ao envelhecimento e a necessidade de ajuste das funções nas equipas.

Capítulo 5

Modelação

5.1 Metodologia

A metodologia adotada para o treino dos modelos de previsão seguiu uma abordagem sistemática e abrangente, com o objetivo de prever se a equipa da casa vence uma partida da NBA, selecionar os melhores algoritmos e otimizá-los para alcançar o máximo desempenho. Inicialmente, foram removidos do *dataset* os campos que representavam identificadores (data do jogo, identificador do jogo, identificadores das equipas e identificador da temporada em que o jogo ocorreu). Prosseguiu-se com a divisão dos dados em dois conjuntos: 70% dos dados foram utilizados para treino e 30% foram reservados para teste. É importante referir que os jogos estavam ordenados cronologicamente, dos mais antigos para os mais recentes, de forma a preservar a sequência temporal dos eventos. Este detalhe é importante, pois eventos desportivos não são eventos completamente independentes. O conjundo de treino é constituído por 11,9 temporadas da NBA, englobando 15250 jogos, ocorridos entre as temporadas 2005/2006 e 2015/2016 e numa parte significativa da temporada 2016/2017. Quanto ao conjunto de teste, é constituído por 5,1 temporadas, englobando 6536 jogos, ocorridos entre as temporadas 2017/2018 e 2021/2022 e na restante parte da temporada 2016/2017 não abrangida no conjunto de treino. Por fim, os dados foram normalizados em ambos os conjuntos, uma etapa essencial especialmente para os algoritmos KNN e SVM, que são sensíveis às escalas das variáveis e que utilizaram esses dados normalizados para o processamento e análise.

Foi realizada uma validação cruzada, apenas com o conjunto de treino, utilizando não só uma variedade de algoritmos base, incluindo RL, AD, NB, KNN, SVM, como também técnicas de *ensemble learning* como *Stacking*, *Voting Classifier*, *Bagging* (*Bagging Classifier* e RF) e *Boosting* (*AdaBoost*, GB, XGBoost e LGBM). A escolha dos algoritmos a serem utilizados como base nos algoritmos de *Stacking* e *Voting Classifier*, foi feita através de uma análise da matriz de correlação entre os algoritmos base, sendo selecionados aqueles com menor correlação entre si.

Posteriormente, os cinco melhores algoritmos foram selecionados com base na métrica Taxa de Acerto Balanceada (TAB), e em seguida, procedeu-se a uma otimização de parâmetros para esses algoritmos selecionados, recorrendo à técnica *RandomizedSearchCV*, com o intuito de aprimorar ainda mais o seu desempenho. Os parâmetros a serem otimizados, para cada algoritmo, estão descritos na Tabela 5.1. É importante referir que, para os algoritmos de *Stacking* e *Voting Classifier*, foram otimizados os parâmetros dos seus algoritmos base.

Apesar da diferença entre o número de jogos em que a equipa da casa ganhou para o número de jogos em que a equipa da casa perdeu não ser significativa, a opção sobre a métrica TAB deveu-se ao facto de se querer atingir uma avaliação mais justa e precisa do desempenho

Tabela 5.1: Parâmetros otimizados

Algoritmo	**Parâmetros**
RL	C *penalty* *solver* *max_iter*
AD	*max_depth* *min_samples_split* *min_samples_leaf* *max_features* *criterion*
NB	Não aplicável.
KNN	*n_neighbors* p *weights* *algorithm*
SVM	*kernel* C
Bagging	*estimator* *n_estimators* *max_samples* *max_features*
RF	*n_estimators* *max_depth* *min_samples_split* *min_samples_leaf* *max_features*
AdaBoost	Não aplicável, pois nunca ficou nos 5 melhores algoritmos.
GB	*learning_rate* *n_estimators* *max_depth*
XGBoost	*eta* *max_depth* *min_child_weight* *subsample* *colsample_bytree* *n_estimators*
LGBM	*num_leaves* *learning_rate* *n_estimators* *max_depth* *min_child_samples*

dos modelos, pois esta métrica ajuda a garantir que são capazes de classificar corretamente ambos os resultados e não apenas o resultado em maioria, ou seja, a vitória da equipa da casa. Por outras palavras, imagine-se que o *dataset* é constituído por 60% de jogos em que a equipa da casa ganhou e 40% de jogos em que a equipa da casa perdeu. Na eventualidade de um modelo classificar todas as instâncias como vitória da equipa da casa, seria atingida uma taxa de acerto de 60% porém, o modelo não seria, de facto, capaz de distinguir entre os dois resultados possíveis.

Após a otimização, foi conduzida uma nova validação cruzada, apenas com o conjunto de treino, utilizando os melhores algoritmos com os parâmetros otimizados.

Finalmente, os três melhores algoritmos foram escolhidos com base nos resultados da etapa

anterior, e uma validação usando a técnica *houldout* foi realizada, utilizando os parâmetros otimizados. A avaliação do desempenho dos modelos resultantes do *holdout* foi feita utilizando as métricas taxa de acerto, TAB, precisão, *F1* e *recall*, garantindo uma análise abrangente de diferentes aspetos do desempenho dos modelos:

- **Taxa de acerto:** representa a percentagem de jogos que o modelo previu corretamente em relação ao total de jogos previstos (Fórmula 5.1).
- **TAB:** Avalia a capacidade do modelo de prever corretamente ambos os resultados possíveis, isto é, vitória ou derrota da equipe da casa (Fórmula 5.2).
- **Precisão:** representa a percentagem de jogos que o modelo previu como vitória da equipa da casa em que, de fato, a equipa da casa ganhou (Fórmula 5.5).
- **Recall:** representa a percentagem de vitórias da equipa da casa corretamente identificadas pelo modelo, em relação ao total de vitórias da equipa da casa contidas no conjunto de dados (Fórmula 5.3).
- **F1:** combina a precisão e o *recall* para oferecer uma visão equilibrada do desempenho do modelo na previsão de vitórias da equipa da casa. É a média harmónica entre essas duas métricas e fornece uma medida única que leva em conta tanto a capacidade do modelo em prever vitórias corretas quanto a de identificar as vitórias reais, garantindo uma avaliação robusta e confiável do modelo (Fórmula 5.6).

A matriz de confusão é uma ferramenta usada para avaliar o desempenho de modelos de classificação, pois permite visualizar o desempenho do modelo em termos de acertos e erros, organizando os resultados das previsões numa tabela que compara as classes reais com as classes previstas pelo modelo. A matriz de confusão no contexto desta tese, está descrita na Tabela 5.2.

Tabela 5.2: Matriz de confusão para o modelo de previsão

	Previsão: Equipa da casa ganha	**Previsão: Equipa da casa perde**
Resultado Real : Equipa da casa ganhou	Verdadeiros Positivos (VP)	Falsos Negativos (FN)
Resultado Real : Equipa da casa perdeu	Falsos Positivos (FP)	Verdadeiros Negativos (VN)

$$\text{Taxa de acerto} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}} \tag{5.1}$$

$$\text{Taxa de Acerto Balanceada} = \frac{\text{Recall} + \text{Especificidade}}{2} \tag{5.2}$$

$$\text{Recall} = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{5.3}$$

$$\text{Especificidade} = \frac{\text{VN}}{\text{VN} + \text{FP}} \tag{5.4}$$

$$\text{Precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}} \tag{5.5}$$

$$F1 = 2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}} \tag{5.6}$$

Foram realizadas quatro iterações de treino de modelos, nas quais foi seguida a metodologia descrita:

- **Primeira iteração:** Na primeira iteração, foi feito o treino de modelos conforme novas características eram adicionadas ao conjunto de dados, de forma a avaliar o impacto que essas novas informações tinham na sua capacidade de previsão.
- **Segunda iteração**: Na segunda iteração, foi repetido o treino dos modelos, com os mesmos conjuntos, porém desta vez com seleção de características.
- **Terceira iteração**: Na terceira iteração, foram experimentados diferentes valores para o número de jogos anteriores a utilizar no cálculo da média das estatísticas nos últimos jogos e no cansaço das equipas.
- **Quarta iteração:** Na quarta iteração, foi avaliado o impacto da remoção de *outliers*, balanceamento do conjunto de dados e uso de validação cruzada estratificada na capacidade de previsão dos modelos.

Ao seguir esta metodologia rigorosa, foi possível identificar os melhores modelos de previsão para o problema em questão, garantindo a sua robustez e capacidade de generalização para dados futuros.

5.2 Primeira iteração

Nesta primeira iteração, foram treinados modelos com quatro diferentes conjuntos de dados:

- **Primeiro conjunto**: O primeiro conjunto incluía as médias das estatísticas das equipas, tanto no geral como em jogos em casa para a equipa anfitriã e em jogos fora para a equipa visitante, além da data de fundação das equipas.
- **Segundo conjunto**: O segundo conjunto continha todas as características do primeiro, mais a informação dos *rankings* nas equipas.
- **Terceiro conjunto**: O terceiro conjunto incluía todos os dados do segundo, além da inclusão do fator cansaço das equipas.
- **Quarto conjunto**: O quarto conjunto incorporava todas as características do terceiro, além de estatísticas adicionais das equipas, como ressaltos ofensivos e defensivos, roubos de bola, bloqueios, perdas de posse de bola, faltas cometidas, e o número de tentativas de lançamento e de lançamentos bem-sucedidos, tanto no geral, quanto em lançamentos três pontos e lançamentos livres. Embora essas estatísticas não estivessem presentes no ficheiro original dos jogos, foram calculadas posteriormente.

5.2.1 Primeiro conjunto

O treino de modelos com este conjunto ocorreu após a junção do ficheiro com a informação dos jogos com o ficheiro com a informação das equipas. A sua constituição pode ser consultada na tabela A.10, presente no anexo A.

Os modelos obtidos com este conjunto estão descritos na tabela 5.3.

Tabela 5.3: Modelos resultantes do treino com o primeiro conjunto

Modelo	**TAB**	**Taxa de acerto**	**Precisão**	**Recall**	**F1**	**Validação cruzada (Desvio padrão)**
RL	0.559	0.596	0.607	0.823	0.698	TAB: 0.592 (0.010)
LGBM	0.542	0.586	0.594	0.858	0.702	TAB: 0.582 (0.015)
GB	0.537	0.580	0.592	0.844	0.696	TAB: 0.584 (0.014)

O modelo RL apresentou a melhor TAB, taxa de acerto e precisão. Em contrapartida, em relação à métrica *recall*, foi o que apresentou pior desempenho.

Todos os modelos apresentaram um desempenho bastante equivalente na métrica F1.

Pelos motivos apresentados, nesta fase, o modelo RL seria a melhor escolha. Não obstante, as métricas deste modelo são bastante baixas, à exceção do *recall* e do F1. O facto de o modelo apresentar um *recall* de 82,3% mas uma precisão de apenas 60,7%, sugere que o modelo poderá estar a classificar praticamente todas as instâncias como vitória da equipa da casa.

5.2.2 Segundo conjunto

O treino de modelos com este conjunto ocorreu após a junção do ficheiro de *rankings* (Subsecção 3.2.5). A constituição do conjunto pode ser consultada na Tabela A.11, presente no anexo A.

Os modelos obtidos com este conjunto estão descritos na tabela 5.4.

Tabela 5.4: Modelos resultantes do treino com o segundo conjunto

Modelo	**TAB**	**Taxa de acerto**	**Precisão**	**Recall**	**F1**	**Validação cruzada (Desvio padrão)**
VCH com NB e RL	0.624	0.623	0.687	0.619	0.651	TAB: 0.640 (0.010)
VCS com NB, RL e SVM	0.620	0.634	0.664	0.720	0.691	TAB: 0.643 (0.006)
VCS com NB e RL	0.619	0.627	0.669	0.677	0.673	TAB: 0.642 (0.007)

O modelo VCS com NB, RL e SVM apresentou a maior taxa de acerto (63,4%), *recall* (72%) e F1 (69,1%), sendo esta última métrica um bom indicador de que existe um bom equilíbrio entre precisão e *recall*.

Todos os modelos apresentaram valores muito semelhantes para a métrica TAB.

Nesta fase, o modelo VCS com NB, RL e SVM seria a melhor escolha devido à sua taxa de acerto, *recall* e F1. Este modelo apresenta um conjunto equilibrado de pontos fortes e fracos que são cruciais para o problema em causa. Entre os pontos fortes, destacam-se a precisão

(66.4%) e o *recall* (72.0%), que indicam que o modelo é eficaz em identificar corretamente as vitórias da equipa da casa, com uma boa percentagem de acerto. O F1 de 69,1% reflete um bom equilíbrio entre essas métricas, tornando-o robusto em termos de desempenho geral. No entanto, a taxa de acerto (63.4%) e a TAB (62.0%) são relativamente baixas, sugerindo que ainda há espaço para melhorias.

Comparativamente ao treino de modelos com o conjunto anterior, foi notada uma clara melhoria, visto que todos os modelos desta fase apresentaram um desempenho em muito superior ao melhor modelo obtido com o primeiro conjunto. Deste modo, foi possível concluir que a inclusão do *ranking* das equipas resultou numa grande melhoria no desempenho das previsões, o que sugere que este é um fator relevante para o sucesso dos modelos preditivos.

5.2.3 Terceiro conjunto

O treino de modelos com este conjunto ocorreu após a implementação do factor de cansaço nas equipas (Subsecção 3.2.6). A constituição deste conjunto pode ser consultada na Tabela A.12, presente no Anexo A.

Os modelos obtidos com este conjunto estão descritos na Tabela 5.5.

Tabela 5.5: Modelos resultantes do treino com o terceiro conjunto

Modelo	**TAB**	**Taxa de acerto**	**Precisão**	**Recall**	**F1**	**Validação cruzada (Desvio padrão)**
VCH com NB e RL	0.625	0.621	0.693	0.597	0.641	TAB: 0.639 (0.010)
VCS com NB, RL e KNN	0.625	0.637	0.669	0.714	0.691	TAB: 0.638 (0.010)
VCS com NB e RL	0.624	0.630	0.677	0.666	0.671	TAB: 0.642 (0.009)

O modelo VCS com NB, RL e KNN apresentou a maior taxa de acerto (63,7%), o maior *recall* (71,4%) e o maior F1 (69,1%), sendo a precisão a única métrica onde não teve o melhor desempenho. Todos os modelos apresentaram valores muito semelhantes para a métrica TAB.

Pelos motivos acima mencionados, este modelo seria a melhor escolha. Este modelo apresenta um conjunto equilibrado de pontos fortes e fracos que são cruciais para o problema da previsão de vitória da equipa da casa na NBA. Entre os pontos fortes, destacam-se a precisão de 66,9% e *recall* de 71,4%, o que sugere que o modelo é eficaz a prever vitórias da equipa da casa, com uma boa proporção de acertos. O F1 de 69,1% confirma que há um bom equilíbrio entre estas duas métricas. No entanto, a taxa de acerto de 63,7% e a TAB de 62,5%, embora aceitáveis, podem ser vistas como relativamente baixas, indicando que há espaço para melhorias no desempenho geral das previsões.

Ao comparar o modelo VCS com NB, RL e KNN com o melhor modelo obtido com o conjunto anterior, o primeiro emerge como a melhor escolha. Apesar de ambos apresentarem um F1 idêntico, existe uma melhoria, apesar de residual, das métricas TAB, taxa de acerto e precisão.

A inclusão do cansaço das equipas resultou numa ligeira melhoria no desempenho das previsões, o que sugere que este é um fator relevante no desempenho das equipas, fornecendo *insights* valiosos que melhoram o desempenho dos modelos preditivos.

5.2.4 Quarto conjunto

O treino de modelos com este conjunto ocorreu após a inclusão de estatísticas de equipa que não estavam presentes no ficheiro dos jogos (Subsecção 3.2.7). A constituição deste conjunto pode ser consultada na Tabela A.13, presente no Anexo A.

O desempenho dos modelos obtidos com este conjunto estão descritos na Tabela 5.6.

Tabela 5.6: Modelos resultantes do treino com o quarto conjunto

Modelo	**TAB**	**Taxa de acerto**	**Precisão**	**Recall**	**F1**	**Validação cruzada (Desvio padrão)**
VCS com NB e RL	0.630	0.630	0.691	0.626	0.657	TAB: 0.641 (0.012)
VCS com NB, RL e KNN	0.628	0.633	0.678	0.669	0.674	TAB: 0.643 (0.011)
VCH com NB e RL	0.628	0.619	0.705	0.562	0.625	TAB: 0.637 (0.014)

O modelo VCS com NB, RL e KNN apresentou os melhores valores em todas as métricas, excetuando a precisão (onde foi o modelo com o valor mais baixo) e TAB, onde ficou empatado com o modelo VCS com NB e RL. Apesar de ser o modelo com a precisão mais baixa, o seu valor de *recall* acaba por compensar essa deficiência conferindo-lhe um maior F1.

Pelos motivos acima mencionados, para este conjunto, este modelo seria a melhor escolha. Apresenta um conjunto equilibrado de pontos fortes e fracos que são cruciais para o problema da previsão de vitória da equipa da casa na NBA. Entre os pontos fortes, destacam-se a precisão de 67,8% . O F1 de 67,4% confirma que há um bom equilíbrio entre estas duas métricas. No entanto, a taxa de acerto de 63,3% e a TAB de 62,8%, embora aceitáveis, podem ser vistas como relativamente baixas, indicando que há espaço para melhorias no desempenho geral das previsões.

Comparativamente ao melhor modelo obtido com o conjunto anterior, considerou-se que o modelo VCS com NB, RL e KNN apresentou um desempenho pior, pois, apesar de uma ligeira melhoria nas métricas TAB e precisão, houve um decréscimo das métricas taxa de acerto, *recall* e F1, sendo estes decréscimos mais significativos do que os aumentos que se verificaram. Deste modo, considera-se que as características incluídas não contribuíram positivamente para a capacidade preditiva dos modelos.

No final da primeira iteração, o melhor modelo obtido até então foi aquele que utilizava o terceiro conjunto de dados e o algoritmo VCS com NB, RL e KNN como algoritmos base.

5.3 Segunda iteração

Na segunda iteração, foi repetido o treino feito na primeira iteração, com os mesmos quatro conjuntos de dados, porém fazendo seleção de características em cada um destes conjuntos.

Este processo de seleção de características, para cada um dos quatro conjuntos, seguiu os seguintes passos:

1. Foram aplicadas 3 técnicas de seleção de características:

- LASSO, com o objetivo de identificar características relevantes, selecionando aquelas cujos coeficientes não eram nulos;
- um método baseado em AD, onde características eram escolhidas com base na sua importância;
- embaralhamento de características, que avaliava o impacto da remoção de cada característica no desempenho do modelo;

2. Após aplicar cada uma destas técnicas, foram obtidos três subconjuntos, além do conjunto original. Para cada um desses quatro subconjuntos (o original e os três resultantes das técnicas), foi realizada validação cruzada utilizando os seguintes algoritmos de ML: *Bagging*, GB, XGBoost e LGBM. O objetivo foi identificar qual subconjunto apresentava o melhor desempenho global, tendo em conta o equilíbrio entre o desempenho e o número de características.
3. Utilizando o melhor subconjunto, foram treinados modelos, seguindo a mesma metodologia da iteração anterior.

Na tabela 5.7 estão descritos os desempenhos dos melhores modelos obtidos com cada conjunto de dados. É importante relembrar que o primeiro conjunto inclui médias de estatísticas gerais e específicas (casa/fora) nos últimos jogos e a data de fundação das equipas, o segundo adiciona os *rankings* das equipas, o terceiro implementa o fator cansaço e o quarto acrescenta estatísticas adicionais. Note-se que a coluna "Características"representa o número de características (excluindo a variável alvo) do conjunto de dados, após a seleção. Por fim, é importante mencionar que o treino dos modelos foi realizado recorrendo aos algoritmos base, *Stacking*, *Voting Classifier*, *Bagging* e *Boosting*.

Tabela 5.7: Melhores modelos obtidos na segunda iteração

Conjunto	Algoritmo	Métricas	Validação cruzada	Características	Seleção
1	RL	TAB: 0.582 Taxa de acerto: 0.614 Precisão: 0.619 *Recall*: 0.828 F1: 0.708	TAB: 0.581 Desvio padrão: 0.010	14	AD
2	LGBM	TAB: 0.600 Taxa de acerto: 0.631 Precisão: 0.632 *Recall*: 0.835 F1: 0.719	TAB: 0.629 Desvio padrão: 0.013	29	AD
3	RL	TAB: 0.612 Taxa de acerto: 0.641 Precisão: 0.641 *Recall*: 0.831 F1: 0.724	TAB 0.631 Desvio padrão::0.009	41	AD
4	VCS com RL, KNN e NB	TAB 0.630 Taxa de acerto: 0.634 Precisão: 0.683 *Recall*: 0.659 F1: 0.671	TAB: 0.643 Desvio padrão: 0.009	61	AD

O modelo relativo ao terceiro conjunto apresentou a maior taxa de acerto (64,1%) e F1 (72,4%).

O modelo relativo ao quarto conjunto teve a melhor TAB (63.0%), porém apresentou o pior F1 e *recall*, sendo esta última extremamente inferior ao obtido pelos restantes modelos.

O modelo relativo ao primeiro conjunto utilizou o menor número de características (14). O modelo obtido com o segundo conjunto utilizou 29 características, com o terceiro conjunto 41 características e com o quarto conjunto 61 características.

Considerando o desempenho e a eficiência, o modelo obtido com o terceiro conjunto destacou-se como a melhor escolha, alcançando a maior taxa de acerto e F1 entre todos os modelos avaliados. Embora tenha apresentado TAB e precisão inferiores ao modelo obtido com o quarto conjunto, superou significativamente este último em *recall* e F1. Apesar de utilizar mais características do que os os modelos obtidos com os conjuntos 1 e 2, o seu desempenho superior justifica essa diferença.

Para além disso, este modelo representou uma melhoria comparativamente ao, até então, melhor modelo, pois, apesar de valores de TAB e precisão um pouco inferiores, apresentou melhorias relativamente à taxa de acerto, *recall* e F1, utilizando apenas 41 características (excluindo a variável alvo), o que representa uma redução de cerca de 45% no número de características. Esta redução torna o modelo mais simples e fácil de utilizar.

Deste modo, a escolha do conjunto de dados a utilizar nas iterações seguintes recaiu sobre o terceiro conjunto, uma vez que foi o que apresentou melhores resultados.

5.4 Terceira iteração

(Chen et al. 2021) enfatiza a importância da escolha apropriada do número de jogos anteriores a utilizar na obtenção de melhores resultados. Até agora, para calcular a média das estatísticas e o cansaço das equipas , haviam sido usados os últimos 4 jogos. No entanto, nesta terceira iteração do treino de modelos, foram testados diferentes números de jogos anteriores a utilizar, com o objetivo de verificar qual valor apresentava melhores resultados.

Para este propósito, foram utilizadas diferentes versões do conjunto de dados que transitou da iteração anterior. Em cada uma destas versões, a única diferença entre os dados estava nas colunas das médias das estatísticas das equipas nos últimos jogos, que variavam conforme o número de jogos considerados para as calcular. É importante referir que, para cada uma destas versões, foi feita a seleção de características, seguindo o mesmo processo utilizado na iteração anterior.

Na tabela 5.8 estão descritos os melhores modelos obtidos com cada número de jogos anteriores utilizados. Note-se que a coluna "Características"representa o número de características (excluindo a variável alvo) do conjunto de dados, após efetuar a seleção.

O modelo que utiliza 5 jogos para o cálculo da médias das estatísticas das equipas apresentou a melhor taxa de acerto (64,2%), sendo este valor bastante semelhante ao obtido pelo modelo que utiliza 4 jogos (64,1%).

O modelo que utiliza 7 jogos apresentou a melhor TAB e a melhor precisão com 62,6% e 67,3%, respetivamente, porém apresentou, de longe, o pior *recall*, com apenas 68,7%.

No que diz respeito à métrica *recall*, houve um empate entre o modelo que usa 4 jogos e o modelo que usa 6 jogos (83,1%).

O modelo que utiliza 4 jogos registou o melhor F1 com 72,4%, porém esta métrica foi bastante equilibrada entre todos os modelos, excetuando o modelo que usa 7 jogos.

A melhor escolha recaiu entre os modelos que utilizam 4 e 5 jogos, devido especialmente ao seu desempenho nas métricas taxa de acerto e F1. A diferença de desempenho entre os dois

Tabela 5.8: Melhores modelos obtidos com diferente número de jogos para calcular a média das estatísticas

Jogos	Algoritmo	Métricas	Validação cruzada	Características	Seleção
4	RL	TAB: 0.612 Taxa de acerto: 0.641 Precisão: 0.641 *Recall*: 0.831 F1: 0.724	TAB: 0.631 Desvio padrão: 009	41	AD
5	RL	TAB: 0.615 Taxa de acerto: 0.642 Precisão: 0.644 *Recall*: 0.823 F1: 0.723	TAB: 0.630 Desvio padrão: 010	40	AD
6	GB	0.602 Taxa de acerto: 0.632 Precisão: 0.634 *Recall*: 0.831 F1: 0.719	TAB: 0.632 Desvio padrão: 0.013	39	AD
7	VCS com RL e NB	TAB: 0.626 Taxa de acerto: 0.634 Precisão:0.673 *Recall*: 0.687 F1: 0.680	TAB: 0.647 Desvio padrão: 0.011	39	AD

modelos foi mínima, sendo que o modelo que utiliza 4 jogos leva uma pequena vantagem em *recall* e F1, enquanto que o modelo que utiliza 5 jogos leva uma pequena vantagem nas métricas TAB, taxa de acerto e precisão. Ambos os modelos seriam uma escolha válida, porém optou-se por escolher o modelo que utiliza 5 jogos, pela preferência pelas métricas TAB e taxa de acerto.

Além de determinar o melhor número de jogos anteriores a serem utilizados no cálculo das médias das estatísticas das equipas nos últimos jogos, nesta fase foi também avaliado o melhor número de jogos anteriores para calcular as características relacionadas ao cansaço das equipas. Para isso, foram testadas diferentes quantidades de jogos anteriores. É importante destacar que, neste contexto, foram utilizados os últimos 5 jogos para o cálculo das médias das estatísticas das equipas, pois foi o valor que demonstrou os melhores resultados no treino anterior. Foram treinados modelos com diferentes versões do conjunto de dados, onde a única variação se encontrava nas colunas relacionadas com o cansaço das equipas, que variavam conforme o número de jogos considerados. A cada uma destas versões do conjunto de dados, foi aplicada seleção de características.

Na tabela 5.9 estão descritos os melhores modelos obtidos com cada quantidade de jogos a utilizar para o cálculo do cansaço das equipas.

O modelo que utiliza 4 jogos para o cálculo do cansaço das equipas apresentou a melhor taxa de acerto, com o valor 64,2%.

O modelo que utiliza 7 jogos para o cálculo do cansaço das equipas obteve a melhor TAB e precisão, com os valores 62% e 65,4%, respetivamente. Apesar do bom desempenho nestas duas métricas, este modelo foi o que apresentou pior desempenho nas métricas *recall* e F1. Em relação à métrica taxa de acerto, este modelo apresentou um desempenho bastante semelhante ao modelo que utiliza 4 jogos, com uma diferença de apenas 0,1%.

O modelo que utiliza 5 jogos apresentou o melhor desempenho nas métricas *recall* e F1, porém a sua precisão e TAB foram as segundas mais baixas, com apenas 63,6% e 60,8%,

Tabela 5.9: Melhores modelos obtidos com diferente número de jogos para calcular o cansaço

Jogos	Algoritmo	Métricas	Validação cruzada	Características	Seleção
4	RL	0.615 Taxa de acerto: 0.642 Precisão:0.644 *Recall*: 0.823 F1: 0.723	TAB: 0.630 Desvio padrão: 0.010	40	AD
5	*Stacking* Base: AD, NB, SVM Estimador Final: SVM	TAB: 0.608 Taxa de acerto: 0.640 Precisão: 0.636 *Recall*: 0.851 F1: 0.728	TAB: 0.630 Desvio padrão: 0.019	41	AD
6	GB	TAB: 0.602 Taxa de acerto: 0.632 Precisão: 0.633 *Recall*: 0.832 F1: 0.719	TAB: 0.632 Desvio padrão: 0.011	42	AD
7	*Stacking* Base: AD, NB Estimador Final: RL	TAB: 0.620 Taxa de acerto:0.641 Precisão: 0.654 *Recall*: 0.777 F1: 0.710	TAB: 0.632 Desvio padrão: 0.018	40	AD

respetivamente. Relativamente à taxa de acerto, foi a terceira melhor, mas com uma diferença de apenas 0,2% em relação à melhor taxa de acerto obtida.

A melhor escolha recaiu entre os modelos que utilizam 4 e 7 jogos, devido especialmente ao seu desempenho nas métricas TAB e taxa de acerto. A diferença de desempenho entre ambos os modelos não foi significativa, visto que um apresentava melhor desempenho em algumas métricas, enquanto o outro apresentava melhor desempenho nas restantes. Ambos os modelos seriam uma escolha válida, porém optou-se por escolher o modelo que utiliza 4 jogos, devido ao seu desempenho nas métricas F1 e *recall*, métricas essas onde se observou maior diferença entre os dois modelos.

O melhor modelo obtido nesta iteração apresentou uma melhoria relativamente ao até então melhor modelo.

5.5 Quarta iteração

Na quarta iteração do treino de modelos, foram aplicados balanceamento dos dados e remoção de *outliers* com o objetivo de melhorar o desempenho dos modelos preditivos. Deste modo, foram treinados modelos, recorrendo a todas as combinações possíveis, passando a enumerar:

- Sem balanceamento dos dados e sem *remoção de outliers*
- Apenas com remoção de *outliers*
- Apenas com balanceamento dos dados
- Com balanceamento dos dados e remoção de *outliers*

É importante referir que o conjunto de dados utilizado foi exatamente igual ao utilizado para o treino do até então melhor modelo, obtido na terceira iteração, inclusive as características selecionadas. Para além disso, é também, importante referir que, nos modelos onde foi

aplicado balanceamento dos dados não foi avaliada a métrica TAB. Por este motivo, nesta iteração, a métrica taxa de acerto foi utilizada como critério na validação cruzada, de forma a manter a consistência no treino dos modelos. Finalmente, todas estas combinações foram também testadas recorrendo a validação cruzada estratificada, em detrimento da validação cruzada tradicional.

Na tabela 5.10 estão descritos os melhores modelos obtidos para cada combinação de fatores.

Tabela 5.10: Melhores modelos obtidos com cada combinação de fatores

Validação	**Modelo**	**Algoritmo**	**TAB**	**Taxa de acerto**	**Precisão**	**Recall**	**F1**	**Validação cruzada (Desvio padrão)**
Tradicional	Nenhum	RL	0.616	0.643	0.645	0.821	0.723	0.666 (0.019)
	Sem *outliers*	RL	0.622	0.637	0.661	0.734	0.696	0.662 (0.013)
	Balanceado	XGBoost	-	0.567	0.662	0.481	0.557	0.746 (0.013)
	Ambos	LGBM	-	0.613	0.648	0.690	0.669	0.670 (0.016)
Estratificada	Nenhum	RL	0.617	0.644	0.645	0.825	0.724	0.669 (0.017)
	Sem *outliers*	RL	0.609	0.638	0.640	0.824	0.720	0.664 (0.012)
	Balanceado	XGBoost	-	0.565	0.665	0.464	0.574	0.747 (0.012)
	Ambos	LGBM	-	0.612	0.659	0.655	0.657	0.674 (0.015)

Em ambos os tipos de validação, o modelo que não implementa nem balanceamento dos dados, nem remoção de *outliers* demonstrou o melhor desempenho, destacando-se com a maior taxa de acerto, *recall* e F1. Comparando os dois tipos de validação, a vantagem recai para a utilização de validação cruzada estratificada, pois o modelo treinado com esta abordagem foi superior, apesar de diferenças muito residuais, ao modelo treinado recorrendo à validação cruzada tradicional em todas as métricas, excetuando a precisão, onde ambos os modelos obtiveram o mesmo valor.

Pelos motivos acima descritos, e pelo facto de haver uma melhoria relativamente ao melhor modelo da iteração passada, a escolha sobre o modelo a implementar num *website* recaiu sobre o modelo que não implementa nenhuma das anteriores técnicas e que recorreu à validação cruzada estratificada. Na tabela A.14, presente no anexo A estão descritas as variáveis que este modelo utiliza para fazer as previsões.

5.6 Implementação do modelo num website

O modelo foi disponibilizado num *website* [1] utilizando a *framework Streamlit* (Streamlit 2024), que permite desenvolver aplicações *web* interativas de forma rápida e com código mínimo. Os dados necessários para realizar as previsões são fornecidos pela NBA-API (NBA-API s.d.). A plataforma *Streamlit* integra-se com o *GitHub*, instala automaticamente todas as dependências e disponibiliza a aplicação *online* de forma gratuita.

A arquitetura pretendida (Figura 5.1) para a aplicação *web* desenvolvida é uma arquitetura simples e integrada. Composta por um único *script Python*, esta aplicação desempenha tanto as funções de *frontend* como de *backend*. A interação com o utilizador ocorre através de uma interface *web* intuitiva, criada com *Streamlit*. O *script* faz requisições à *NBA-API* para recolher os dados necessários, que são posteriormente processados e transformados

[1] https://nbaprevision.streamlit.app/

para alimentar o modelo de previsão. As previsões geradas são, então, apresentadas ao utilizador na interface gráfica.

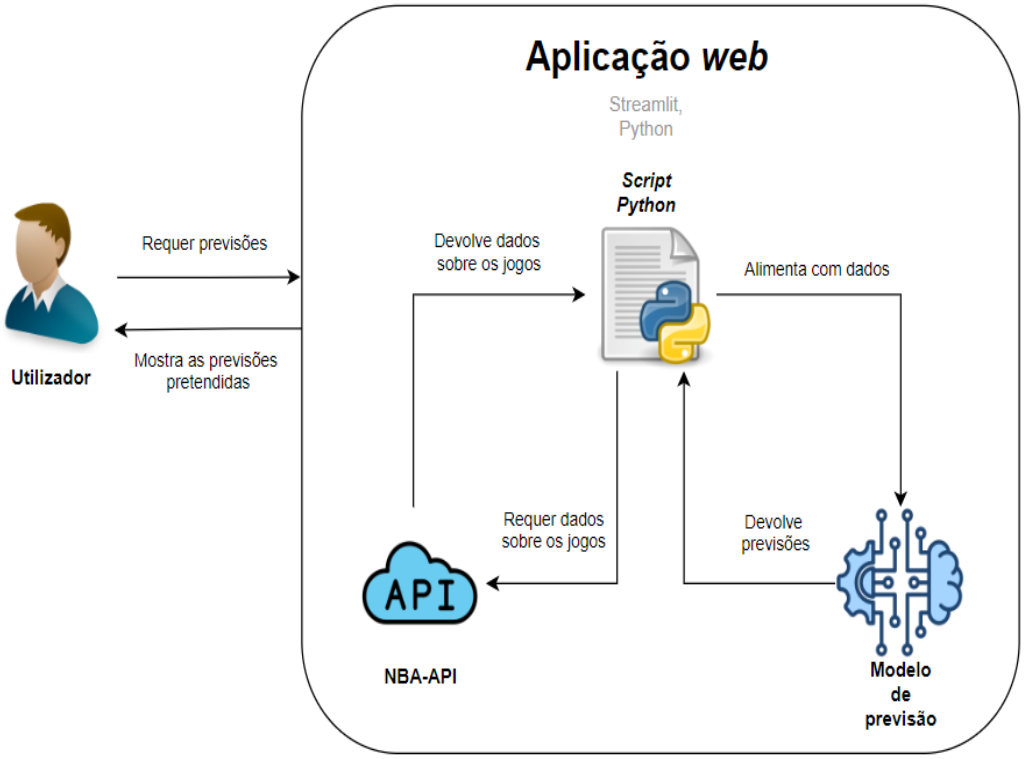

Figura 5.1: Arquitetura pretendida para a aplicação *web*

A implementação desta arquitetura encontrou obstáculos, devido às limitações da *NBA-API*. A *API* não disponibiliza um *endpoint* para obter diretamente todos os dados dos jogos ocorridos até a data pretendida; em vez disso, fornece apenas um *endpoint* que retorna os identificadores dos jogos passados. Para obter os dados completos de cada jogo, era necessário realizar múltiplos pedidos, um para cada jogo, através de um *endpoint* que recebe o identificador de um jogo e devolve os seus dados. Este processo era ineficiente, pois envolvia múltiplos pedidos consecutivos à *API*, resultando num tempo de processamento muito longo e num risco elevado de atingir limites de taxa, além de aumentar o tempo de resposta, o que poderia levar a *timeouts*, comprometendo assim a obtenção dos dados necessários para as previsões.

Para ultrapassar as limitações da arquitetura pretendida, optou-se por obter previamente os dados dos jogos, através da *NBA-API*, e armazená-los em ficheiros *Excel*. Com esta abordagem, a arquitetura foi modificada para utilizar esses ficheiros em vez de depender diretamente da *API*. Embora a estrutura geral da aplicação permaneça a mesma, esta modificação permitiu evitar os problemas mencionados previamente, garantindo um acesso mais rápido e confiável aos dados necessários para as previsões. A arquitetura final da aplicação pode ser visualizada na Figura 5.2.

Apesar das vantagens desta abordagem, existem também limitações. A principal é que os ficheiros *Excel* precisam de ser atualizados manualmente, de forma a incluir novos jogos e dados recentes. Isto pode resultar em atrasos ou na falta de dados atualizados, o que

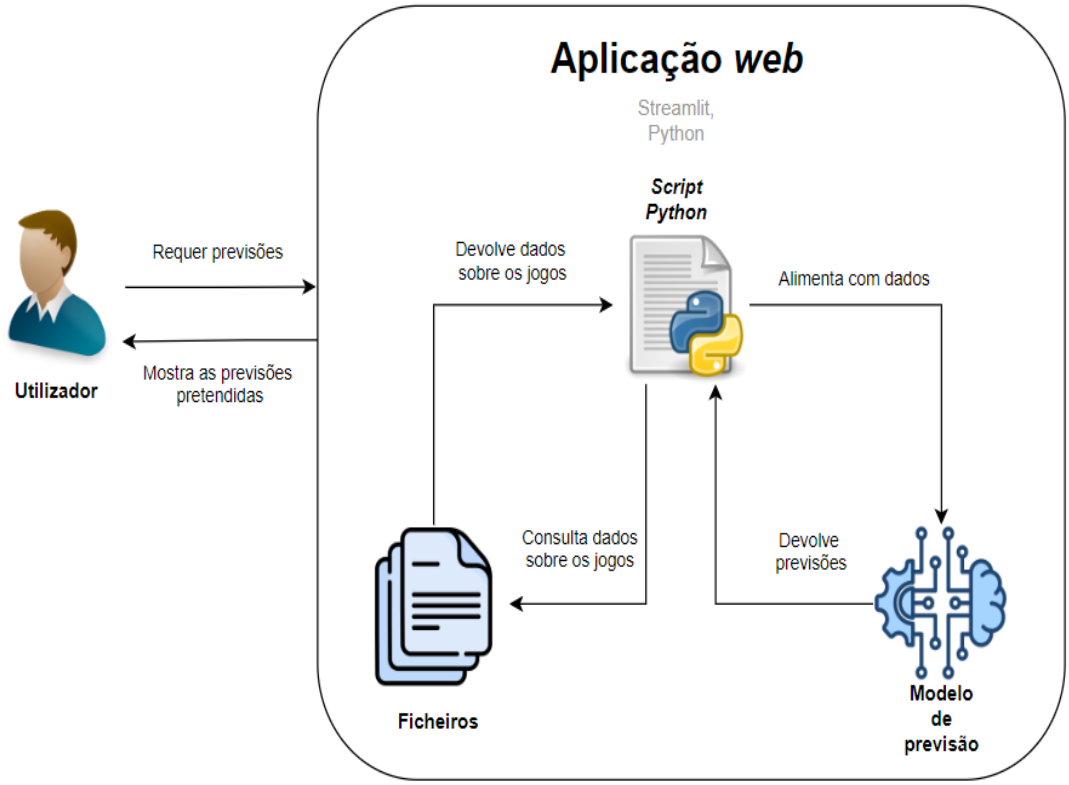

Figura 5.2: Arquitetura final da aplicação *web*

pode impactar a precisão das previsões. Como trabalho futuro, propõe-se a atualização automática destes ficheiros, através de um *script*. Com esta melhoria, a aplicação poderá manter-se sempre atualizada com os dados mais recentes e sem a necessidade de intervenção manual.

A aplicação foi projetada para realizar o caso de uso principal: gerar previsões (Figura 5.3). A interface intuitiva e fácil de usar permite que os utilizadores selecionem a data desejada e visualizem as previsões dos jogos programados para esse dia. Além das previsões futuras, o utilizador pode também visualizar previsões passadas, de forma a comparar com os resultados reais. O fluxo que o utilizador segue para obter as previsões está descrito na Tabela 5.11. A Figura 5.4 demonstra a exibição das previsões na interface gráfica da aplicação.

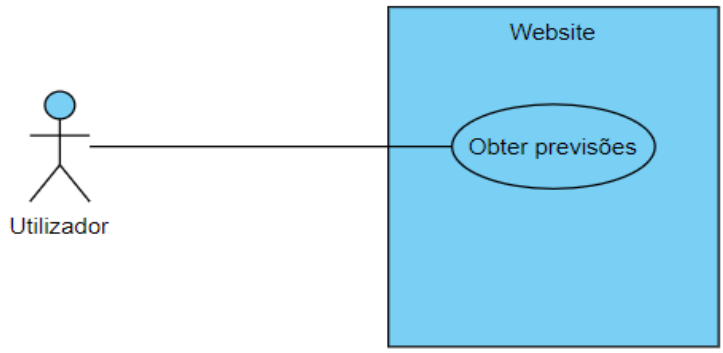

Figura 5.3: Casos de Uso da aplicação

Tabela 5.11: Caso de Uso Obter previsões

ID	UC001
Nome	Obter previsões
Ator	Utilizador
Descrição	O utilizador pretende obter as previsões para os jogos da NBA de determinado dia.
Pré-condições	Não aplicável
Fluxo	1: O utilizador seleciona a data para a qual quer obter previsões. 2: O utilizador carrega no botão "Get predictions". 3: A aplicação gera e exibe as previsões para os jogos agendados para o dia selecionado.
Pós-condições	As previsões para os jogos da NBA agendados para o dia selecionado são apresentadas ao utilizador.
Fluxo alternativo	1: O utilizador seleciona uma data sem jogos agendados 2: A aplicação exibe uma mensagem a informar que não há jogos agendados para o dia selecionado.

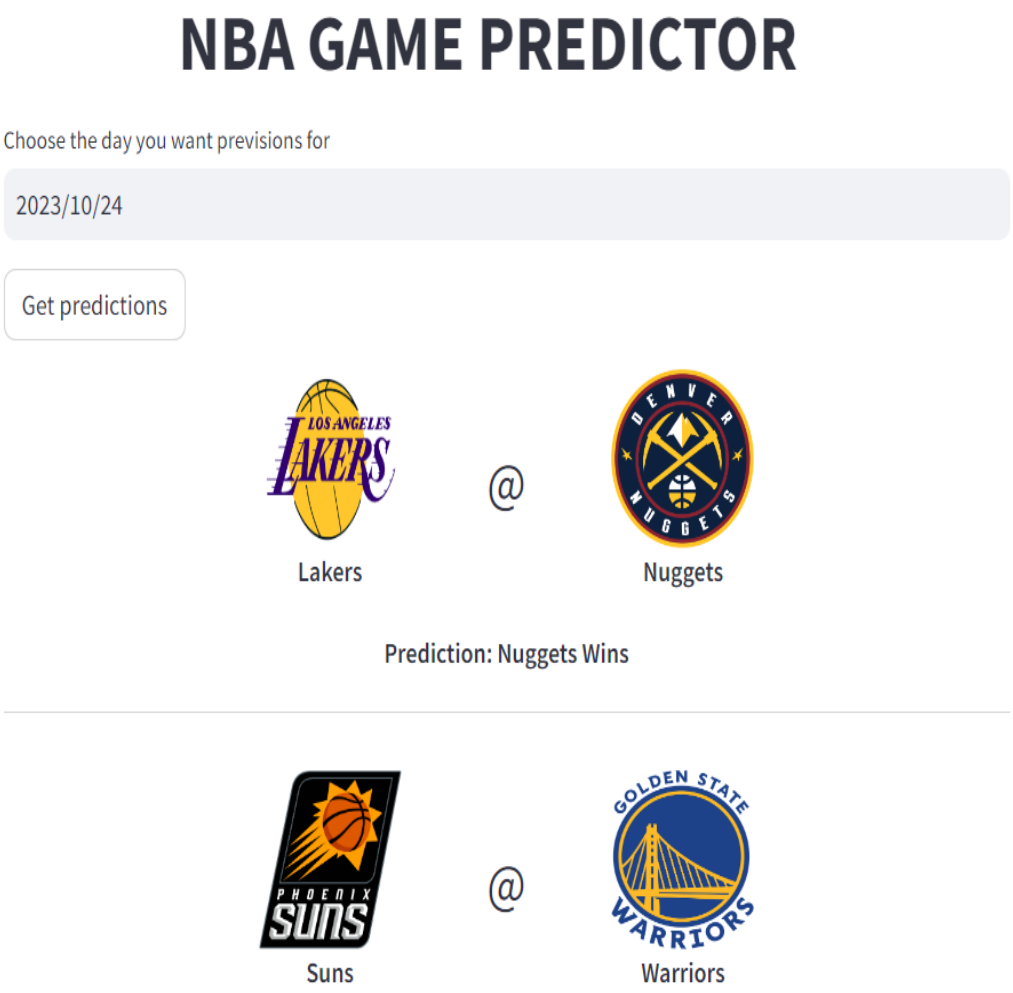

Figura 5.4: Previsões realizadas no *website*

Capítulo 6

Análise de Resultados

Prever o desfecho de uma partida da NBA com precisão é uma tarefa desafiadora, muito devido à sua imprevisibilidade, que pode ser atribuída a vários fatores. Primeiramente, o *basketball* é um desporto onde qualquer equipa pode vencer num determinado dia, independentemente da sua forma recente. Além disso, existem métricas e variáveis que são difíceis de quantificar e que afetam diretamente o resultado dos jogos. Fatores como o *"clutch factor"*, que se refere ao desempenho em momentos críticos do jogo, e as oscilações de forma ("cold"e "hot streaks"), tanto para as equipas, como para os jogadores, podem influenciar os resultados de maneira imprevisível. A presença de lesões e o fator arbitragem também introduzem variáveis que são quase impossíveis de prever com precisão. Outro aspecto importante é a sorte, que pode desempenhar um papel significativo em qualquer jogo. Dada a complexidade e a quantidade de variáveis envolvidas, desde estatísticas e desempenhos individuais até circunstâncias imprevistas, a previsibilidade dos resultados na NBA é intrinsecamente limitada.

O modelo escolhido apresentou um desempenho moderado, com uma taxa de acerto de 64,4% e F1=72,4%. Estas métricas sugerem que, embora o modelo acerte a maioria das previsões, ainda há margem para melhorias. O recall elevado de 82,5% indica que o modelo é eficaz a identificar vitórias da equipa da casa, no entanto, a precisão de 64,5% revela que há um número significativo de falsos positivos, o que compromete a fiabilidade das previsões. A média harmónica destas duas métricas é refletida no F1 de 72,4%, que, embora razoável, evidencia a necessidade de um equilíbrio mais robusto entre estas métricas.

Uma das limitações do modelo é a ausência do fator de desempenho individual dos jogadores. A abordagem pretendida era utilizar as médias das estatísticas dos últimos jogos para cada jogador que seria titular na partida. No entanto, isso não foi possível, pois os alinhamentos das equipas apenas são confirmados muito próximo da hora de início dos jogos. Como alternativa, considerou-se utilizar a moda dos alinhamentos dos últimos jogos para estimar quais jogadores estariam em campo. Contudo, esta alternativa não era completamente viável, pois poderiam ocorrer alterações de última hora no cinco inicial, resultando na utilização de dados incorretos pelo modelo. O desempenho individual dos jogadores é um fator importante que poderia ter melhorado a capacidade preditiva do modelo, uma vez que pode influenciar consideravelmente o resultado das partidas.

Foi feita uma análise de como seria o desempenho do modelo ao longo da época 2023/2024 da NBA, época em vigor na altura da escrita desta tese. É importante referir que esta época não estava contida no conjunto de dados usados no treino do modelo. Os resultados desta análise estão descritos na Tabela 6.1. Para além disso, foi também analisado o desempenho do modelo para a época regular e para os *playoffs* desta temporada, estando os resultados desta análise descritos na Tabela 6.2.

Tabela 6.1: Desempenho do modelo com o decorrer da época 2023/2024

Mês	Número de jogos	Jogos acertados	Taxa de acerto
outubro	54	29	53,7%
novembro	219	142	64,84%
dezembro	208	141	67,79%
janeiro	231	157	67,97%
fevereiro	174	106	60,92%
março	230	147	63,91%
abril	157	106	67,52%
maio	41	25	60,98%
junho	5	4	80%
Total	1319	857	64,97%

Tabela 6.2: Desempenho do modelo para a época regular e *playoffs* 2023/2024

Tipologia	Número de jogos	Jogos acertados	Taxa de acerto
Época regular	1237	800	64,67%
Playoffs	82	57	69,51%

Ao analisar a taxa de acerto do modelo ao longo da época, várias tendências significativas emergem quando se considera o contexto da temporada. A época começou a 24 de outubro, o que explica o número reduzido de jogos nesse mês e a baixa taxa de acerto de 53,7%, possivelmente devido à imprevisibilidade típica do início da temporada, quando as equipas ainda estão a ajustar-se. Em fevereiro, a paragem para o *All-Star Weekend* pode ter contribuído para a queda no desempenho do modelo, que apresentou uma taxa de acerto de 60,92%, o que sugere que a interrupção na regularidade dos jogos e possíveis alterações no ritmo das equipas após o retorno da pausa impactou a capacidade preditiva do modelo. O início dos *playoffs* a 20 de abril não resultou numa diminuição significativa no desempenho, com o modelo a manter uma taxa de acerto de 67,52% nesse mês. Nos meses finais, maio e junho, o modelo continuou a apresentar resultados consistentes, apesar do número reduzido de jogos, com taxas de acerto de 60,98% e 80%, respetivamente. No entanto, é importante notar que o alto desempenho em junho deve ser interpretado com cautela devido ao pequeno número de jogos, mas pode indicar que o modelo se saiu bem na previsão dos resultados finais. É interessante notar que, nos *playoffs*, a taxa de acerto do modelo foi de 69,51%, superior à taxa de 64,67% obtida durante a época regular, o que pode indicar uma melhor adaptação do modelo às dinâmicas das equipas nas fases decisivas. No geral da época, o modelo alcançou uma taxa de acerto de 64,97% em 1319 jogos, indicando um desempenho razoável ao longo de toda a temporada e em linha com os resultados obtidos no Capítulo 5. Observa-se que o desempenho do modelo variou ao longo dos meses, mas manteve uma tendência geral de estabilização e melhoria após um início mais incerto, destacando-se especialmente durante os *playoffs*.

Ao comparar o desempenho do modelo escolhido com os modelos abordados no estado da arte, pretende-se analisar como o modelo se posiciona em termos de eficácia e robustez. A análise considera diferentes aspectos, como taxas de acerto e riscos de *overfitting*, com o objetivo de avaliar a capacidade de generalização do modelo obtido. Esta comparação é crucial para identificar os pontos fortes e compreender o seu desempenho em diferentes contextos e condições. A Tabela 6.3 faz uma comparação entre os resultados obtidos nesta tese e os resultados dos estudos abordados no Estado da Arte.

Tabela 6.3: Resumo das abordagens existentes

Artigo	Temporadas	Resultados
(Horvat, Job et al. 2023)	2013 a 2018	Taxa de Acerto Média: 66% Taxa de Acerto Máxima: 78%
(Ozkan 2020)	2015/2016 (240 jogos)	Taxa de Acerto: 79,2% Sensibilidade: 72,7% Especificidade: 79,1%
(Zhao, Du e G. Tan 2023)	2012 a 2018	Taxa de Acerto Média: 71,54% Taxa de Acerto Máxima: 73,78%
(Chen et al. 2021)	2018/2019	MAPE: 0.0818
(Cheng et al. 2016)	2007 a 2015	Taxa de Acerto: 74,4%
(Horvat, Hava e Srpak 2020)	2009 a 2018	Taxa de Acerto Média: 60,01% Taxa de Acerto Máxima: 60,82%
(Zheng 2022)	2012 a 2021	Taxa de Acerto: 67,98%
Modelo obtido	2005/2006 a 2021/2022	TAB: 61,7% Taxa de acerto: 64,4% Precisão: 64,5% *Recall*: 82,5% F1: 72,4%

A comparação dos resultados revela que, embora o modelo desenvolvido não alcance as taxas de acerto mais altas entre os modelos analisados, ele apresenta uma abordagem mais robusta e abrangente. A maioria dos modelos analisados apenas se foca numa única métrica de desempenho, como a taxa de acerto ou o MAPE, o que pode oferecer uma visão limitada do desempenho real. Em contraste, o modelo desenvolvido avalia múltiplas métricas de desempenho, permitindo uma análise mais completa da sua capacidade.

Embora alguns estudos relatem taxas de acerto superiores, essas métricas são frequentemente obtidas em configurações que podem limitar a generalização, como conjuntos de dados mais restritos ou configurações de treino que favorecem *overfitting*.

Em suma, apesar de o modelo não alcançar as taxas de acerto mais elevadas em comparação com os outros modelos, a sua abordagem robusta e avaliação abrangente de múltiplas métricas de desempenho garantem uma análise mais completa e confiável, favorecendo uma maior capacidade de generalização e aplicabilidade a diferentes cenários de jogos.

Capítulo 7

Conclusão

O principal objetivo desta tese foi o desenvolvimento de um modelo preditivo capaz de prever se a equipa da casa vencerá um jogo da NBA, com base em variáveis estatísticas e históricas. O modelo alcançou uma taxa de acerto de 64,4% e F1=72,4%, o que indica um desempenho frágil em termos percentuais, mas robusto, considerando a complexidade e a competitividade dos jogos na NBA.

Além do modelo preditivo, esta pesquisa abordou cinco questões complementares de grande relevância para a análise da NBA: como o fator casa afeta o desempenho das equipas, a forma como a dinâmica da liga mudou ao longo do tempo, o impacto da mudança de equipa no desempenho dos jogadores, a evolução das posições no jogo e a relação entre a idade e o desempenho dos atletas. Conclui-se que o fator casa é determinante na NBA, devido ao apoio do público, da familiaridade com o campo e possíveis efeitos psicológicos e viés dos árbitros. Atualmente a NBA apresenta um ritmo de jogo mais acelerado e um maior ênfase em lançamentos de três pontos, refletindo uma evolução significativa nas capacidades ofensivas e defensivas das equipas. As trocas de equipa afetam mais o número de jogos jogados pelos atletas do que o seu desempenho, sugerindo uma adaptação estável. As posições no jogo tornaram-se mais versáteis, com os jogadores a adaptarem-se ao estilo moderno e a desempenhar múltiplas funções. Quanto à relação entre idade e desempenho, observa-se que o auge ocorre entre os 28 e 30 anos, e que após esta fase, ocorre um declínio gradual.

Estes resultados, além de responderem às questões propostas, contribuem para uma compreensão mais aprofundada dos fatores que afetam o desempenho na NBA, tanto a nível de equipas como individual. Reconhecem-se, no entanto, algumas limitações na pesquisa. O modelo preditivo pode ser melhorado com a incorporação de variáveis adicionais e técnicas mais avançadas, como o uso de redes neuronais. Em investigações futuras, sugere-se explorar também fatores qualitativos, como por exemplo, aspetos emocionais e classificação ELO das equipas, que podem impactar o desempenho das equipas e dos jogadores.

Para além disso, sugere-se também, como trabalho futuro, a atualização automática dos ficheiros utilizados pela aplicação *web* para realizar previsões, através de um *script*. Esta melhoria permitirá que a aplicação se mantenha sempre atualizada com os dados mais recentes, sem necessidade de intervenção manual.

Em resumo, esta tese não só oferece uma ferramenta de previsão útil para analisar jogos da NBA, como também esclarece a evolução da liga e dos jogadores ao longo do tempo. As descobertas servem como uma base sólida para análises preditivas mais refinadas e para a tomada de decisões, tanto no ambiente desportivo como em estudos académicos.

Bibliografia

Alexandropoulos, Stamatios-Aggelos et al. (mai. de 2019). «Stacking Strong Ensembles of Classifiers». Em: pp. 545–556. isbn: 978-3-030-19822-0. doi: 10.1007/978-3-030-19823-7_46.

Ali, Maria et al. (2022). «Stacking Classifier with Random Forest functioning as a Meta Classifier for Diabetes Diseases Classification». Em: *Procedia Computer Science* 207. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 26th International Conference KES2022, pp. 3459–3468. issn: 1877-0509. doi: https://doi.org/10.1016/j.procs.2022.09.404. url: https://www.sciencedirect.com/science/article/pii/S1877050922012959.

AS, Diario (2023). *Do NBA stars play in pre-season games?* Acesso em 16 de maio de 2024. url: https://en.as.com/nba/do-nba-stars-play-in-pre-season-games-n/.

Aziz, NorShakirah et al. (out. de 2020). «A Study on Gradient Boosting Algorithms for Development of AI Monitoring and Prediction Systems». Em: pp. 11–16. doi: 10.1109/ICCI51257.2020.9247843.

Basketball, Breakthrough (s.d.). *The Critical Importance of Shooting*. Acesso em 4 de junho de 2024. url: https://www.breakthroughbasketball.com/fundamentals/shooting-importance.html.

Betclic (2021). *24*. Anúncio de televisão. url: https://www.youtube.com/watch?v=No5YflV_VfM.

– (2023). *BEM-VINDA À FAMÍLIA, TICHA PENICHEIRO!* Acesso em 26 de dezembro de 2023. url: https://www.youtube.com/watch?v=RHvz7TJME84.

Bwin (2023). *Como Apostar na NBA*. Acesso em 12 de dezembro de 2023, 2023. url: https://blog.bwin.pt/como-apostar-na-nba/?wm=5386884&tdpeh=_20004519540_____c&utm_source=search_google&utm_campaign=20004519540&utm_content=&utm_medium=&utm_term=&sb=1&gclid=CjwKCAiAp5qsBhAPEiwAP0qeJju5RNYhQEDZaN-_Zgvp9-NgvtMpi5iOVnuWuywbNjNs8OUT7idnTRoCtoUQAvD_Bw.

Cao, C. (2012). «Sports data mining technology used in basketball outcome prediction». Masters Dissertation. Technological University Dublin.

Cervantes, Jair et al. (2020). «A comprehensive survey on support vector machine classification: Applications, challenges and trends». Em: *Neurocomputing* 408, pp. 189–215. issn: 0925-2312. doi: https://doi.org/10.1016/j.neucom.2019.10.118. url: https://www.sciencedirect.com/science/article/pii/S0925231220307153.

Chen, Wei-Jen et al. (2021). «Hybrid Basketball Game Outcome Prediction Model by Integrating Data Mining Methods for the National Basketball Association». Em: *Entropy* 23.4. issn: 1099-4300. doi: 10.3390/e23040477. url: https://www.mdpi.com/1099-4300/23/4/477.

Cheng, Ge et al. (2016). «Predicting the Outcome of NBA Playoffs Based on the Maximum Entropy Principle». Em: *Entropy* 18.12. issn: 1099-4300. doi: 10.3390/e18120450. url: https://www.mdpi.com/1099-4300/18/12/450.

Coach&A.D. (s.d.). *Unleashing the power of the 3-pointers*. Acesso em 4 de junho de 2024. url: https://coachad.com/play/unleashing-the-power-of-a-proven-3-point-shot-attack/.

Deng, Haowen et al. (dez. de 2021). «Ensemble learning for the early prediction of neonatal jaundice with genetic features». Em: *BMC Medical Informatics and Decision Making* 21. doi: 10.1186/s12911-021-01701-9.

Developers, XGBoost (s.d.). *XGBoost Documentation*. https://xgboost.readthedocs.io/en/stable/tutorials/model.html. Acesso em 30 de dezembro de 2023.

DN/Lusa (2023). *Apostas desportivas online movimentam 4.000 milhões de euros em cinco anos*. Acesso em 11 de dezembro de 2023. url: https://www.dn.pt/desporto/apostas-desportivas-online-movimentam-4000-milhoes-de-euros-em-cinco-anos-15836237.html.

ESPN (s.d.). *ESPN*. Acesso em 22 de maio de 2024. url: https://www.espn.com/.

FiveThirtyEight (2014). *The Hidden Value of the NBA Steal*. Acesso em 3 de junho de 2024. url: https://fivethirtyeight.com/features/the-hidden-value-of-the-nba-steal/.

Forebet (s.d.). Acesso em 11 de dezembro de 2023. url: https://www.forebet.com/.

FPB (2019). *TICHA PENICHEIRO NO WOMENS BASKETBALL HALL OF FAME*. Acesso em 26 de dezembro de 2023. url: https://www.fpb.pt/noticia/ticha-penicheiro-no-womens-basketball-hall-of-fame-2/.

Hoop, Home School (2024). *What is a Turnover in Basketball? From Basics to Strategy*. Acesso em 3 de junho de 2024. url: https://homeschoolhoop.com/turnover-in-basketball/.

HoopSocial (2023). *The Role of Coaching and Its Impact on Team Success in the NBA*. Acesso em 21 de maio de 2024. url: https://hoop-social.com/the-role-of-coaching-and-its-impact-on-team-success-in-the-nba/.

Hornyák, Olivér e László Barna Iantovics (2023). «AdaBoost Algorithm Could Lead to Weak Results for Data with Certain Characteristics». Em: *Mathematics* 11.8. issn: 2227-7390. doi: 10.3390/math11081801. url: https://www.mdpi.com/2227-7390/11/8/1801.

Horvat, Tomislav, Ladislav Hava e Dunja Srpak (2020). «The Impact of Selecting a Validation Method in Machine Learning on Predicting Basketball Game Outcomes». Em: *Symmetry* 12.3. issn: 2073-8994. doi: 10.3390/sym12030431. url: https://www.mdpi.com/2073-8994/12/3/431.

Horvat, Tomislav, Josip Job et al. (2023). «A Data-Driven Machine Learning Algorithm for Predicting the Outcomes of NBA Games». Em: *Symmetry* 15.4. issn: 2073-8994. doi: 10.3390/sym15040798. url: https://www.mdpi.com/2073-8994/15/4/798.

Imtiaz Khan, Nafiz et al. (nov. de 2020). «Prediction of Cesarean Childbirth using Ensemble Machine Learning Methods». Em: doi: 10.1145/3428757.3429138.

Kaggle (s.d.). *Kaggle: Your Machine Learning and Data Science Community*. Acesso em 15 de maio de 2024. url: https://www.kaggle.com/.

Khatun, Rabea et al. (2023). «Cancer Classification Utilizing Voting Classifier with Ensemble Feature Selection Method and Transcriptomic Data». Em: *Genes* 14.9. issn: 2073-4425. doi: 10.3390/genes14091802. url: https://www.mdpi.com/2073-4425/14/9/1802.

Khushaktov, Mr Farkhod (2023). *Introduction to Random Forest Classification by Example*. Acesso em 31 de dezembro de 2023. url: https://medium.com/@mrmaster907/introduction-random-forest-classification-by-example-6983d95c7b91.

Koutsouridis, Christos et al. (dez. de 2020). «Original Article Effect of offensive rebound on the game outcome during the 2019 Basketball World Cup». Em: *Journal of Physical Education and Sport* 20, pp. 3651–3659. doi: 10.7752/jpes.2020.06492.

Lauga, Nathan (s.d.). *NBA games data*. Acesso em 15 de maio de 2024. url: https://www.kaggle.com/datasets/nathanlauga/nba-games.

learn, scikit (s.d.). *BaggingClassifier*. Acesso em 16 de junho de 2024. url: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html.

LightGBM (s.d.). *LightGBM*. Acesso em 7 de julho de 2024. url: https://lightgbm.readthedocs.io/en/latest/Features.html.

Lusa (2016). *Placard já conquistou 920 mil apostadores e distribuiu 200 milhões em prémios*. Acesso em 11 de dezembro de 2023. url: https://www.dn.pt/sociedade/placard-ja-conquistou-920-mil-apostadores-e-distribuiu-200-milhoes-em-premios-5378654.html.

– (2021). *Neemias Queta. O primeiro português a jogar na NBA*. Acesso em 26 de dezembro de 2023. url: https://www.dn.pt/desporto/neemias-queta-tornou-se-o-primeiro-portugues-a-jogar-na-nba-14421592.html.

Medium (2023a). *Understanding the AdaBoost Algorithm*. Acesso em 16 de junho de 2024. url: https://medium.com/@datasciencewizards/understanding-the-adaboost-algorithm-2e9344d83d9b.

– (2023b). *What is Hard and Soft Voting in Machine Learning?* Acesso em 15 de junho de 2024. url: https://ilyasbinsalih.medium.com/what-is-hard-and-soft-voting-in-machine-learning-2652676b6a32.

– (s.d.[a]). *Bagging: Boosting Model Performance through Ensemble Learning*. Acesso em 16 de junho de 2024. url: https://medium.com/@tejaswaroop2310/bagging-boosting-model-performance-through-ensemble-learning-2fbe8ebbb494.

– (s.d.[b]). *Stacking to Improve Model Performance: A Comprehensive Guide on Ensemble Learning in Python*. Acesso em 15 de junho de 2024. url: https://medium.com/@brijesh_soni/stacking-to-improve-model-performance-a-comprehensive-guide-on-ensemble-learning-in-python-9ed53c93ce28.

Mokray, William George, Larry W. Donald e Robert G. Logan (2023). *Basketball*. Encyclopedia Britannica, 26 Dec. 2023. url: https://www.britannica.com/sports/basketball (acedido em 26/12/2023).

Natekin, Alexey e Alois Knoll (dez. de 2013). «Gradient Boosting Machines, A Tutorial». Em: *Frontiers in neurorobotics* 7, p. 21. doi: 10.3389/fnbot.2013.00021.

NBA (2023). *2023-24 NBA Season Official Playing Rules*. Acesso em 21 de dezembro de 2023. url: https://official.nba.com/wp-content/uploads/sites/4/2023/10/2023-24-NBA-Season-Official-Playing-Rules.pdf.

– (s.d.). *NBA*. Acesso em 16 de maio de 2024. url: https://www.nba.com/.

NBA-API (s.d.). *NBA-API*. Acesso em 25 de julho de 2024. url: https://nba-apidocumentation.knowledgeowl.com/help.

News, Deseret (2023). *Is home-court advantage real? And if it is what explains it?* Acesso em 23 de julho de 2024. url: https://www.deseret.com/sports/2023/10/6/23787371/is-home-court-advantage-real/.

Ngo, Giang, Rodney Beard e Rohitash Chandra (out. de 2022). «Evolutionary bagging for ensemble learning». Em: *Neurocomputing* 510, pp. 1–14. issn: 0925-2312. doi: 10.1016/j.neucom.2022.08.055. url: http://dx.doi.org/10.1016/j.neucom.2022.08.055.

obviously.ai (2022). *Data Cleaning: The Most Important Step in Machine Learning*. Acesso em 16 de maio de 2024. url: https://www.obviously.ai/post/data-cleaning-in-machine-learning.

Ozkan, Ilker Ali (ago. de 2020). «A Novel Basketball Result Prediction Model Using a Concurrent Neuro-Fuzzy System». Em: *Applied Artificial Intelligence* 34, pp. 1–17. doi: 10.1080/08839514.2020.1804229.

Ramalingam, Karthikeyan et al. (mar. de 2024). «BMC Oral Health Light gradient boosting-based prediction of quality of life among oral cancer-treated patients». Em: *BMC Oral Health* 24. doi: 10.1186/s12903-024-04050-x.

RealGM (s.d.). *RealGM*. Acesso em 16 de maio de 2024. url: https://basketball.realgm.com/nba/preseason/schedule.

Reference, BasketBall (s.d.[a]). *BasketBall Reference*. Acesso em 16 de maio de 2024. url: https://www.basketball-reference.com/.

– (s.d.[b]). *Calculating PER*. Acesso em 28 de agosto de 2024. url: https://www.basketball-reference.com/about/per.html.

Report, Bleacher (2013a). *How Important Is Home-Court Advantage in the NBA?* Acesso em 23 de julho de 2024. url: https://bleacherreport.com/articles/1520496-how-important-is-home-court-advantage-in-the-nba.

– (2013b). *Why Ownership Makes All the Difference in the NBA*. Acesso em 21 de maio de 2024. url: https://bleacherreport.com/articles/1865757-why-ownership-makes-all-the-difference-in-the-nba.

Sá, José et al. (mar. de 2016). «Lightning Forecast Using Data Mining Techniques On Hourly Evolution Of The Convective Available Potential Energy». Em: pp. 1–5. doi: 10.21528/CBIC2011-27.1.

Shariah, Mohammad et al. (jun. de 2022). «SOFT VOTING MACHINE LEARNING CLASSIFICATION MODEL TO PREDICT AND EXPOSE LIVER DISORDER FOR HUMAN PATIENTS». Em: *Journal of Theoretical and Applied Information Technology* 100, pp. 4554–4564.

Sherazi, Syed Waseem Abbas, Jang-Whan Bae e Jong Yun Lee (jun. de 2021). «A soft voting ensemble classifier for early prediction and diagnosis of occurrences of major adverse cardiovascular events for STEMI and NSTEMI during 2-year follow-up in patients with acute coronary syndrome». Em: *PLOS ONE* 16.6, pp. 1–20. doi: 10.1371/journal.pone.0249338. url: https://doi.org/10.1371/journal.pone.0249338.

Sohrabi, C. et al. (2021). «PRISMA 2020 statement: Whats new and the importance of reporting guidelines». Em: *International Journal Of Surgery* 88, p. 105918. doi: 10.1016/j.ijsu.2021.105918.

Sportskeeda (2023). *The Role of Coaching in the NBA: Analyzing the impact of coaches on team performance*. Acesso em 21 de maio de 2024. url: https://www.sportskeeda.com/basketball/the-role-coaching-nba-analyzing-impact-head-coaches-team-performance.

SportyTrader (s.d.). Acesso em 11 de dezembro de 2023, 2023. url: https://www.sportytrader.com/en/.

SRIJ (2023). *Entidades licenciadas*. Acesso em 11 de dezembro de 2023. url: https://www.srij.turismodeportugal.pt/pt/jogos-e-apostas-online/entidades-licenciadas.

StatsMuse (s.d.). *StatsMuse*. Acesso em 20 de julho de 2024. url: https://www.statmuse.com/.

Streamlit (2024). *Streamlit*. Acesso em 25 de julho de 2024. url: https://streamlit.io/.

Student, Hoop (s.d.). *Steal in Basketball: Basic Information Explained*. Acesso em 3 de junho de 2024. url: https://hoopstudent.com/basketball-steal-basics/#:~:text=The%20primary%20benefit%20of%20stealing,than%20average%20scorers%2Fperimeter%20shooters..

Tan, Pang-Ning, Michael Steinbach, Anuj Karpatne et al. (2018). *Introduction to Data Mining (2nd Edition)*. 2nd. Pearson. isbn: 0133128903.

Tan, Pang-Ning, Michael Steinbach e Vipin Kumar (2013). *Introduction to Data Mining: Pearson New International Edition*. 1ª ed. Pearson.

Ting, Wen-Chien et al. (2020). «Developing a Novel Machine Learning-Based Classification Scheme for Predicting SPCs in Colorectal Cancer Survivors». Em: *Applied Sciences* 10.4. issn: 2076-3417. doi: 10.3390/app10041355. url: https://www.mdpi.com/2076-3417/10/4/1355.

Vidhya, Analytics (s.d.). *Guide on Support Vector Machine (SVM) Algorithm*. Acesso em 14 de junho de 2024. url: https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/.

Vitibet (s.d.). Acesso em 11 de dezembro de 2023. url: https://www.vitibet.com/.

Wirth, R. e Jochen Hipp (jan. de 2000). «CRISP-DM: Towards a standard process model for data mining». Em: *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*.

Wong, Glenn e Chris Deubert (dez. de 2010). «National Basketball Association General Managers: An Analysis of the Responsibilities, Qualifications and Characteristics». Em: *SSRN Electronic Journal*. doi: 10.2139/ssrn.2282583.

Zhao, Kai, Chunjie Du e Guangxin Tan (2023). «Enhancing Basketball Game Outcome Prediction through Fused Graph Convolutional Networks and Random Forest Algorithm». Em: *Entropy* 25.5. Acesso em 11 de dezembro de 2023. issn: 1099-4300. doi: 10.3390/e25050765.

Zheng, Xi (2022). «NBA Winner Prediction: A Hybrid Framework Incorporating Internal and External Factors». Em: *Proceedings of the 4th International Conference on Big Data Engineering*. BDE '22. Beijing, China: Association for Computing Machinery, pp. 71–80. isbn: 9781450395632. doi: 10.1145/3538950.3538960. url: https://doi.org/10.1145/3538950.3538960.

Apêndice A

Constituição dos conjuntos de dados

Tabela A.1: Constituição do ficheiro *games*

Coluna	Informação
GAME_DATE_EST	Data em que o jogo ocorreu
GAME_ID	Identificador do jogo
GAME_STATUS_TEXT	Estado do jogo
HOME_TEAM_ID e *TEAM_ID_home*	Identificador da equipa da casa
VISITOR_TEAM_ID e *TEAM_ID_away*	Identificador da equipa visitante
SEASON	Identificador da temporada em que o jogo ocorreu
PTS_home e *PTS_away*	Pontos marcados no jogo pela equipa da casa e pela equipa visitante, respetivamente
FG_PCT_home e *FG_PCT_away*	Percentagem de acerto de lançamentos da equipa da casa e visitante, respetivamente
FT_PCT_home e *FT_PCT_away*	Percentagem de acerto de lançamentos livres da equipa da casa e visitante, respetivamente
FG3_PCT_home e *FG3_PCT_away*	Percentagem de acerto de lançamentos de três pontos da equipa da casa e visitante, respetivamente
AST_home e *AST_away*	Assistências feitas no jogo pela equipa da casa e pela equipa visitante, respetivamente
REB_home e *REB_away*	Ressaltos feitos no jogo pela equipa da casa e pela equipa visitante, respetivamente
HOME_TEAM_WINS	Se a equipa da casa ganhou ou não. Variável alvo

Tabela A.2: Constituição do ficheiro *players*

Coluna	Informação
PLAYER_NAME	Nome do jogador
TEAM_ID	Identificador da equipa onde o jogador joga
PLAYER_ID	Identificador do jogador
SEASON	Temporada à qual o registo pertence

Tabela A.3: Constituição do ficheiro *ranking*

Coluna	**Informação**
LEAGUE_ID	Identificador da liga
TEAM_ID	Identificador da equipa
SEASON_ID	Identificador da temporada à qual o registo pertence
STANDINGSDATE	Data à qual o registo pertence
CONFERENCE	Conferência em que a equipa joga
TEAM	Nome da equipa
G, *W*, *L*	Jogos realizados, vitórias e derrotas da equipa, respetivamente
W_PCT	Percentagem de jogos em que a equipa saiu vitoriosa
HOME_RECORD	Registo da equipa a jogar em casa, no formato Vitórias-Derrotas
ROAD_RECORD	Registo da equipa a jogar como visitante, no formato Vitórias-Derrotas
RETURNTOPLAY	Indica se foi uma data de retorno à competição, após paragem

Tabela A.4: Constituição do ficheiro *teams*

Coluna	**Informação**
LEAGUE_ID	Identificador da liga
TEAM_ID	Identificador da equipa
MIN_YEAR e *YEARFOUNDED*	Ano de fundação da equipa
MAX_YEAR	Ano até ao qual a equipa participou
ABBREVIATION e *NICKNAME*	Abreviatura e alcunha da equipa, respetivamente
CITY	Cidade sede da equipa
ARENA e *ARENACAPACITY*	Arena onde a equipa joga e respetiva capacidade
OWNER, *GENERALMANAGER* e *HEADCOACH*	Dono, GM e treinador da equipa, respetivamente
DLEAGUEAFFILIATION	Nome da equipa B

Tabela A.5: Constituição do ficheiro *game_details*

Coluna	**Informação**
GAME_ID	Identificador do jogo
TEAM_ID	Identificador da equipa
TEAM_ABBREVIATION	Abreviatura do nome da equipa
TEAM_CITY	Cidade sede da equipa
PLAYER_ID	Identificador do jogador
PLAYER_NAME	Nome completo do jogador
NICKNAME	Nome que o jogador usa no equipamento
START_POSITION	Posição em que o jogador jogou. Esta coluna apenas está preenchida caso o jogador tenha sido titular
COMMENT	Motivo da ausência, se aplicável
MIN	Minutos jogados
FGA, *FGM* e *FG_PCT*	Tentativas de lançamento, lançamentos bem sucedidos e percentagem de lançamentos bem sucedidos, respetivamente
FG3A, *FG3M* e *FG3_PCT*	Tentativas de lançamento de 3 pontos, lançamentos de 3 pontos bem sucedidos e percentagem de lançamentos de 3 pontos bem sucedidos, respetivamente
FTA, *FTM* e *FT_PCT*	Lançamentos livres feitos, lançamentos livres bem sucedidos e percentagem de lançamentos livres bem sucedidos, respetivamente
REB, *OREB* e *DREB*	Total de ressaltos, ressaltos ofensivos e ressaltos defensivos, respetivamente
AST	Assistências realizadas
STL	Roubos de bola realizados
BLK	Bloqueios realizados
TO	Perdas da posse de bola
PF	Faltas cometidas
PTS	Pontos marcados
PLUS_MINUS	Diferença entre os pontos marcados e sofridos durante o tempo em que o jogador esteve em campo.

Tabela A.6: Constituição do ficheiro *games* após limpeza e tratamento

Informação	**Descrição**
Data do jogo	Data em que o jogo ocorreu
ID	Identificador do jogo
Equipa da casa e equipa visitante	Identificadores das equipas que participaram no jogo
Temporada	Identificador da temporada em que o jogo ocorreu
Estatísticas das equipas	Média nos últimos 4 jogos das estatísticas (pontos marcados e sofridos, assistências, percentagens de lançamento, ressaltos) de ambas as equipas
Estatísticas como equipa da casa	Média das estatísticas nos últimos 4 jogos realizados em casa (pontos marcados e sofridos, assistências, percentagens de lançamento, ressaltos) pela equipa da casa
Estatísticas como equipa visitante	Média das estatísticas nos últimos 4 jogos realizados fora de casa (pontos marcados e sofridos, assistências, percentagens de lançamento, ressaltos) pela equipa visitante
Equipa Vencedora	Se a equipa da casa ganhou ou não. Variável alvo

Tabela A.7: Constituição do ficheiro *teams* após limpeza e tratamento

Informação	**Descrição**
Equipa	Identificador da equipa
Fundação	Ano de fundação da equipa

Tabela A.8: Constituição do ficheiro *ranking* após limpeza e tratamento

Informação	**Descrição**
Equipa	Identificador da equipa
Temporada	Identificador da temporada à qual o registo pertence
Data	Data à qual o registo pertence
Conferência	Conferência em que a equipa joga
Registos da equipas	Jogos realizados, vitórias, percentagem de jogos ganhos e derrotas de ambas as equipas
Registo da equipa em casa	Jogos realizados, vitórias e derrotas da equipa a jogar em casa
Registo da equipa fora de casa	Jogos realizados, vitórias e derrotas da equipa a jogar como visitante

Tabela A.9: Constituição do ficheiro *game_details* após limpeza e tratamento

Informação	**Descrição**
Jogo	Identificador do jogo
Equipa	Identificador da equipa
Jogador	Identificador do jogador
Posição	Posição em que o jogador jogou, caso tenha sido titular. Se começou no banco tem o valor *DIDN'T START*
Minutos	Minutos jogados pelo jogador
Estatísticas do jogador	Pontos, assistências, ressaltos (ofensivos, defensivos e total), roubos de bola, bloqueios, perdas de posse, faltas cometidas e diferença entre os pontos marcados e sofridos durante o tempo em que o jogador esteve em campo
Estatísticas de lançamento do jogador	Tentativas de lançamento, lançamentos bem sucedidos e percentagem de lançamentos bem sucedidos; tentativas de lançamento livres, lançamentos livres bem sucedidos e percentagem de lançamentos livres bem sucedidos; tentativas de lançamento de 3 pontos, lançamentos de 3 pontos bem sucedidos e percentagem de lançamentos de 3 pontos bem sucedidos

Tabela A.10: Constituição do conjunto de dados após junção da informação das equipas

Informação	**Descrição**
Data do jogo	Data em que o jogo ocorreu
ID	Identificador do jogo
Equipa da casa e equipa visitante	Identificadores das equipas que participaram no jogo
Temporada	Identificador da temporada em que o jogo ocorreu
Estatísticas das equipas	Média nos últimos 4 jogos das estatísticas (pontos marcados e sofridos, assistências, percentagens de lançamento, ressaltos) de ambas as equipas
Estatísticas como equipa da casa	Média das estatísticas nos últimos 4 jogos realizados em casa (pontos marcados e sofridos, assistências, percentagens de lançamento, ressaltos) pela equipa da casa
Estatísticas como equipa visitante	Média das estatísticas nos últimos 4 jogos realizados fora de casa (pontos marcados e sofridos, assistências, percentagens de lançamento, ressaltos) pela equipa visitante
Fundação	Ano de fundação das equipas que participaram no jogo
Equipa Vencedora	Se a equipa da casa ganhou ou não. Variável alvo

Tabela A.11: Constituição do conjunto de dados após junção da informação dos *rankings*

Informação	**Descrição**
Data do jogo	Data em que o jogo ocorreu
ID	Identificador do jogo
Equipa da casa e equipa visitante	Identificadores das equipas que participaram no jogo
Temporada	Identificador da temporada em que o jogo ocorreu
Estatísticas das equipas	Média nos últimos 4 jogos das estatísticas (pontos marcados e sofridos, assistências, percentagens de lançamento, ressaltos) de ambas as equipas
Estatísticas como equipa da casa	Média das estatísticas nos últimos 4 jogos realizados em casa (pontos marcados e sofridos, assistências, percentagens de lançamento, ressaltos) pela equipa da casa
Estatísticas como equipa visitante	Média das estatísticas nos últimos 4 jogos realizados fora de casa (pontos marcados e sofridos, assistências, percentagens de lançamento, ressaltos) pela equipa visitante
Fundação	Ano de fundação das equipas que participaram no jogo
Conferências	Conferências às quais as equipas que participam no jogo pertencem
Registo das equipas	Jogos realizados, vitórias, percentagem de jogos ganhos e derrotas das equipas que participam no jogo
Registo da equipa da casa em casa	Jogos realizados, vitórias e derrotas da equipa da casa a jogar em casa
Registo da equipa visitante fora de casa	Jogos realizados, vitórias e derrotas da equipa visitante a jogar fora de casa
Sequências de vitórias/derrotas	Sequência de vitórias/derrotas de ambas as equipas, sequência de vitórias/derrotas da equipa da casa em jogos em casa e sequência de vitórias/derrotas da equipa visitante em jogos fora de casa
Equipa Vencedora	Se a equipa da casa ganhou ou não. Variável alvo

Tabela A.12: Constituição do conjunto de dados após implementação do cansaço

Informação	**Descrição**
Data do jogo	Data em que o jogo ocorreu
ID	Identificador do jogo
Equipa da casa e equipa visitante	Identificadores das equipas que participaram no jogo
Temporada	Identificador da temporada em que o jogo ocorreu
Estatísticas das equipas	Média nos últimos 4 jogos das estatísticas (pontos marcados e sofridos, assistências, percentagens de lançamento, ressaltos) de ambas as equipas
Estatísticas como equipa da casa	Média das estatísticas nos últimos 4 jogos realizados em casa (pontos marcados e sofridos, assistências, percentagens de lançamento, ressaltos) pela equipa da casa
Estatísticas como equipa visitante	Média das estatísticas nos últimos 4 jogos realizados fora de casa (pontos marcados e sofridos, assistências, percentagens de lançamento, ressaltos) pela equipa visitante
Fundação	Ano de fundação das equipas que participaram no jogo
Conferências	Conferências às quais as equipas que participam no jogo pertencem
Registo das equipas	Jogos realizados, vitórias, percentagem de jogos ganhos e derrotas das equipas que participam no jogo
Registo da equipa da casa em casa	Jogos realizados, vitórias e derrotas da equipa da casa a jogar em casa
Registo da equipa visitante fora de casa	Jogos realizados, vitórias e derrotas da equipa visitante a jogar fora de casa
Sequências de vitórias/derrotas	Sequência de vitórias/derrotas de ambas as equipas, sequência de vitórias/derrotas da equipa da casa em jogos em casa e sequência de vitórias/derrotas da equipa visitante em jogos fora de casa
Cansaço das equipas	Para ambas as equipas: quantidade de jogos nos últimos 7,10 e 14 dias; quantidade de jogos fora de casa nos últimos 7,10 e 14 dias; sequência de jogos fora de casa; média de jogadores utilizados nos últimos 4 jogos; média de minutos por jogador nos últimos 4 jogos; média de minutos jogados pelos titulares nos últimos 4 jogos
Equipa Vencedora	Se a equipa da casa ganhou ou não. Variável alvo

Tabela A.13: Constituição do conjunto de dados após implementação de mais estatísticas

Informação	**Descrição**
Data do jogo	Data em que o jogo ocorreu
ID	Identificador do jogo
Equipa da casa e equipa visitante	Identificadores das equipas que participaram no jogo
Temporada	Identificador da temporada em que o jogo ocorreu
Estatísticas das equipas	Média nos últimos 4 jogos das estatísticas (pontos marcados e sofridos, assistências, lançamentos (tentativas, bem sucedidos e percentagem de sucesso), lançamentos livres (tentativas, bem sucedidos e percentagem de sucesso), lançamentos de 3 pontos (tentativas, bem sucedidos e percentagem de sucesso), ressaltos (ofensivos, defensivos, total), faltas, perdas de bola, roubos de bola, bloqueios) de ambas as equipas
Estatísticas como equipa da casa	Média das estatísticas nos últimos 4 jogos realizados em casa (pontos marcados e sofridos, assistências, lançamentos (tentativas, bem sucedidos e percentagem de sucesso), lançamentos livres (tentativas, bem sucedidos e percentagem de sucesso), lançamentos de 3 pontos (tentativas, bem sucedidos e percentagem de sucesso), ressaltos (ofensivos, defensivos, total), faltas, perdas de bola, roubos de bola, bloqueios) pela equipa da casa
Estatísticas como equipa visitante	Média das estatísticas nos últimos 4 jogos realizados fora de casa (pontos marcados e sofridos, assistências, lançamentos (tentativas, bem sucedidos e percentagem de sucesso), lançamentos livres (tentativas, bem sucedidos e percentagem de sucesso), lançamentos de 3 pontos (tentativas, bem sucedidos e percentagem de sucesso), ressaltos (ofensivos, defensivos, total), faltas, perdas de bola, roubos de bola, bloqueios) pela equipa visitante
Fundação	Ano de fundação das equipas que participaram no jogo
Conferências	Conferências às quais as equipas que participam no jogo pertencem
Registo das equipas	Jogos realizados, vitórias, percentagem de jogos ganhos e derrotas das equipas que participam no jogo
Registo da equipa da casa em casa	Jogos realizados, vitórias e derrotas da equipa da casa a jogar em casa
Registo da equipa visitante fora de casa	Jogos realizados, vitórias e derrotas da equipa visitante a jogar fora de casa
Sequências de vitórias/derrotas	Sequência de vitórias/derrotas de ambas as equipas, sequência de vitórias/derrotas da equipa da casa em jogos em casa e sequência de vitórias/derrotas da equipa visitante em jogos fora de casa
Cansaço das equipas	Para ambas as equipas: quantidade de jogos nos últimos 7,10 e 14 dias; quantidade de jogos fora de casa nos últimos 7,10 e 14 dias; sequência de jogos fora de casa; média de jogadores utilizados nos últimos 4 jogos; média de minutos por jogador nos últimos 4 jogos; média de minutos jogados pelos titulares nos últimos 4 jogos
Equipa Vencedora	Se a equipa da casa ganhou ou não. Variável alvo

Tabela A.14: Variáveis utilizadas pelo modelo preditivo final

Informação	**Descrição**
Estatísticas da equipa da casa	Média nos últimos 5 jogos de pontos marcados e sofridos, assistências, percentagem de lançamentos bem sucedidos, percentagem de lançamentos livres bem sucedidos, percentagem de lançamentos de 3 pontos bem sucedidos e ressaltos. Média nos últimos 5 jogos em casa de pontos marcados e sofridos, assistências e ressaltos
Estatísticas da equipa visitante	Média nos últimos 5 jogos de pontos marcados e sofridos, assistências, percentagem de lançamentos bem sucedidos, percentagem de lançamentos livres bem sucedidos, percentagem de lançamentos de 3 pontos bem sucedidos e ressaltos. Média nos últimos 5 jogos fora de casa de pontos marcados e sofridos, assistências, percentagem de lançamentos de 3 pontos bem sucedidos e ressaltos
Fundação	Ano de fundação de ambas as equipas
Registo da equipa da casa	Jogos realizados, vitórias, percentagem de jogos ganhos e número de derrotas. Derrotas em casa, percentagem de jogos ganhos em casa. Sequência de vitórias/derrotas em jogos em casa
Registo da equipa visitante	Vitórias, derrotas, percentagem de jogos ganhos e percentagem de jogos ganhos fora de casa.
Cansaço das equipas	Para ambas as equipas: média de minutos por jogador nos últimos 4 jogos e média de minutos jogados pelos titulares nos últimos 4 jogos