UNICAMP

UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Matemática, Estatística e Computação Científica

AMAURÍ HIDEO SAIJO

Um estudo sobre modelos preditivos para resultados de jogos da NBA

Campinas

2023

Amaurí Hideo Saijo

Um estudo sobre modelos preditivos para resultados de jogos da NBA

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Matemática Aplicada e Computacional.

Orientador: Cristiano Torezzan

Este trabalho corresponde à versão final da dissertação defendida pelo aluno Amaurí Hideo Saijo e orientada pelo Prof. Dr. Cristiano Torezzan.

Campinas

2023

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

Saijo, Amaurí Hideo, 1992-
Sa21e Um estudo sobre modelos preditivos para resultados de jogos da NBA / Amaurí Hideo Saijo. – Campinas, SP : [s.n.], 2023.

Orientador: Cristiano Torezzan.
Dissertação (mestrado profissional) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Basquetebol. 2. Python (Linguagem de programação de computador). 3. Aprendizado de máquina. 4. Processo decisório por critério múltiplo. 5. Desempenho esportivo. I. Torezzan, Cristiano, 1976-. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Informações Complementares

Título em outro idioma: A study on predictive models for NBA game results
Palavras-chave em inglês:
Basketball
Python (Computer program language)
Machine learning
Multiple criteria decision making
Sports performance
Área de concentração: Matemática Aplicada e Computacional
Titulação: Mestre em Matemática Aplicada e Computacional
Banca examinadora:
Cristiano Torezzan [Orientador]
João Eloir Strapasson
Luciano Allegretti Mercadante
Data de defesa: 19-06-2023
Programa de Pós-Graduação: Matemática Aplicada e Computacional

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0009-0000-8667-7423
- Currículo Lattes do autor: https://lattes.cnpq.br/9783156536891540

Dissertação de Mestrado Profissional defendida em 19 de junho de 2023 e aprovada pela banca examinadora composta pelos Profs. Drs.

Prof(a). Dr(a). CRISTIANO TOREZZAN

Prof(a). Dr(a). JOÃO ELOIR STRAPASSON

Prof(a). Dr(a). LUCIANO ALLEGRETTI MERCADANTE

A Ata da Defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós-Graduação do Instituto de Matemática, Estatística e Computação Científica.

Resumo

A previsão de resultados em esportes é um tema que desafia especialistas, fãs e pesquisadores de diversas áreas. A atenção por esse problema aumentou ainda mais a partir dos recentes avanços na área de aprendizado de máquina. Nesse contexto, este trabalho de dissertação investiga a eficácia de modelos orientados a dados para a previsão de resultados de jogos de basquetebol da NBA. Foram propostos e analisados modelos que visam prever o número de pontos de cada time em um dado confronto, a partir de estatísticas prévias dos times. O trabalhou utilizou dados reais das temporadas 2018-2019, 2019-2020, 2021-2021 e 2021-2022 da NBA, incluindo estatísticas de ataque e defesa das equipes, além de fatores extras como mando de casa e o histórico de vitórias de cada time. Os modelos propostos foram programados em linguagem Python e os resultados foram comparados com trabalhos de referência da área. Uma particular atenção foi dedicada para a comparação entre a acurácia dos modelos quando se utiliza, ou não, dados da partida em andamento para realizar a previsão. Em todos os testes realizados, identificamos uma queda significativa no desempenho dos algoritmos quando não se utiliza dados da partida em andamento, fato que é muito relevante em vários contextos. Além disso, os algoritmos propostos neste trabalho obtiveram desempenho competitivo quando comparados à modelos da literatura.

Palavras-chaves: Previsão de pontos na NBA; regressão; modelos preditivos.

Abstract

The prediction of results in sports is a topic that challenges specialists, fans and researchers from different areas. Attention to this problem has increased even more since recent advances in machine learning. In this context, this dissertation investigates the effectiveness of data-driven models for predicting the results of NBA basketball games. Models that aim to predict the number of points for each team in a given match were proposed and analyzed, based on previous team statistics. The work used real data from the 2018-2019, 2019-2020, 2021-2021 and 2021-2022 NBA seasons, including team attack and defense statistics, as well as extra factors such as home field and each team's winning history. The proposed models were programmed in Python language and the results were compared with reference works in the area. Particular attention was dedicated to the comparison between the accuracy of the models when using, or not, data from the match in progress to perform the forecast. In all the tests carried out, we identified a significant drop in the performance of the algorithms when data from the game in progress is not used, a fact that is very relevant in several contexts. Furthermore, the algorithms proposed in this work obtained competitive performance when compared to literature models.

Keywords: Game points predictions in the NBA; regression; predictive models.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Nos últimos anos, a importância das estatísticas no mundo do esporte tem crescido significativamente. A análise detalhada dos dados se tornou fundamental para entender e aprimorar a performance dos atletas de alto rendimento. Constantemente, novas pesquisas e estudos são conduzidos em várias modalidades esportivas, com o intuito de analisar tanto o desempenho individual dos jogadores quanto o desempenho coletivo das equipes. Assim como é possível ver em Mercadante [2021], onde foi feito um estudo geral do basquete do *jogo livre ao alto rendimento* ou em Oliver [2004], onde se estudou algumas estatísticas, tais como arremessos de 2 pontos e 3 pontos, em conjunto do trabalho em equipe, analisando como eles impactam no resultado final da partida.

O presente estudo pretende focar as análises no basquete, mais especificamente nos jogos da NBA. As principais análises aqui realizadas são para tentar prever os resultados de um jogo antes mesmo dele acontecer, utilizando apenas dados históricos das equipes e jogadores. Apesar de existirem diversos estudos relacionados, tal como em [Huang and Lin, 2020], [Jain and Kaur, 2017] e em [Pai et al., 2017], a grande maioria fica restrita em testar os métodos preditivos com apenas os dados históricos dos jogos e não uma simulação do jogo a ser previsto. A principal motivação desse estudo surge desse contra-senso: qual a razão de prever o resultado de um jogo se eu tenho todas as estatísticas desde jogo? Por isso, nos aprofundamos em analisar a eficiência de tentar prever um jogo sem ter as estatísticas daquele jogo, apenas com as estatísticas históricas. Além disso, analisamos também o impacto na acurácia dos modelos da inclusão ou não das estatísticas dos jogos que estão sendo previstos.

1.1 O que é a NBA?

A NBA (*National Basketball Association*) é a principal liga profissional de basquete dos Estados Unidos da América. Fundada em 1946, a NBA se tornou uma das ligas esportivas mais populares e influentes do mundo. Ela é composta por 30 equipes, sendo 29 dos Estados Unidos e 1 do Canadá, e atrai alguns dos melhores jogadores de basquete do planeta.

A temporada regular da NBA ocorre anualmente, geralmente de outubro a abril, com cada equipe disputando 82 jogos. Após a temporada regular, os playoffs são realizados, envolvendo as equipes com melhor desempenho, para determinar o campeão da NBA. Os playoffs são conhecidos por sua intensidade e emocionantes séries de confrontos

eliminatórios.

A NBA é conhecida por seu alto nível de habilidade, jogadas espetaculares, jogadores carismáticos e uma cultura única que envolve não apenas o esporte, mas também moda, música e entretenimento.

Em linhas gerais, o torneio é composto por 30 times que são divididos em 2 conferências, a do Leste e a do Oeste. Na temporada regular, cada equipe joga 82 partidas (contra equipes da sua conferência e da outra) e os 6 melhores de cada conferência se classificam direto para os *playoffs*. As equipes entre a sétima e a décima posição de cada conferência jogam um mini torneio por conferência chamado de *play-in*. Os 2 melhores de cada conferência do *play-in* se juntam aos que já estavam previamente classificados.

As equipes de cada conferência se enfrentam em série de jogos melhor de 7 partidas no estilo 'mata-mata' e o campeão de cada conferência se enfrentam na grande final da NBA. A Figura 1.1 ajuda a ilustrar melhor como se dão as partidas da fase eliminatória da NBA.

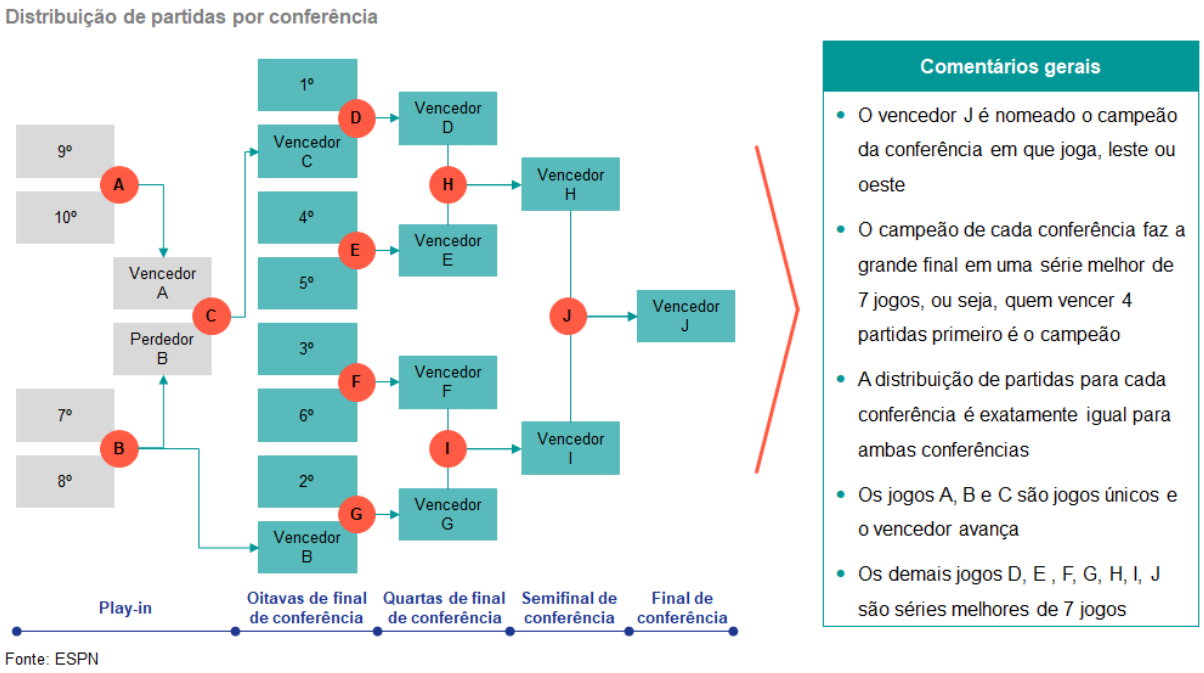

Figura 1.1 – Distribuição de partidas da NBA na fase eliminatória

1.2 Uso das estatísticas na NBA

A estatística desempenha um papel fundamental no basquete, e a NBA é um exemplo notável disso. Por meio da análise estatística, é possível obter *insights* valiosos sobre o desempenho dos jogadores, o impacto tático das equipes e as tendências do jogo. Estatísticas como pontos marcados, rebotes, assistências, porcentagem de arremes-

sos convertidos e eficiência geral são utilizadas para avaliar a contribuição individual dos jogadores. Além disso, estatísticas dos jogos, como percentual de acerto, eficiência de arremessos e rating ofensivo/defensivo, fornecem uma compreensão mais profunda do valor e impacto de um jogador em diferentes aspectos do jogo. Os times da NBA baseiam suas estratégias, escalações e trocas em análises estatísticas detalhadas, permitindo uma abordagem mais fundamentada e objetiva na tomada de decisões dentro e fora da quadra.

Atualmente, todos os times da NBA possuem equipes focadas nas análises das estatísticas dos jogos. Segundo a [NBA Stuffer, 2022], algumas equipes podem ter até 10 pessoas trabalhando com dados estatísticos e análise de desempenho. Em jogos como os de basquete, onde há bastante repetibilidade de eventos (passes, arremessos, assistências etc), é evidente que a análise desses dados é fundamental.

Só na temporada de 2020-2021, foram 326 jogos que terminaram com no máximo 3 pontos de diferença, isto é, apenas uma posse de bola.

Nesse contexto, este trabalho de dissertação investiga a eficácia de modelos orientados a dados para a previsão de resultados de jogos de basquetebol da NBA. Foram propostos e analisados modelos que visam prever o número de pontos de cada time em um dado confronto, a partir de estatísticas prévias dos times. O trabalhou utilizou dados reais das temporadas 2018-2019, 2019-2020, 2021-2021 e 2021-2022 da NBA, incluindo estatísticas de ataque e defesa das equipes, além de fatores extras como mando de casa e o histórico de vitórias de cada time. Uma aplicação muito útil desse estudo pode ser a análise prévia da comissão técnica da equipe e tentar mudar jogadores e esquemas táticos ou focar os treinos para melhorar pontos vulneráveis que aumentem as chances de vitória da equipe. Por exemplo, se historicamente a equipe tiver um baixo aproveitamento de 3 pontos, pode ser que ao aumentar o percentual de aproveitamento, as chances de vitória aumentem significativamente.

Uma outra motivação para esse estudo é que, com um bom modelo preditivo, pode-se utilizar para comparar com sites de apostas e eventualmente ser um modelo lucrativo a médio e longo prazo. Os sites de aposta foram legalizados em dezembro de 2018 e, apesar da precária regulamentação, já movimentam bastante dinheiro. Segundo o [El pais, 2022], em 2021, foram movimentados nos sites de aposta do Brasil cerca de R$ 12 bilhões, um mercado bastante expressivo.

As demais seções deste trabalho estão organizadas da seguinte forma: No Capítulo 2, apresentamos alguns estudos presentes na literatura atual, no Capítulo 3, as metodologias que iremos estudar nessa dissertação, no Capítulo 4, apresentamos de onde os dados foram extraídos e como foram tratados, no Capítulo 5, analisamos os principais resultados e, por fim, no Capítulo 6, apresentamos a conclsusão desse estudo.

2 Modelos e métodos preditivos em basquetebol

Nos últimos anos, o uso de modelos preditivos tem se tornado cada vez mais comum no basquete. Com o avanço da tecnologia, e a disponibilidade de dados detalhados, as equipes e os analistas têm explorado o poder dos algoritmos e da análise estatística para prever resultados, identificar tendências e tomar decisões estratégicas mais embasadas. Modelos preditivos no basquete podem abranger desde a previsão de desempenho individual dos jogadores até a estimativa de resultados de partidas, a otimização de escalações e até mesmo a análise de risco de lesões. Essas abordagens preditivas têm se mostrado valiosas para as equipes da NBA e outras ligas, permitindo uma compreensão mais aprofundada do jogo e uma vantagem competitiva no planejamento e na tomada de decisões.

Como mencionado anteriormente, hoje em dia há muitos estudos sobre diversos métodos de predição. Apesar do avanço da área, muitos modelos ainda dependem do uso de estatísticas da partida que será prevista como *input* para a predição. Apesar disso aumentar a acurácia dos modelos, o uso de estatísticas do próprio jogo inviabiliza algumas aplicações, como seu uso para preparar treinamentos, fazer modificações táticas durante o jogo e, até mesmo, sua utilização para planejar apostas esportivas.

Na próxima seção apresentamos uma breve revisão sobre alguns dos principais modelos utilizados para a previsão de resultados de basquete. Esses modelos serviram de referência bibliográfica para a proposição do modelo apresentado no Capítulo 3 e, também, como base de comparação.

2.1 Modelos de regressão e de inteligência artificial

Em [Huang and Lin, 2020], foram utilizados alguns métodos preditivos como Regressão Linear, Árvore de regressão M5P e Regressão de Suporte de Vetores (SVR) para prever a pontuação entre 2 equipes que se enfrentaram. Os autores mostram que o método Árvore de regressão M5P alcançou a acurácia de 87,5%, o que é um resultado bastante interessante. Porém, em todas as análises, foram utilizadas as estatísticas do próprio jogo que se deseja prever, o que torna falho no sentido de tentar prever um jogo futuro, onde não se sabe ainda as estatísticas do jogo. Isso significa que os modelos foram treinados com um histórico de partidas e o teste para ver se o método estava funcionando também utilizou os dados da própria partida. A Tabela 2.1 apresenta um exemplo de

resultados obtidos no estudo de [Huang and Lin, 2020] com 16 jogos da equipe *Golden State Warriors* na temporada 2017-2018.

Time 1	**Pts time 1**	**Pts time 2**	**Time 2**	**Previsão**	**Real**
GSW	108	106	MIN	Vitória	Derrota
GSW	118	106	LAL	Vitória	Vitória
GSW	96	100	SAC	Derrota	Derrota
GSW	122	111	PHX	Vitória	Vitória
GSW	78	86	SAS	Derrota	Derrota
GSW	104	99	ATL	Vitória	Vitória
GSW	86	109	UTH	Derrota	Derrota
GSW	81	93	IND	Derrota	Derrota
GSW	107	114	MIL	Derrota	Derrota
GSW	112	97	SAC	Vitória	Vitória
GSW	117	114	PHX	Empate	Vitória
GSW	111	104	OCT	Vitória	Vitória
GSW	106	131	IND	Derrota	Derrota
GSW	120	129	NOP	Derrota	Derrota
GSW	117	100	PHX	Vitória	Vitória
GSW	79	119	UTH	Derrota	Derrota

Tabela 2.1 – Resultados de Huang

Uma abordagem semelhante ocorre em estudos como [Jain and Kaur, 2017] e [Pai et al., 2017]. No caso de [Jain and Kaur, 2017], métodos *Hybrid Fuzzy Support vector machines* (HFSVM) e *Support vector machines* (SVM) são utilizados, além de dar uma atenção especial para analisar quais variáveis impactam mais o resultado final da partida. Ao treinar o algoritmo com uma amostra de 640 partidas e testar em 160, é possível observar que os rebotes defensivos apresentam cerca de 40% de importância, o maior dentre todas as estatísticas analisadas, contra menos de 5% de importância do percentual de acertos de lances livres.

Os métodos HFSVM e SVM utilizam os dados de treinamento para encontrar vetores que minimizam as distâncias entre os pontos. No caso de análises mais complexas com várias variáveis, os vetores se tornam hiperplanos devido a quantidade de variáveis que os vetores armazenam, mas seguem o mesmo princípio.

Na Tabela 2.2, é possível observar os resultados médios de cada um dos métodos apresentados no estudo de [Jain and Kaur, 2017]. O método HFSVM é o que apresenta o melhor resultado, com mais de 88% de assertividade. O método SVM apresenta um resultado muito parecido mas um pouco inferior com 86% de assertividade.

Método	**Acurácia**	**Número de vetores**
HFSVM	88,26%	391
SVM	86,21%	547

Tabela 2.2 – Resultados de Jain

No caso de [Pai et al., 2017], há uma mistura de métodos como SVM e HSVMDT que tem uma abordagem mais de árvore de regressão. Os resultados são bastante interessantes, chegando a 85% de assertividade.

Em [Cao, 2012], há uma abordagem um pouco diferente testando outros métodos tais como *Simple Logistics*, *Native Bayes*, SVM e *Neural Network*. Os resultados obtidos são um pouco inferiores aos dos estudos mencionados anteriormente, não superando os 70% de assertividade como podemos ver na tabela 2.3.

Método	**Acurácia**
Simple Logistics	69,67%
Native Bayes	66,25%
SVM	67,70%
Neural Network	68,01%

Tabela 2.3 – Resultados de Chenjie

A abordagem de [Puranmalka, 2013] vai um pouco além das que vimos até agora. Nesse estudo, usa-se as estatísticas também por jogador que certamente tem um peso interessante no resultado final. É de se imaginar por exemplo que se o melhor jogador do time não jogar, o time vai ter um desempenho pior, ou pelo menos terá que se adaptar para conseguir manter a performance.

Os resultados gerais ficaram em torno de 70% como pode ser visto na tabela 2.4.

Ano	**Acurácia SVM**
2003	73,45%
2004	73,20%
2005	72,45%
2006	72,95%
2007	69,95%
2008	72,70%
2009	69,45%
2010	71,20%
2011	70,20%
2012	67,20%

Tabela 2.4 – Resultados de Puranmalka

Um outro estudo que também serviu como inspiração para algumas de nossas análises foi o [Fadi Thabtah1, 2019]. Além do estudo sobre a previsão de resultados de basquete com os métodos *Naive Bayes*, *Artificial Neural Networks* (ANN) e *Logistical Model Tree* (LMT), ele faz variações com a quantidade de variáveis a serem analisadas pelos métodos. Mais especificamente, o estudo analisou 5 grupo de variáveis, sendo que algumas variáveis estão em mais de um grupo como pode ser visto na tabela 2.5. Os resultados variam entre os métodos e os grupos de variáveis como apresentado na tabela 2.6.

Grupo	**Quantidade de variáveis**
A	22
B	7
C	7
D	8
E	5

Tabela 2.5 – Grupo de variáveis Fadi

Grupo	**Método**	**Acurácia**
A	*Naive Bayes*	76%
A	ANN	83%
A	LMT	82%
B	*Naive Bayes*	73%
B	ANN	71%
B	LMT	75%
C	*Naive Bayes*	77%
C	ANN	74%
C	LMT	78%
D	*Naive Bayes*	80%
D	ANN	80%
D	LMT	83%
E	*Naive Bayes*	78%
E	ANN	76%
E	LMT	79%

Tabela 2.6 – Resultados de Fadi

Uma outra possível abordagem encontrada em [Lieder, 2018] é a inclusão de variáveis externas tais como: dias de descanso entre uma partida e outra, sequência de vitórias etc. Os resultados foram razoáveis não ultrapassando os 70% de acurácia.

Como podemos ver, há inúmeras possibilidades de abordagens e métodos a serem utilizados para prever os resultados das partidas. Há ainda outros estudos que buscam analisar outras perspectivas do jogo, como por exemplo em [Soliman et al., 2017] onde é estudada as chances do jogador ser escolhido como um *All Star*, isto é, como um dos melhores jogadores da liga.

Todos os estudos mencionados anteriormente serviram de certa forma como inspiração para essa dissertação, de tal forma a guiar uma linha de raciocínio para aprofundar os estudos relacionados a previsão de resultados no basquete.

O desenvolvimento de métodos mais precisos de predição de resultados pode ter grande utilidade para o entendimento da performance das equipes e até mesmo para focar no desenvolvimento de atributos que podem levar a equipe se tornar mais vencedora no médio e longo prazo. Um exemplo muito prático é simular após o jogo se a equipe tivesse alguns rebotes a mais, um aproveitamento melhor de 3 pontos, ou se tivesse feito menos faltas, o que teria acontecido com o resultado da partida? Muitas das vezes poderia levar o time de uma derrota para uma vitória.

A partir dessas reflexões, procurou-se seguir uma outra linha de raciocínio, de forma que fosse possível prever os resultados antes das partidas acontecerem. Como já foi dito, isso poderia ser de grande valia para as equipes e outros setores interessados nos resultados. Antes das partidas, seria possível montar esquemas táticos que fossem mais eficazes contra outras equipes, ou jogadores que são mais eficientes em determinadas estatísticas (roubo de bola, percentual de acerto de arremesso etc). Não só isso, mas também seria possível treinar a equipe em fundamentos em que seriam determinantes para vencer tal partida analisada, pois com uma breve simulação da partida melhorando alguma estatística, seria possível ver o impacto na pontuação dos times e, consequentemente, o resultado final.

Os estudos que serviram de inspiração tentam de certa forma se aprofundar nas previsões das partidas, mas como utilizam as estatísticas do próprio jogo que desejam prever, não conseguem ter uma aplicação prática para a comissão técnica do time agir antes da partida. Por isso, nosso estudo pretende se apoiar em métodos de *machine learning* para conseguir prever os resultados das partidas antes mesmo delas acontecerem. Mais do que apenas prever os resultados, o estudo traz uma visão comparativa clara da lacuna de oportunidade para futuros estudos em elaborar métodos cada vez mais sofisticados para serem cada vez mais assertivos.

3 Metodologia

A partir do que foi encontrado na literatura, 3 métodos foram escolhidos para serem testados quanto a previsão de resultados dos jogos de basquete da NBA. Tais métodos serão detalhados ao longo deste capítulo, mas antes apresentaremos uma breve apresentação de todos os métodos citados nesta dissertação.

3.1 Métodos de aprendizado de máquina utilizados como referência

3.1.1 SVM

SVM (Support Vector Machine), ou Máquina de Vetores de Suporte, é um algoritmo de aprendizado de máquina amplamente utilizado em diversos campos, incluindo análise de dados esportivos como o basquete. O SVM é uma técnica de classificação que busca encontrar um hiperplano ótimo para separar dados em diferentes classes. No contexto do basquete, o SVM pode ser aplicado para classificar jogadores em diferentes categorias, como por posição ou estilo de jogo. Ele é capaz de analisar diversos atributos dos jogadores, como estatísticas individuais, habilidades específicas e características físicas, para criar fronteiras de decisão que melhor separam os jogadores em grupos distintos. Com isso, o SVM permite uma análise mais precisa e uma compreensão mais aprofundada das características individuais e da dinâmica coletiva das equipes no basquete.

3.1.2 Naive Bayes

O método Naive Bayes é um algoritmo de aprendizado de máquina baseado no Teorema de Bayes, que busca realizar classificações utilizando a probabilidade condicional das variáveis. No contexto do basquete, o Naive Bayes pode ser aplicado para diferentes propósitos. Por exemplo, pode ser utilizado para prever a probabilidade de sucesso em arremessos, considerando variáveis como distância do arremesso, posição do jogador, marcação defensiva, entre outros. Além disso, o Naive Bayes pode ser empregado para classificar jogadas ofensivas ou defensivas, considerando informações como ação tomada pelos jogadores, posicionamento em quadra e resultado final da jogada. Com sua capacidade de realizar classificações baseadas em probabilidades, o Naive Bayes proporciona uma abordagem analítica interessante para analisar o basquete e tomar decisões estratégicas com base em dados estatísticos.

3.2 Métodos de aprendizado de máquina utilizados neste trabalho

Os métodos utilizados (M5P, *Random Forests* e XGBoost) são métodos que apresentaram bons resultados na literatura, então por terem boas referências, é um bom ponto de partida. Apesar da lógica do uso dos métodos serem parcialmente diferentes do que encontrado na literatura, é esperado que eles mantenham bons desempenhos e por isso foram escolhidos.

Em todos os métodos e na extração de dados, foram utilizados a linguagem de programação Python por se tratar de uma poderosa linguagem muito utilizada hoje em dia. Além disso, possui uma comunidade na internet muito disposta a interagir e colaborar.

O uso dos métodos abaixo tinha como objetivo trazer as diferenças que o mesmo método pode ter quando utilizado exclusivamente com os dados reais e quando utilizado com os dados estimados das partidas, isto é, usando estimativas das principais estatísticas uma vez que a partida ainda não ocorreu. As diferenças ficarão mais claras no capítulo em que mostraremos os resultados.

A ideia é que os métodos consigam prever através da pontuação dos times o resultado da partida. Por exemplo, se o time A estiver enfrentando o time B, os métodos preveem a pontuação do time A e a pontuação do time B e, consequentemente, o resultado final da partida comparando as 2 pontuações.

3.2.1 M5P

O método M5P é um algoritmo de aprendizado de máquina baseado em árvores de regressão. Ele é utilizado para realizar previsões numéricas em conjunto com a exploração de regras de decisão. No contexto do basquete, o método M5P pode ser aplicado para realizar previsões de variáveis contínuas, como a pontuação média de uma equipe em uma partida, com base em diferentes atributos, como estatísticas de jogadores, histórico de confrontos e condições de jogo. O M5P busca identificar os relacionamentos e padrões entre as variáveis para construir uma árvore de decisão que permita realizar previsões precisas. Essa abordagem auxilia na compreensão dos fatores que influenciam o desempenho das equipes, auxiliando na análise estratégica do basquete e no planejamento tático durante as partidas.

O M5P é um método desenvolvido por [Wang and Witten, 1997] e que na verdade foi um aprimoramento do método M5 desenvolvido por [Quinlan, 1992]. Esse método se utiliza usando modelo de árvore de regressão binária tendo no último nó da árvore a função que gera a previsão do resultado desejado.

O método M5P é dividido basicamente em 4 etapas. A primeira etapa busca dividir a árvore em vários subespaços da matriz de *input* a fim de minimizar o erro padrão dos nós, isto é, ela procura dividir a árvore de forma que o erro padrão diminua e consequentemente tenha uma melhor precisão na previsão dos resultados. A segunda etapa busca encontrar uma regressão linear em cada subespaço criado na etapa anterior. A terceira etapa define as 'podas' dos galhos das árvores a fim de não acontecer o chamado *overtraining*, que é quando os modelos preditivos ficam viciados na matriz de *input* e acabam sendo bons preditores apenas daquela matriz de *input*. A quarta e última etapa é um processo para suavizar possíveis interrupções abrutas obtidas na etapa anterior.

A figura 3.1 mostra de forma simples e didática como funciona a lógica da árvore de regressão. Em cada nó da árvore busca-se um 'caminho' para minimizar o erro padrão. Ao final de cada caminho é possível obter a equação que otimiza a predição de acordo com os valores da variáveis da matriz de *input*.

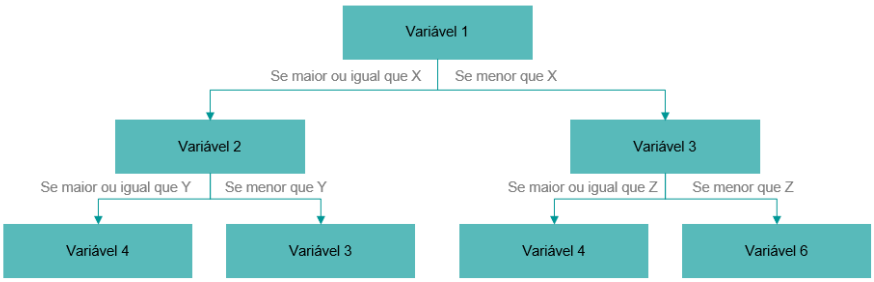

Figura 3.1 – Exemplo de esquema de árvore de regressão

Neste estudo foram analisadas 3 parametrizações do método: A primeira parametrização testada foi tirando a etapa de suavização e de poda; a segunda parametrização testada foi tirando apenas a etapa de suavização e a terceira parametrização foi considerando suavização e a poda. Para executar o método, utilizou-se da biblioteca 'm5py' no Python.

3.2.2 Random Forests

O método Random Forests, ou Florestas Aleatórias, é uma poderosa abordagem de aprendizado de máquina que utiliza conjuntos de árvores de decisão para realizar tarefas de classificação e regressão. O diferencial das Florestas Aleatórias é sua capacidade de criar múltiplas árvores de forma independente e combiná-las para obter uma predição final mais precisa e robusta. Cada árvore é treinada em um subconjunto aleatório de dados de treinamento e usa uma combinação de características selecionadas aleatoriamente, garantindo diversidade e reduzindo a probabilidade de *overfitting*, que é quando o método fica viciado na amostra de treinamento e acaba não representando um caso geral e sim um

caso específico daquela amostra de treinamento. Logo, os resultados de teste começam a apresentar uma assertividade bem inferior ao da amostra de treinamento. Durante a predição, as árvores individuais votam em uma classe (no caso da classificação) ou fornecem uma estimativa (no caso da regressão), e o resultado final é determinado pela maioria dos votos ou pela média das estimativas. As *Random Forests* são amplamente aplicadas em diversas áreas, como ciência de dados, bioinformática e detecção de fraudes, devido à sua eficiência, precisão e capacidade de lidar com conjuntos de dados complexos. No contexto do basquete, o método *Random Forests* pode ser aplicado de diversas maneiras. Por exemplo, ele pode ser utilizado para prever o desempenho de jogadores em diferentes estatísticas, como pontos, rebotes ou assistências, levando em consideração variáveis como histórico de partidas, idade, posição em quadra e outras características relevantes. Além disso, o *Random Forests* pode ser empregado para classificar jogadores em diferentes grupos ou estilos de jogo, identificar padrões de sucesso em jogadas ofensivas ou defensivas e até mesmo para prever resultados de partidas. Sua capacidade de lidar com múltiplas variáveis e a robustez do método, que combina múltiplas árvores de decisão, tornam o *Random Forests* uma ferramenta valiosa para a análise e tomada de decisões estratégicas no basquete.

O método é bastante sofisticado e muito utilizado na literatura, como em [Soliman et al., 2017]. O método *Random Forests* treina cada árvore de maneira independente usando uma amostra aleatória da matriz de *input*, dessa forma, ela evita o *overtraining*, como mencionado na seção anterior. Além disso, como ela gera uma grande quantidade de árvores, é possível minimizar o erro de forma mais eficiente e encontrar uma boa predição. Para executar este método, utilizou-se da biblioteca sklearn, a função RandomForestRegressor no Python.

3.2.3 XGBoost

O método XBG, também conhecido como Extreme Boosted Gradient [Chen and Guestrin, 2016], é uma técnica avançada de aprendizado de máquina que combina os princípios do algoritmo de impulso (boosting) e do gradiente extremo (extreme gradient). Ele se destaca por sua capacidade de lidar com problemas complexos de classificação e regressão, onde a precisão e a interpretabilidade são essenciais. O XBG utiliza uma combinação de árvores de decisão altamente adaptativas, que são treinadas sequencialmente para corrigir os erros cometidos pelos modelos anteriores. Além disso, o método incorpora estratégias de reamostragem inteligentes e regularização para evitar *overfitting*, resultando em modelos altamente precisos e robustos. O XBG tem sido amplamente utilizado em uma variedade de campos, incluindo finanças, medicina e ciência de dados, demonstrando sua eficácia e versatilidade em diversas aplicações. Ele costuma ter um desempenho superior,

principalmente, em casos de previsão envolvendo dados não estruturados ou com dados faltantes. Também é um método que em sua natureza evita *overtraining*.

O XGBoost é também um método de aprendizado de máquina baseado em árvore de decisão e que utiliza uma estrutura de *Gradient Boosting*, que é um reforço em que os erros são minimizados pelo algoritmo de gradiente descendente (*gradient descent*) [Datageeks, 2022]. O método amplifica a estrutura de *Gradient Boosting* tornando um método bastante eficiente.

3.2.4 Comitê de escolha

Como pôde ser visto, os métodos são diversos entre si e consequentemente apresentam resultados um tanto quanto diferentes, isto é, para a mesma partida o método A pode ter acertado e o método B pode ter errado ou vice-versa. Dado isso, o objetivo dessa seção é tentar criar uma lógica que funcionaria como comitê de escolha. Ou seja, a partir do resultado dos métodos mencionado acima, haveria uma lógica para escolher o resultado ou melhor método.

3.2.4.1 Escolha pela média

Nessa opção, foram coletados os placares previstos das 5 opções e feito uma média dos 5 placares. Evidentemente que não era esperado uma mudança muito brusca dos principais resultados, pois nesse critério leva-se muito em conta os resultados obtidos nas seções anteriores. A Figura 3.2 ajuda a ilustrar como se deu a elaboração desse comitê de escolha.

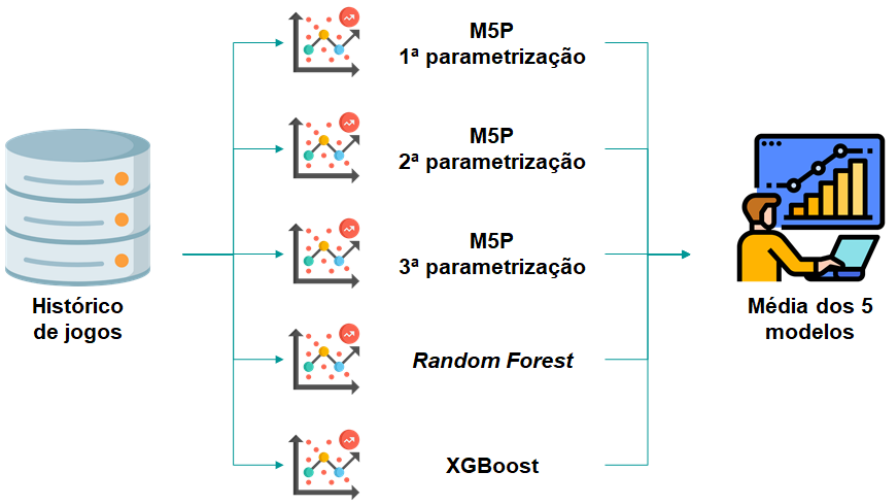

Figura 3.2 – Esquema do comitê de escolha baseado na média dos resultados dos modelos

3.2.4.2 Escolha pela maioria

Para esse caso, a ideia era analisar o resultado de cada uma das 5 opções das seções anteriores e decidir pela maioria. Ou seja, se 3 das opções estavam prevendo que o time A iria vencer, a previsão final seria vitória para o time A. Com isso, não fazia sentido analisar os grupos com a diferença de pontos entre os times, pois aqui o *output* é apenas a previsão de quem venceu a partida. A Figura 3.3 ajuda a ilustrar como se deu a elaboração desse comitê de escolha.

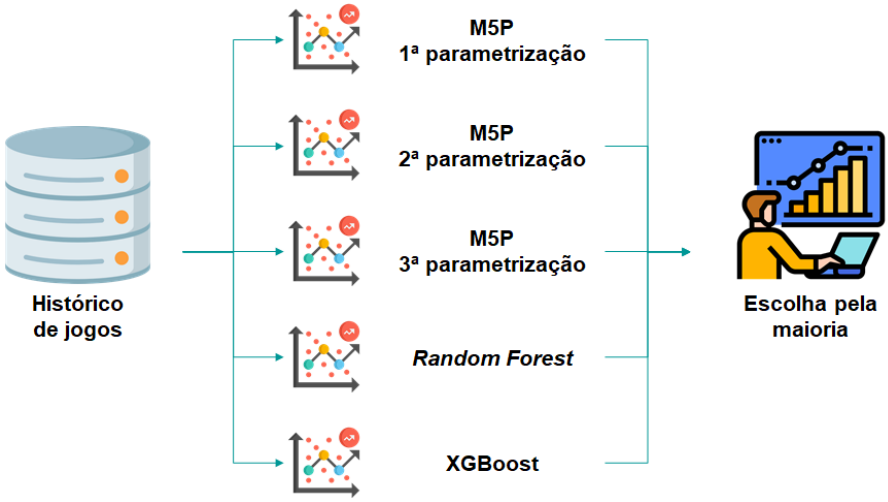

Figura 3.3 – Esquema do comitê de escolha baseado na maioria como critério de escolha

3.2.4.3 Escolha por comportamento histórico

Uma outra abordagem escolhida foi um pouco mais complexa. A ideia principal era tentar encontrar no histórico jogos 'parecidos' e avaliar qual método havia acertado naquele jogo 'parecido' e assumir que esse era o melhor método.

Para definir um jogo 'parecido', foi utilizado a distância entre as estatísticas do jogo que estamos tentando prever e a as estatísticas do jogos históricos. Para calcular a distância, foi feita a soma dos quadrados da diferença entre cada estatística e depois tirado raiz. Aquele que tivesse o menor resultado, seria considerado o jogo mais 'parecido'. Se por acaso o jogo mais 'parecido' não tivesse nenhum acerto entre os 5 métodos, seria escolhido o segundo jogo mais 'parecido' e assim sucessivamente. A equação 3.1 ajuda a ilustrar como seria a conta para comparar jogo A e o jogo B, e cada letra (J, K, Z) seria as respectivas variáveis analisadas.

$$\Delta = \sqrt{(J_A - J_B)^2 + (K_A - K_B)^2 + ... + (Z_A - Z_B)^2} \tag{3.1}$$

Para que fosse justa a conta, todas as estatísticas foram normalizadas entre 0 e 1, dando pesos iguais para cada uma delas. Por exemplo, para tentativas de arremessos

de 2 pontos, o jogo que teve o maior número de arremessos terá 1 como o valor da variável normalizada e o jogo que tiver a menor quantidade de arremessos de 2 pontos terá 0 como valor da variável normalizada. Todos os valores intermediários terão valores proporcionais entre 0 e 1.

Após definido o jogo mais parecido, como os métodos estavam com desempenhos gerais relativamente parecidos, foi escolhido arbitrariamente o método do jogo a ser previsto. A lógica ficou assim: se o método M5P (3ª parametrização) tiver acertado, ele seria escolhido, caso contrário seria checado o método *Random Forests* e caso contrário o método XGBoost.

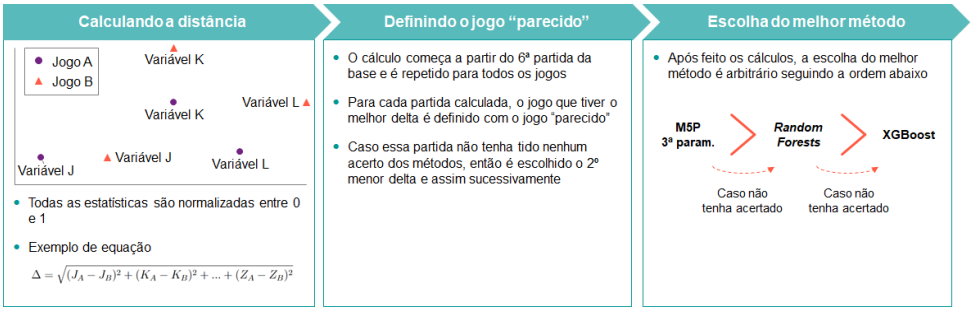

Figura 3.4 – Esquema do comitê de escolha baseado no histórico como critério de escolha

4 Estudo de caso: Uma aplicação com base em dados reais da NBA

Neste Capítulo apresentamos uma aplicação dos métodos apresentados no Capítulo 3 com base em dados reais da NBA das temporadas 2018-2019 até 2021-2022.

Para esse estudo foi necessário a extração de alguns dados, como por exemplo as estatísticas dos jogos, o histórico de vitórias e derrotas dos jogos, e também as *odds* de apostas do jogos analisados. Além disso, alguns dados precisaram de alguns tratamentos para que pudessem ser analisados corretamente. Todo esse processo de extração e tratamento dos dados será explicado nas próximas seções.

4.1 Estatísticas dos jogos

A coleta de dados se iniciou com uma pesquisa de sites que poderiam fornecer as estatísticas necessárias para análise histórica. Dentre diversos sites que existem hoje em dia tal como o próprio site da NBA, o site da ESPN entre outros, optou-se pelo site [Basketball Reference, 2022]. Esse site possui diversas visualizações dos dados, além de possuir um histórico razoável e possuir todas as estatísticas necessárias para as análises.

A partir de algumas referências bibliográficas começou a seleção de quais estatísticas deveriam ser coletadas. Essa seleção foi feita em conjunto com quais dados a fonte [Basketball Reference, 2022] teria à disposição para se iniciar a coleta. Depois de feita essa análise, foram selecionadas as estatísticas da 4.1 para coleta dos dados.

Variável	Descrição	Tipo
2PA	Arremessos de 2 pontos	Ataque
2P%	Percentual de acerto de arremessos de 2 pontos	Ataque
3PA	Arremessos de 3 pontos	Ataque
3P%	Percentual de acerto de arremessos de 3 pontos	Ataque
FTA	Arremessos de lance livre	Ataque
FT%	Percentual de acerto de arremessos de lance livre	Ataque
ORB	Rebote ofensivo	Ataque
AST	Assistências	Ataque
TOV	*Turnover* - erro que implica em perda da posse de bola	Ataque
H/A	Se está jogando em casa ou fora de casa	Ataque
PTS	Quantidade de pontos feitos	Ataque
STREAK	Quantidade de vitórias/derrotas seguidas nos últimos jogos	Ataque
DRB	Rebote defensivo	Defesa
STL	Roubo de bola	Defesa
BLK	*Block* ou toco	Defesa
TOV-FORC	Quantas vezes forçou *turnover* do oponente	Defesa
PF	Faltas	Defesa
PTS-CON	Quantos pontos o oponente fez no jogo	Defesa

Tabela 4.1 – Estatísticas extraídas para análise

Após a seleção de quais estatísticas seriam extraídas, foi feito o *web scraping* das estatísticas de cada time por jogo da temporada regular ao longo das temporadas 2018-2019 até 2021-2022. Os dados foram extraídos com a ajuda da linguagem Python aplicada na ferramenta do Google colab. O uso gratuito dessa ferramenta permite acesso a todos e democratiza bastante o acesso a diversos dados e análises para quem quiser. Além disso, apesar de não ter sido encontrado um guia específico de como executar o *web scraping* do site [Basketball Reference, 2022] há diversos *sites* e *blogs* que dão várias dicas e ajudam a guiar o caminho para conseguir construir o algoritmo.

Depois de feita a extração dos dados, a ideia era poder prever os pontos de ambos os time de um confronto. Por exemplo, imaginando que o time A fosse jogar contra o time B, a ideia era colocar nos modelos preditivos as estatísticas ofensivas do time A e as estatísticas defensivas do time B para poder prever os pontos do time A e, de maneira análoga, para prever os pontos do time B usar as estatísticas ofensivas do time B e as estatísticas defensivas do time A, como pode ser visto na Figura 4.1. A partir daí, comparar a pontuação do time A com a pontuação do time B para então, definir qual time seria o vencedor.

A extração é feita time a time, então para poder juntar as estatísticas ofensivas do time A e defensivas do time B foi utilizado a ferramenta Excel utilizando uma chave única contendo as 2 equipes e a data de cada jogo. A escolha do Excel se deu por uma maior afinidade com essa ferramenta, mas que também poderia ser facilmente substituída

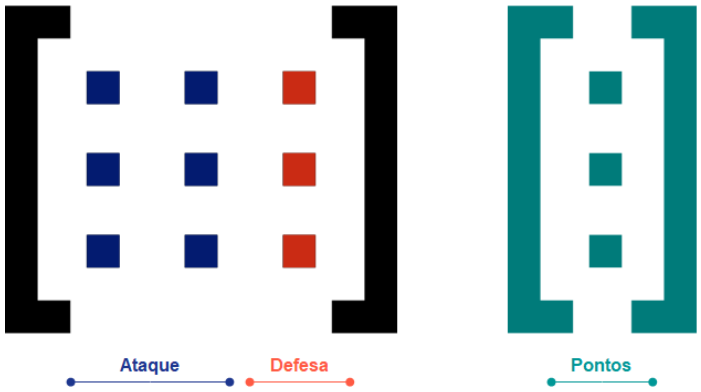

Figura 4.1 – Ilustração das matrizes analisadas pelos modelos

pela combinação de Python e Google colab.

4.1.1 Streak de jogos

Uma das estatísticas mencionadas anteriormente era a de vitórias ou derrotas seguidas pela mesma equipe. A ideia era ter uma métrica que fosse capaz de trazer o momento de cada equipe, e caso o momento fosse favorável isso deveria pesar a favor caso o momento não fosse favorável deveria pesar contra. Por exemplo, se uma equipe viesse embalada de muitas vitórias deveria haver uma maneira de metrificarmos isso, da mesma forma que se a equipe viesse embalada de muitas derrotas.

Uma forma encontrada para conseguir aplicar isso foi avaliar quantas vitórias ou derrotas consecutivas cada equipe tinha até o momento da partida a ser analisada. Nesse caso foi necessário uma extração diferente no próprio site [Basketball Reference, 2022] para conseguir o histórico de cada time jogo a jogo. Depois de extraído essa informação, foi feito no Excel um cruzamento dos jogos para inserir essa estatística na tabela geral com as demais estatísticas.

A lógica criada para mensurar essa variável foi para somar 1 a cada vitória consecutiva e para cada derrota consecutiva, se subraía 1 valor unitário. Ou seja, se o time estivesse com 3 derrotas consecutivas, sua estatística seria -3 e se tivesse com 5 vitórias consecutivas seria +5.

4.1.2 *Dataset* reduzido

Ao buscar mais referências bibliográficas, foi encontrado o estudo de [Fadi Thabtah1, 2019] no qual simula alguns *datasets* diferentes para poder avaliar os desempenho com um número reduzido de estatísticas. Essa redução pode ser benéfica, pois muitas

variáveis juntas podem acabar trazendo algumas que não são tão relevantes e acabar 'atrapalhando' o desempenho dos algoritmos preditivos.

A ideia foi então testar alguns *datasets* reduzidos e analisar a performance das previsões para poder avaliar se os resultados melhorariam de acordo com a variação desses *datasets*. Para a escolha dos *datasets*, foi utilizado como referência o próprio estudo de [Fadi Thabtah1, 2019].

Nesse caso foram escolhidos 2 *datasets* com bom desempenho de acordo com o artigo de [Fadi Thabtah1, 2019]. O primeiro dele consiste em 8 estatísticas conforme tabela abaixo. Vale ressaltar aqui que as estatísticas de defesa nesse caso são do próprio time, ao contrário do que fora feito nas análises anteriores onde as estatísticas de defesa são do time adversário.

Variável	**Descrição**	**Tipo**
2PA	Arremessos de 2 pontos	Ataque
2P%	Percentual de acerto de arremessos de 2 pontos	Ataque
3P%	Percentual de acerto de arremessos de 3 pontos	Ataque
FT	Pontos de lance livre	Ataque
TOV	*Turnover* - erro que implica em perda da posse de bola	Ataque
DRB	Rebote defensivo	Defesa
TRB	Total de rebotes	Defesa
PF	Faltas	Defesa

Tabela 4.2 – Estatísticas extraídas para análise com *dataset* reduzido

O segundo *dataset* considera apenas 5 estatísticas conforme tabela abaixo. De maneira similar ao item anterior, as estatísticas de defesa também são do próprio time.

Variável	**Descrição**	**Tipo**
2P%	Percentual de acerto de arremessos de 2 pontos	Ataque
3P%	Percentual de acerto de arremessos de 3 pontos	Ataque
FT	Pontos de lance livre	Ataque
DRB	Rebote defensivo	Defesa
TRB	Total de rebotes	Defesa

Tabela 4.3 – Estatísticas extraídas para análise com *dataset* reduzido

Dá para notar que as variáveis da tabela 4.3 já aparecem na tabela 4.2 mas de forma reduzida para o grupo de *dataset*.

4.2 Tratamento dos dados

Após a extração dos dados, a primeira coisa a ser feita foi a checagem geral dos dados para garantir que eles estavam fazendo sentido e que havia ali todas as partidas de

todas as temporadas. Além disso, também foi checado se as estatísticas extraídas estavam fazendo sentido, por exemplo, o percentual de acertos estava entre 0 e 1? A ideia aqui era garantir que os dados estavam consistentes para sequência das análises.

Como mencionado, a ideia era poder prever a pontuação, e consequentemente o resultado da partida, sem ter nenhuma estatística da partida, como se fôssemos prever um jogo que ainda não aconteceu de fato. A maneira como iria ser feito não teve muita referência bibliográfica, uma vez que os estudos costumam usar as próprias estatísticas históricas para alimentar os modelos.

A escolha feita nesse estudo foi de usar a média dos últimos 10 jogos para poder comparar o time A e o time B. Dessa forma é possível diluir alguns *outliers* e ainda assim ter um histórico consistente. A Figura 3.3 ilustra como é feito, puxando as estatísticas de ataque do time A e as estatísticas de defesa do time B para ter as estatísticas da partida a ser prevista.

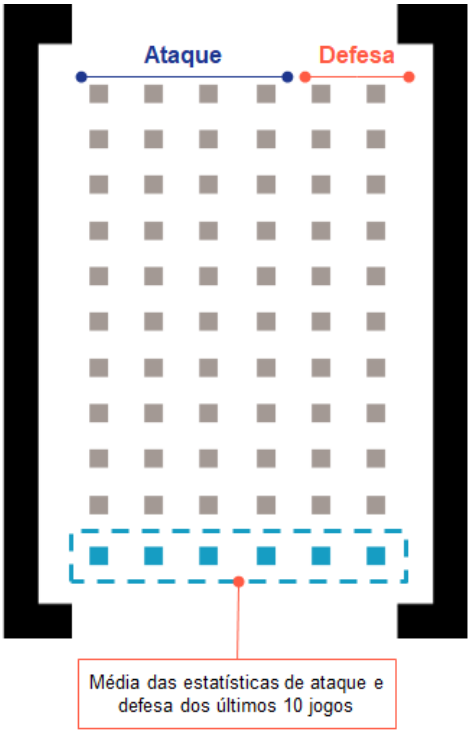

Figura 4.2 – Ilustração da média utilizada para predição dos resultados

Para alguns casos, tais como o *streak* de jogos é utilizado a soma as vitórias ou derrotas consecutivas até aquele jogo. No caso da variável que avalia se o jogo é em casa ou fora também é avaliado apenas aquele jogo em específico e não se usa a média e sim o dado daquela partida uma vez que sempre saberemos se a partida será dentro ou fora de casa. Para partidas dentro de casa foi considerado 1 e para partidas fora de casa foi considerado 0,5. Alguns estudos colocavam 0 quando a partida era fora de casa, mas acreditamos que apesar de haver uma desvantagem em jogar fora de casa, não deveria ser

tão discrepante assim. Além disso, quando os modelos forem treinar eles irão adaptar o peso da variável para compensar essa diferença.

Um outro ponto que vale observar é que alguns jogos estavam com algumas estatísticas faltando e por isso tiveram que ser desconsideradas. Além disso, como é preciso analisar a média dos últimos 10 jogos, os primeiros jogos da base de cada time tiveram que ser desconsideradas também uma vez que ainda não havia histórico suficiente. Com isso, algumas partidas da amostra extraída tiveram que ser desconsideradas, mas sem grandes prejuízos uma vez que o tamanho da amostra é relativamente bom.

4.3 Odds dos jogos

Ao analisar a assertividade dos métodos era difícil de saber se aquele número era bom ou não. Por exemplo, se um método acerta 60% é pouco ou muito? e se acertar 70%? Por isso, uma forma de balizar se os resultados estavam sendo satisfatórios ou não era supor uma aposta de R$ 1,00 para cada jogo previsto e avaliar se seria lucrativo ou não essas apostas.

Para isso, foi utilizado como referência as *odds* dos jogos. As *odds* é o valor de quanto um site de aposta paga para cada dinheiro apostado naquele jogo. De certa maneira, ele tem a ver com a probabilidade daquele evento acontecer. Por exemplo, se as *odds* para vitória do time A vencer estiver bem alta, significa que para cada real apostado o retorno tenderá a ser bem alto, pois a probabilidade disso acontecer deve ser relativamente baixa. Esse valor é dinâmico e varia de acordo com o quanto os apostadores estão acreditando mais ou menos em cada evento.

Para simplificar foi utilizado apenas as *odds* de vitória para cada equipe, seja do time A ou do time B. O mundo de apostas possui uma imensa variedade possível, como diferença de placar, quanto cada jogador irá pontuar ou qual o placar no intervalo etc.

Como podemos ver na figura 4.3, o time A com *odds* de 1,72 representa 58% de chances de vitória. Esse percentual é calculado pela divisão 1 por 1,72. Por outro lado, o time B apresenta 2,21 de *odds* e, calculando da mesma maneira, 45% de probabilidade de vitória. Ao somar as duas probabilidades de vitória, podemos chegar em 103% o que é de cara um pouco estranho. Essa 'sobra' que existe, é a margem que as casas de apostas possuem e por isso raramente chega-se em 100% essa soma.

Além disso, fazer a comparação com as *odds* é uma forma, ainda que simplista, de comparar com outras ferramentas de predição que possam existir no mercado, uma vez que as *odds* acabam variando com o que as pessoas acreditam que irá acontecer. É claro que a especulação impacta bastante o valor das *odds* e por isso é uma maneira bastante

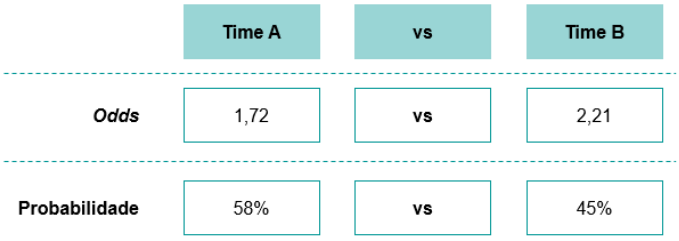

Figura 4.3 – Ilustração das *odds* entre o time A e o time B

simplista.

Como falamos inicialmente, comparar com as *odds* é uma forma de saber se a aplicação desses métodos para apostas esportivas poderia ser uma fonte lucrativa de dinheiro.

Para pegar as *odds* dos jogos, isto é, quanto que um site de apostas pagaria para cada aposta feita, foi utilizado como apoio o site [Odds Portal, 2022]. Como foi utilizado um valor histórico, foram utilizados os dados médios do [Odds Portal, 2022]. Lá no site, é possível obter as *odds* de diversos sites de apostas, mas para essa análise foi utilizado um valor médio geral.

Usando uma lógica similar de *web scraping* utilizado para extrair as estatísticas, foi extraído as *odds* de cada partida. Isto é, foi utilizado a linguagem de progrmação Python combinada com o uso da ferramenta Google colab. Após a extração dos dados, essas *odds* foram cruzadas com as previsões dos jogos através do excel, novamente por questão de afinidade com tal ferramenta.

5 Resultados e discussões

Nesse capítulo, mostraremos os principais resultados dos modelos escolhidos para análise. O principal objetivo era analisar se a previsão de vitória ou derrota estava certa. Houve ainda um refinamento para tentar entender se a assertividade aumenta para jogos mais óbvios, isto é, aqueles em que a diferença de pontos é muito alta. Para isso, os jogos foram classificados em 4 grupos: jogos que a previsão terminaram com a diferença de pontos entre 0 e 10, jogos que a previsão terminaram com a diferença de pontos entre 10 e 20, jogos que a previsão terminaram com a diferença de pontos entre 20 e 30 e jogos que a previsão terminaram com a diferença acima de 30 pontos, caso houvesse.

Em todas as seções, serão apresentadas tabelas contendo uma visão geral dos resultados que inclui, além dos grupos explicados anteriormente, a quantidade de previsões corretas dos jogos, a quantidade de jogos analisados e a assertividade das previsões por grupo. Também traremos em alguns casos as comparações por temporada para podermos analisar a consistência entre as temporadas.

Para fins comparativos e até para enriquecimento desse estudo, foi também analisado os resultados com os *datasets* reduzidos, conforme explicado anteriormente e, também, caso fosse usado sempre as estatísticas reais do jogo ao invés da média dos últimos jogos. Essa segunda análise é importante para termos uma referência se os resultados obtidos estão razoáveis ou não.

5.1 M5P

Conforme mencionado anteriormente, 3 parametrizações diferentes foram analisadas no método M5P visando entender as variações e qual seria a melhor opção.

Na tabela 5.1, podemos ver que a assertividade da primeira parametrização vai aumentando conforme a diferença entre os pontos prevista também aumenta, como era de se esperar. Para jogos com diferença prevista entre 20 e 30 pontos entre as duas equipes, a assertividade chega a 66%. Se considerarmos todos os jogos, a assertividade cai para 58%.

Grupo	Previsões corretas	Jogos analisados	Assertividade
0-10	1546	2755	56%
10-20	383	603	64%
20-30	19	29	66%
Total	1948	3387	58%

Tabela 5.1 – Resultados do modelo M5P - Parametrização 1

Já na tabela 5.2, podemos ver que a segunda parametrização tem um resultado geral levemente abaixo do que a anterior, obtendo um resultado geral de 57% de assertividade. De maneira geral, os resultados foram muito próximos ao da primeira parametrização.

Grupo	Previsões corretas	Jogos analisados	Assertividade
0-10	1533	2779	55%
10-20	375	589	64%
20-30	13	19	68%
Total	1921	3387	57%

Tabela 5.2 – Resultados do modelo M5P - Parametrização 2

Na terceira parametrização, um fato curioso pôde ser observado, a assertividade foi exatamente a mesma. Apesar de haver 3 jogos não coincidentes, os resultados gerais foram os mesmos, mostrando que essa parametrização, neste contexto, pouco influenciou no resultado final de predição. Os resultados podem ser vistos na tabela 5.3.

Grupo	Previsões corretas	Jogos analisados	Assertividade
0-10	1533	2779	55%
10-20	375	589	64%
20-30	13	19	68%
Total	1921	3387	57%

Tabela 5.3 – Resultados do modelo M5P - Parametrização 3

Tentando olhar por uma perspectiva de temporadas, gostaríamos de entender se alguma temporada divergia muito das outras ou se o método era consistente ao longo do tempo. Para todas as parametrizações, desconsideramos a primeira temporada de 2018-2019 porque haviam apenas 18 jogos previstos por conta dos dados utilizados para treinamento do algoritmo, portanto as previsões mostradas abaixo começam a partir da temporada 2019-2020.

Na tabela 5.4 podemos ver que os resultados da primeira parametrização foram bastante consistentes entre as temporadas, com uma assertividade entre 57% e 58%.

Temporada	Previsões corretas	Jogos analisados	Assertividade
2019-2020	611	1059	58%
2020-2021	630	1080	58%
2021-2022	698	1230	57%
Total	1939	3369	58%

Tabela 5.4 – Resultados do modelo M5P - Parametrização 1 por temporada

Na tabela 5.5 podemos ver que os resultados da segunda parametrização também foram bastante consistentes entre as temporadas, com uma assertividade entre 56% e 58%. Além disso, mostrou uma consistência em relação a primeira parametrização, pois de maneira geral tiveram resultados parecidos na mesma ordem de grandeza.

Temporada	Previsões corretas	Jogos analisados	Assertividade
2019-2020	598	1059	56%
2020-2021	627	1080	58%
2021-2022	688	1230	56%
Total	1913	3369	57%

Tabela 5.5 – Resultados do modelo M5P - Parametrização 2 por temporada

Na tabela 5.6 podemos ver que os resultados da terceira parametrização apresenta novamente um desempenho muito próximo ao da segunda parametrização, como era de se esperar. A assertividade ficou entre 56% e 58%.

Temporada	Previsões corretas	Jogos analisados	Assertividade
2019-2020	598	1059	56%
2020-2021	626	1080	58%
2021-2022	689	1230	56%
Total	1913	3369	57%

Tabela 5.6 – Resultados do modelo M5P - Parametrização 3 por temporada

5.2 Random Forests

O método *Random Forests* foi inicialmente testado com 10 árvores, mas em seguida foi-se aumentando gradualmente a quantidade de árvores do método. Os resultados aqui expostos na tabela 5.7, consideram 100 árvores em sua 'floresta', pois essa parametrização foi a que apresentou o melhor resultado. Vale ressaltar que os números variaram pouco e, por isso, não estão aqui os resultados com menos árvores.

Podemos ver na tabela 5.7, que nenhum jogo foi previsto com mais do que 30 pontos de diferença e apenas 22 foram previstos com uma diferença de 20 a 30 pontos. Entretanto, apesar dos resultados previstos apresentarem uma baixa diferença entre os

pontos de cada time (cerca de 81% dos jogos com menos de 10 pontos de diferença), a assertividade geral foi de 57%, isto é, praticamente igual ao do método M5P. Além disso, se considerarmos apenas os jogos com uma diferença de 10 a 20 pontos, a assertividade vai para 64%, o que não é tão ruim dado que 611 jogos se encaixam nesse intervalo de pontos.

Grupo	**Previsões corretas**	**Jogos analisados**	**Assertividade**
0-10	1525	2754	55%
10-20	391	611	64%
20-30	15	22	68%
Total	1931	3387	57%

Tabela 5.7 – Resultados do modelo Random Forests

Assim como no método M5P, podemos ver na tabela 5.8 que o método *Random Forests* também apresenta uma boa consistência nos resultados ao longo das temporadas, com uma assertividade variando entre 56% e 58%. Na tabela 5.8 podemos ver os resultados por temporada do método.

Temporada	**Previsões corretas**	**Jogos analisados**	**Assertividade**
2019-2020	605	1059	57%
2020-2021	625	1080	58%
2021-2022	693	1230	56%
Total	1923	3369	57%

Tabela 5.8 – Resultados do modelo Random Forests por temporada

5.3 XGBoost

De maneira parecida ao do *Random Forests*, o método XGBoost não teve jogos previstos com uma diferença no placar de mais de 30 pontos, poucos jogos previstos com uma diferença de placar entre 20 e 30 pontos além de uma assertividade geral de apenas 57%. Podemos ver na tabela 5.9, os resultados em cada intervalo de pontos.

Grupo	**Previsões corretas**	**Jogos analisados**	**Assertividade**
0-10	1526	2782	55%
10-20	381	585	65%
20-30	13	20	65%
Total	1920	3387	57%

Tabela 5.9 – Resultados do modelo XGBoost

Novamente, os resultados por temporada também foram bastante consistentes ficando entre 55% e 58% como pode ser visto na tabela 5.10.

Temporada	Previsões corretas	Jogos analisados	Assertividade
2019-2020	598	1059	56%
2020-2021	631	1080	58%
2021-2022	682	1230	55%
Total	1911	3369	57%

Tabela 5.10 – Resultados do modelo XGBoost por temporada

5.4 Comitê de escolha

A fim de tentar melhorar os resultados obtidos neste capítulo, buscou-se testar alternativas chamadas de 'Comitê de escolha'. Esse comitê busca encontrar, de maneira relativamente simples, uma escolha neutra entre as 5 opções analisadas anteriormente.

5.4.1 Escolha pela média

Na tabela 5.11 pode-se ver que de fato o resultado geral não difere muito dos analisados anteriormente, chegando a uma assertividade de 57%. Um resultado bastante esperado uma vez que as assertividades dos métodos anteriores foram próximas umas dos outras.

Grupo	Previsões corretas	Jogos analisados	Assertividade
0-10	1538	2771	56%
10-20	379	595	64%
20-30	14	21	67%
Total	1931	3387	57%

Tabela 5.11 – Resultados do comitê pela média

Do mesmo modo, era de se esperar que os resultados por temporada fossem bastante parecidos com os dos métodos anteriores, e isso é possível observar na tabela 5.12.

Temporada	Previsões corretas	Jogos analisados	Assertividade
2019-2020	602	1059	57%
2020-2021	631	1080	58%
2021-2022	690	1230	56%
Total	1923	3369	57%

Tabela 5.12 – Resultados do comitê pela média por temporada

5.4.2 Escolha pela maioria

Na tabela 5.13 podemos ver novamente que a assertividade foi bastante em linha com as opções anteriores, chegando a uma acurácia total de 57%. Nesse caso, esse

% de acerto mostra que os métodos, em sua grande maioria, tendem a ter resultados parecidos, isto é, costumam acertar e errar juntos.

Um ponto importante para ressaltar é que aqui não faz sentido olharmos pela diferença de pontos pois aqui o método segue pela maioria, isto é, se a maioria dos métodos indica vitória, esse comitê irá assumir vitória e o mesmo para derrota. Portanto, não faz sentido olharmos a pontuação, uma vez que esse método via comitê não prevê a pontuação.

Grupo	**Previsões corretas**	**Jogos analisados**	**Assertividade**
Total	1925	3387	57%

Tabela 5.13 – Resultados do comitê pela maioria

Da mesma maneira, os resultados por temporada seguem um comportante bastante parecido, como podemos observar na tabela 5.14.

Temporada	**Previsões corretas**	**Jogos analisados**	**Assertividade**
2019-2020	600	1059	57%
2020-2021	628	1080	58%
2021-2022	688	1230	56%
Total	1916	3369	57%

Tabela 5.14 – Resultados do comitê pela maioria por temporada

5.4.3 Escolha por comportamento histórico

Como explicado anteriormente, uma outra alternativa analisada era criar um algoritmo que fosse capaz de escolher o melhor método a partir do histórico de partidas. Ou seja, buscou-se partidas com características semelhantes na qual algum dos métodos havia acertado e assumir que esse método acertaria novamente.

Na tabela 5.14 podemos ver que os resultados não são muito diferentes dos apresentados anteriormente. Com uma assertividade geral de 56%, os resultados ficaram um pouco aquém do esperado para esse comitê.

Temporada	**Previsões corretas**	**Jogos analisados**	**Assertividade**
2019-2020	564	1055	53%
2020-2021	632	1080	59%
2021-2022	698	1230	57%
Total	1894	3365	56%

Tabela 5.15 – Resultados do comitê pelo comportamento histórico por temporada

5.5 Resultado geral considerando o dataset reduzido

Os *datasets* foram testados a fim de observar se reduzindo a quantidade de estatísticas analisadas poderia melhorar a assertividade dos métodos, pois em alguns casos, muitas variáveis podem ocasionar no *overfitting* e atrapalhar na assertividade geral. De qualquer maneira, o *dataset* reduzido pode trazer ganhos computacionais interessantes, pois poderia reduzir o tempo de processamento dos dados e, eventualmente, baratear os custos das análises.

Nas tabelas abaixo, tem-se os resultados do primeiro *dataset* testado como exemplificado na tabela 4.2. Os resultados encontrados não foram satisfatório e de maneira geral pioraram os resultados apresentados anteriormente quando foi utilizado todas as variáveis extraídas. Nenhum dos métodos obteve mais do que 54% de assertividade geral, o que é relativamente baixo.

Grupo	**Previsões corretas**	**Jogos analisados**	**Assertividade**
0-10	742	1456	51%
10-20	567	1071	53%
20-30	310	593	52%
+30	144	267	54%
Total	1763	3387	52%

Tabela 5.16 – Resultados do primeiro dataset do método M5P - 1ª parametrização

Grupo	**Previsões corretas**	**Jogos analisados**	**Assertividade**
0-10	771	1481	52%
10-20	576	1063	54%
20-30	278	524	53%
+30	180	319	56%
Total	1805	3387	53%

Tabela 5.17 – Resultados do primeiro dataset do método M5P - 2ª parametrização

Grupo	**Previsões corretas**	**Jogos analisados**	**Assertividade**
0-10	1055	1993	53%
10-20	567	1018	56%
20-30	161	288	56%
+30	46	88	52%
Total	1829	3387	54%

Tabela 5.18 – Resultados do primeiro dataset do método M5P - 3ª parametrização

Grupo	Previsões corretas	Jogos analisados	Assertividade
0-10	1313	2484	53%
10-20	468	808	58%
20-30	53	94	56%
+30	1	1	100%
Total	1835	3387	54%

Tabela 5.19 – Resultados do primeiro dataset do método Random Forests

Grupo	Previsões corretas	Jogos analisados	Assertividade
0-10	1053	2064	51%
10-20	576	1034	56%
20-30	146	257	57%
+30	17	32	53%
Total	1792	3387	53%

Tabela 5.20 – Resultados do primeiro dataset do método XGBoost

O mesmo aconteceu no segundo *dataset* conforme as estatísticas apresentadas pela tabela 4.3. Os resultados também foram relativamente baixos, podemos ver nas tabelas abaixo que novamente não passou 54% de assertividade geral mostrando que ambas tentativas dos *datasets* não trouxeram melhoras na previsibilidade das partidas.

Grupo	Previsões corretas	Jogos analisados	Assertividade
0-10	756	1520	50%
10-20	547	1068	51%
20-30	294	539	55%
+30	146	260	56%
Total	1743	3387	51%

Tabela 5.21 – Resultados do segundo dataset do método M5P - 1ª parametrização

Grupo	Previsões corretas	Jogos analisados	Assertividade
0-10	920	1803	51%
10-20	517	1007	51%
20-30	213	384	55%
+30	108	193	56%
Total	1758	3387	52%

Tabela 5.22 – Resultados do segundo dataset do método M5P - 2ª parametrização

Grupo	Previsões corretas	Jogos analisados	Assertividade
0-10	1167	2300	51%
10-20	501	894	56%
20-30	95	168	57%
+30	18	25	72%
Total	1781	3387	53%

Tabela 5.23 – Resultados do segundo dataset do método M5P - 3ª parametrização

Grupo	Previsões corretas	Jogos analisados	Assertividade
0-10	1286	2436	53%
10-20	487	849	57%
20-30	56	98	57%
+30	2	4	50%
Total	1831	3387	54%

Tabela 5.24 – Resultados do segundo dataset do método Random Forests

Grupo	Previsões corretas	Jogos analisados	Assertividade
0-10	1073	2048	52%
10-20	551	1022	54%
20-30	158	269	59%
+30	26	48	54%
Total	1808	3387	53%

Tabela 5.25 – Resultados do segundo dataset do método XGBoost

5.6 Resultado geral considerando as estatísticas reais de cada jogo

Vimos nas seções anteriores que os resultados obtidos, apesar de não serem catastróficos, estão longe do esperado para um método de previsão de excelência. A grande maioria dos resultados obtidos ficaram em torno de 60% de assertividade tendo alguns casos com resultados superiores e outros inferiores.

Nesta seção traremos os resultados nos quais aplicamos os mesmos métodos seguindo as mesmas metodologias das seções anteriores mas utilizando exclusivamente os dados reais. Isto é, ao invés de usarmos a média dos últimos 10 jogos, aqui foi utilizado a estatística real da partida.

Esses resultados servirão de base para podermos entender se a metodologia utilizada nesse estudo é adequada para previsão de jogos. Se os resultados previstos com os dados reais forem satisfatórios, o *gap* para ter uma boa previsão estará entre a média dos últimos 10 jogos e as estatísticas reais de cada partida. Todos esses pontos serão discutidos mais a fundo na sequência.

5.6.1 M5P

Na primeira parametrização do método M5P já podemos ver uma grande diferença na assertividade, mas ainda com resultados bastante tímidos apesar de uma melhora de 11 pontos percentuais. A tabela 5.26 mostra a assertividade por diferença de pontos.

Grupo	Previsões corretas	Jogos analisados	Assertividade
0-10	919	1625	57%
10-20	828	1115	74%
20-30	482	563	86%
+30	218	239	91%
Total	2447	3542	69%

Tabela 5.26 – Resultados do modelo M5P com dados reais - Parametrização 1

Na segunda parametrização já vemos uma melhora expressiva com resultados muito interessantes. Podemos ver na tabela 5.27 uma assertividade geral de 88%, um número muito bom para um modelo preditivo. Obviamente, os jogos previstos com maiores intervalo de pontos são os que apresentam os melhores índices de assertividade superando os 95% de acerto.

Grupo	Previsões corretas	Jogos analisados	Assertividade
0-10	1395	1736	80%
10-20	1081	1134	95%
20-30	463	476	97%
+30	191	196	97%
Total	3130	3542	88%

Tabela 5.27 – Resultados do modelo M5P com dados reais - Parametrização 2

A terceira parametrização novamente segue um valor geral próximo da segunda parametrização, mas nesse caso consegue superar em 2 pontos percentuais no resultado geral, chegando a 90% de assertividade. Na tabela 5.28 podemos ver que os jogos previstos com uma diferença de mais de 10 pontos chega a superar 98% de assertividade. Levando em conta que 50% dos jogos previstos possuem mais de 10 pontos de diferença, esse resultado não é nada desprezível e bastante interessante.

Grupo	Previsões corretas	Jogos analisados	Assertividade
0-10	1452	1767	82%
10-20	1102	1128	98%
20-30	479	482	99%
+30	165	165	100%
Total	3198	3542	90%

Tabela 5.28 – Resultados do modelo M5P com dados reais - Parametrização 3

5.6.2 Random Forests

Podemos ver na tabela 5.29 que a assertividade ficou bem mais alta do que utilizando a média dos últimos 10 jogos, com cerca de 23 pontos percentuais na média

geral (57% *vs* 80%). Porém ainda assim ficou abaixo do ótimo resultado que o método M5P apresentou de 90% de assertividade geral.

Podemos observar também que para os jogos previstos com diferença de pontos acima de 10 pontos o desempenho foi muito próximo ao resultado da parametrização 3 do método M5P. Enquanto que no M5P tinha um desempenho acima de 98% no *Random Forests* ficou acima de 95%, ou seja, muito próximo.

Grupo	**Previsões corretas**	**Jogos analisados**	**Assertividade**
0-10	1574	2233	70%
10-20	995	1051	95%
20-30	223	226	99%
+30	32	32	100%
Total	2824	3542	80%

Tabela 5.29 – Resultados do modelo Random forests com dados reais

5.6.3 XGBoost

O método *XGBoost* também ficou com um desempenho abaixo do resultado apresentado na parametrização 3 do método M5P. Porém, como podemos ver na tabela 5.30, apresentou uma leve melhora com relação ao método 5.29 de 2 pontos percentuais na assertividade geral.

Assim como o método *Random Forests*, o método *XGBoost* teve um bom desempenho na média geral mas também teve um desempenho excelente para jogos previstos com uma diferença acima de 10 pontos.

Grupo	**Previsões corretas**	**Jogos analisados**	**Assertividade**
0-10	1416	1999	71%
10-20	1085	1150	94%
20-30	322	324	99%
+30	69	69	100%
Total	2892	3542	82%

Tabela 5.30 – Resultados do modelo XGBoost com dados reais

5.7 Discussão de resultados

Os métodos analisados nesse estudo, obtiveram uma assertividade que variou bastante entre os grupos de pontos analisados. Entretanto, na média geral os resultados ficaram bastante próximos entre si. A maior assertividade geral encontrada foi a primeira parametrização do método M5P, mas com uma diferença marginal entre os demais métodos (58% vs 57%). Entre os intervalos de pontos tivemos um empate entre os métodos M5P e *Random Forests* com 68% de assertividade.

Na Figura 5.1, podemos ver um breve comparativo da assertividade geral dos métodos analisados.

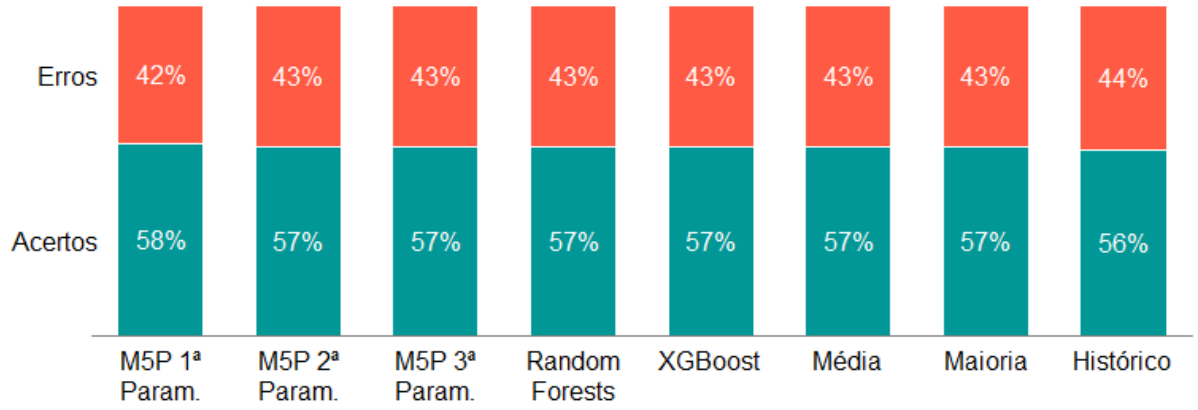

Figura 5.1 – Comparativo dos métodos analisados

Todos os métodos foram avaliados caso fossem feitas apostas nos jogos analisados. Para simplificar, foi avaliado o resultado financeiro obtido, caso tivesse sido apostado R$ 1,00 em cada um dos jogos previstos. Como era de se esperar, com essa assertividade inferior a 60%, em todos os casos o retorno era negativo, isto é, haveria prejuízo nessas circunstâncias. Mesmo simulando alguns grupos de jogos em que houveram uma maior assertividade, o resultado não foi muito satisfatório.

Para as equipes que os métodos apresentaram uma assertividade alta, as análises com as *odds* eram bastante interessantes, porém muito difícil de serem aplicadas na prática, uma vez que seria muito complicado descobrir qual equipe seria mais previsível dado que a cada temporada as equipes 'mais previsíveis' mudavam.

Os *datasets* reduzidos também mostraram não ter um bom desempenho geral. Pelo contrário, pudemos ver que o desempenho dos métodos pioraram. Muito provavelmente foram retirado estatísticas relevantes para previsão e acabou comprometendo o aprendizado dos algoritmos.

Talvez se testássemos exaustivamente diversos *datasets*, poderíamos encontrar algum *dataset* com poucas variáveis e com desempenho similar. Mas tudo indica que os ganhos seriam muito baixos.

A tentativa do comitê de escolha para tentar escolher 'democraticamente' um resultado final não foi bem. Adicionar outros métodos robustos poderiam ajudar principalmente o método baseado no histórico de partidas pois daria mais insumo mas também não mudaria drasticamente o resultado final.

A única análise que teve um desempenho excelente foi quando utilizamos os dados reais ao invés da média das últimas partidas. Isso indica que a previsão pela média

simples dos últimos jogos está distorcendo bastante a realidade e afetando o desempenho dos algoritmos preditivos.

Nas figuras abaixo podemos ver a diferença de assertividade entre os métodos quando utilizado os valores reais e quando utilizado a média dos últimos 10 jogos para poder prever o resultado.

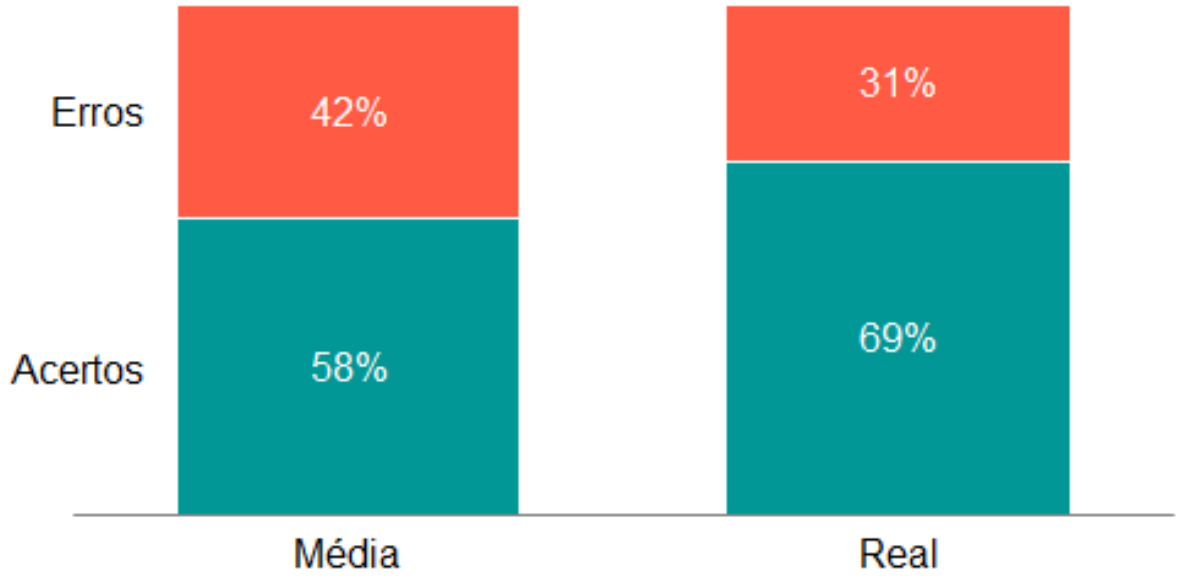

Figura 5.2 – Comparativo entre a média e as estatísticas reais - M5P 1ª parametrização

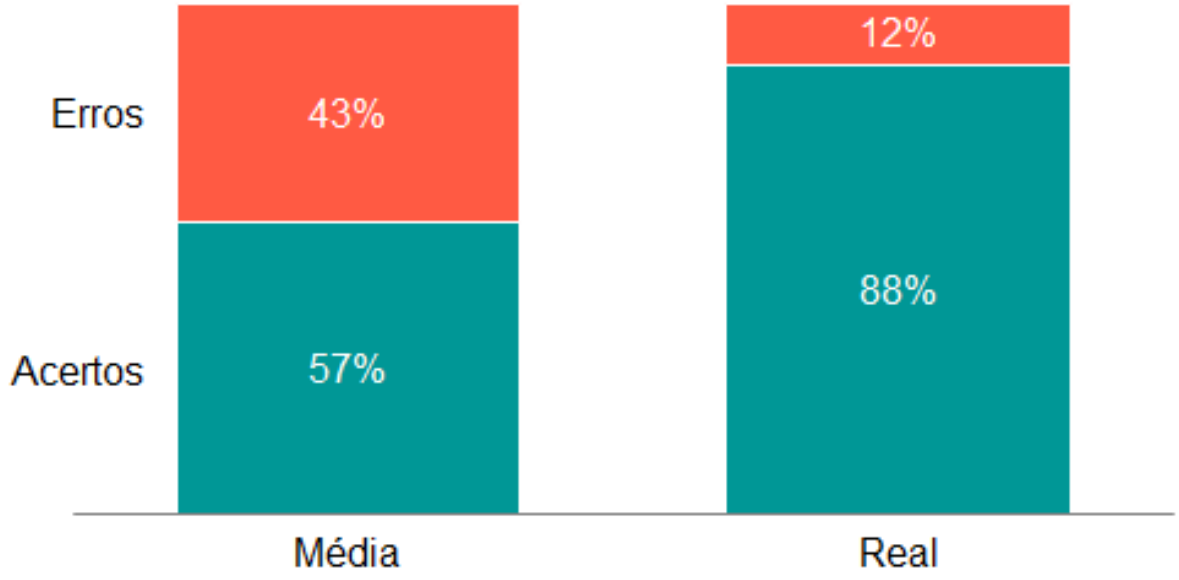

Figura 5.3 – Comparativo entre a média e as estatísticas reais - M5P 2ª parametrização

Podemos ver que quando utilizamos os dados reais, os valores da previsão ficaram superiores aos encontrados em outros estudos da literatura. Esse desempenho superior se dá por dois principais motivos.

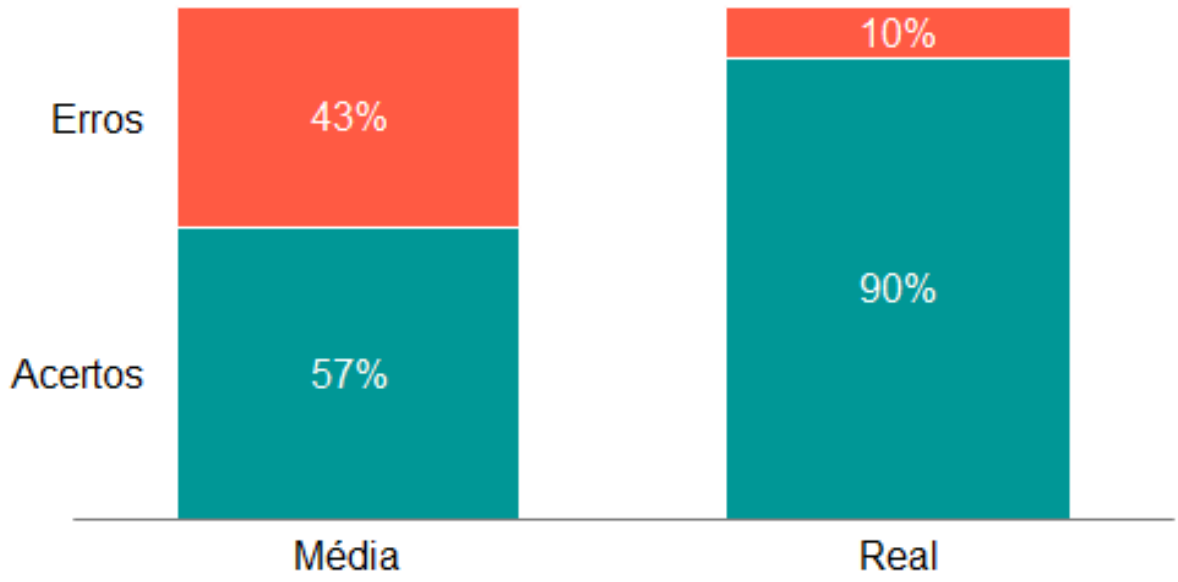

Figura 5.4 – Comparativo entre a média e as estatísticas reais - M5P 3ª parametrização

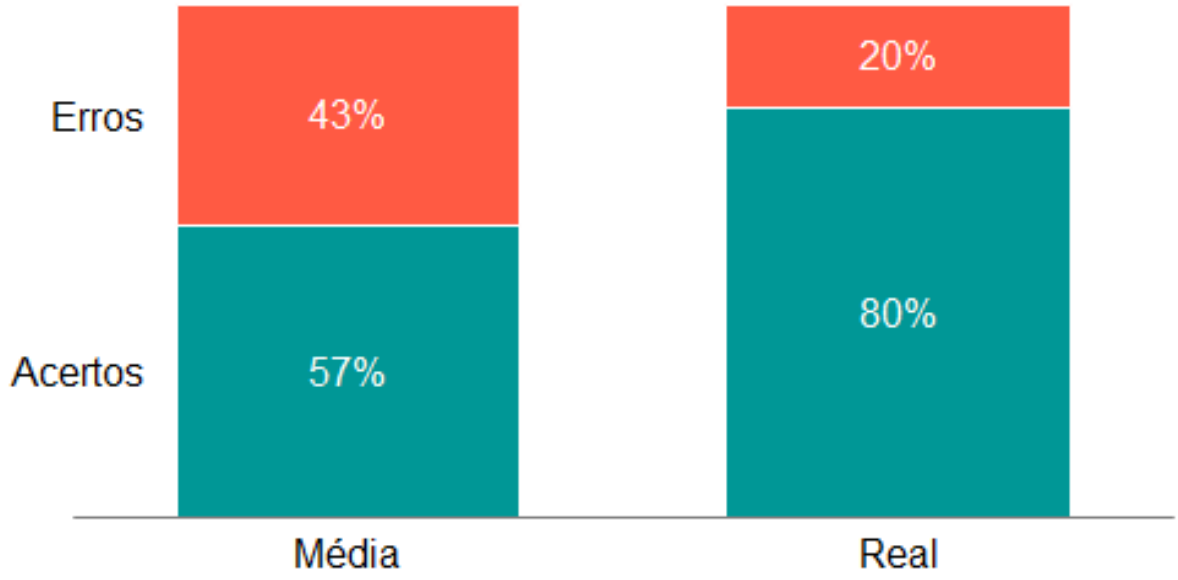

Figura 5.5 – Comparativo entre a média e as estatísticas reais - Random Forests

O primeiro deles é que aqui foram utilizados uma grande quantidade de estatísticas e muitas delas extremamente relevantes para o placar final. Alguns dos estudos utilizaram uma quantidade reduzida de estatísticas, dificultando o aprendizado dos métodos.

O segundo deles, que é o principal motivo, é que aqui fizemos a predição jogo a jogo. A grande maioria dos estudos treinava o método com uma amostra de partidas e aplicava nas demais partidas. No presente estudo, pegávamos uma amostra com as últimas 70 partidas de cada uma das equipes que iriam jogar e tentávamos prever a partida 71.

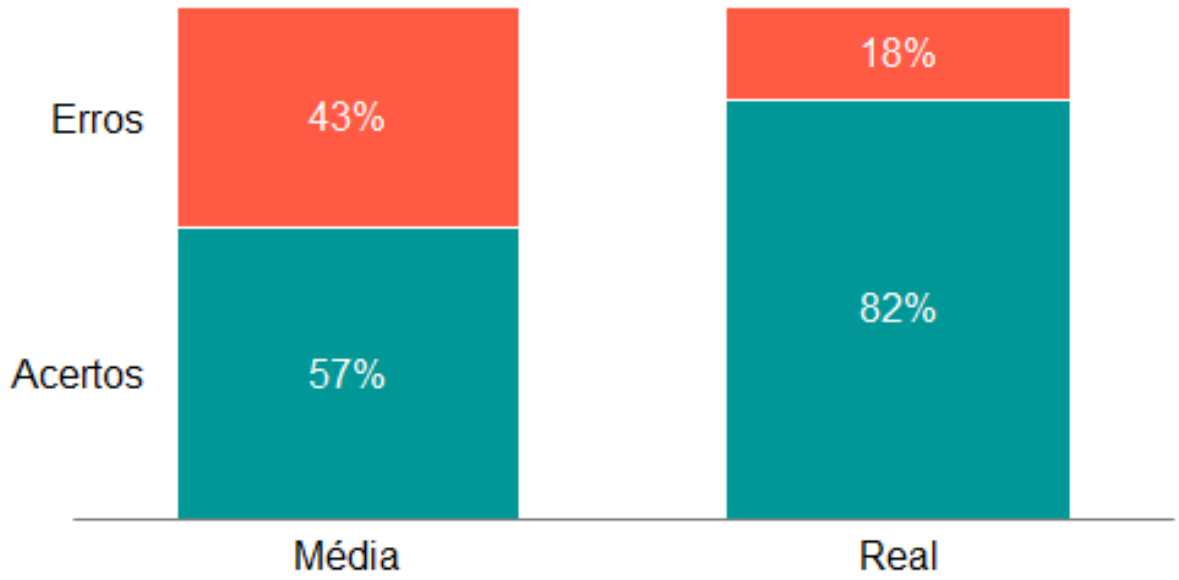

Figura 5.6 – Comparativo entre a média e as estatísticas reais - XGBoost

Depois pegávamos novamente as últimas 70 partidas a tentávamos prever a partida 72. Dessa forma os algoritmos de *machine learning* sempre tinham as últimas partidas para prever apenas 1 partida e não criar um método para tentar prever diversas maneiras. Isso é bem mais custoso operacionalmente falando, porém é muito mais assertivo pois as equipes mudam ao longo do campeonato por conta de contratações, lesões e ajustes técnicos e táticos.

Esse estudo mostrou uma maneira muito eficaz de prever as partidas utilizando alguns métodos sofisticados de *machine learning*. Os resultados mostraram que esses métodos são sim muito eficientes, porém, precisam de estatísticas relevantes e precisas sobre o possível desempenho das equipes durante a partida.

Vimos que há uma lacuna bastante grande de predição ao utilizar a média dos últimos 10 jogos nas estatísticas e as estatísticas reais de cada partida. Se a predição das estatísticas de cada partida for bem trabalhada, certamente sairá uma ferramenta extremamente robusta de predição.

Os resultados obtidos aqui nesse estudo fazem parte de uma primeira etapa de análise de um mundo bastante complexo e desafiador que é o basquete. Isto é, como já fora mencionado, os modelos de previsão de resultados ainda podem ser melhorados e ficarão como motivação para futuros aprimoramentos e avanços.

5.7.1 Aplicações

As aplicações dessa metodologia é bem ampla. Como mencionamos anteriormente, 2 principais aplicações são bastante relevantes mas uma em especial vale ser

mencionada com mais detalhes.

A aplicação que vale um destaque extra é o uso dessa metodologia pela comissão técnica de uma equipe de basquetebol. Os modelos poderiam ser treinados em alguns minutos ou em poucas horas, a depender da capacidade computacional, e testados prontamente em alguns segundos. Por exemplo, os treinos dos métodos poderiam ser feitos antes das partidas e, ao final do primeiro quarto, a equipe de inteligência poderia executar o modelo com as estatísticas reais do primeiro quarto em poucos segundos e entender quais mudanças impactariam de fato o resultado do jogo. O uso da metodologia seria uma ferramenta extra no apoio da tomada de decisão da comissão técnica em optar por uma eventual mudança de estratégia da partida. Ainda nesse exemplo, se a comissão técnica acredita que uma determinada substituição de jogadores pode melhorar o % de acerto de 3 pontos com uma leve piora nos rebotes defensivos, basta executar o método com as respectivas mudanças das estatísticas para saber o impacto no resultado do jogo (de possível derrota para uma vitória). Então, caso seja uma mudança positiva, essa substituição poderia ser feita.

Ainda nessa mesma abordagem, uma utilização clara das comissões técnicas, seria a utilização para o planejamento do treinamento da equipe. Ao analisar previamente uma partida, o uso dessa metodologia determinará qual estatística é determinante para a vitória da partida. Portanto, a melhora em determinada estatística pode fazer com que uma possível derrota se torne uma vitória.

6 Conclusões e perspectivas

Este trabalho de dissertação apresenta um estudo sobre o uso de métodos de aprendizado de máquina para a previsão de resultados de jogos de basquetebol.

O estudo conseguiu se aprofundar nas análises de previsão de resultados do basquete, principalmente trazendo alternativas para o que é encontrado na maioria dos trabalhos encontrados na bibliografia atual.

Os dados utilizados foram as estatísticas das temporadas de 2018-2019 até 2021-2022 da *NBA*. Em algumas análises utilizaram-se os dados reais da partida e em outros os dados foram trabalhados para um melhor seguimento das análises.

A metodologia adotada neste estudo permitiu que os métodos de *machine learning* fossem treinados com um conjunto de dados atualizado, incluindo informações das últimas 70 partidas para cada jogo previsto. Essa abordagem resultou em níveis de acurácia significativamente altos, alcançando até 90% na análise da 3ª parametrização do método M5P utilizando os dados da própria partida. Esses resultados são particularmente relevantes quando comparados aos estudos anteriores mencionados na revisão bibliográfica, que apresentaram acurácia em torno de 80%. Isso destaca a eficácia e o avanço proporcionados pela metodologia utilizada, proporcionando *insights* valiosos para análise e previsão no contexto esportivo.

As previsões que não utilizaram os dados da partida que se pretendia prever tiveram um resultado relativamente baixo, com cerca de 60% evidenciando a queda de acurácia quando comparado com métodos que utilizam a média dos últimos 10 jogos. Isso mostra que a média dos últimos 10 jogos, de maneira geral, não está evidenciando o real comportamento das partidas.

Com os resultados obtidos, é evidente que utilizar os dados reais para prever o resultado da partida leva a uma acurácia alta dos métodos de predição. Porém, se de fato quisermos prever uma partida antes mesmo dela acontecer, ou seja, sem ter os dados reais da partida, a acurácia é reduzida significativamente. Ficam de motivações para futuros estudos outras abordagens e possibilidades para melhorar a assertividade na predição dos jogos, como descrito na seção seguinte.

6.1 Perspectivas futuras

Era esperado que os resultados obtidos neste estudo fossem ligeiramente inferiores aos encontrados na revisão bibliográfica, uma vez que os estudos anteriores se baseavam em estatísticas reais do jogo, enquanto este estudo se baseia em estimativas das estatísticas da partida. Essa diferença foi comprovada nas análises comparativas com as estatísticas reais. No entanto, o aprofundamento contínuo deste estudo tem como objetivo aprimorar os resultados e torná-lo mais eficaz, com a possibilidade de se tornar lucrativo quando aplicado em contextos como sites de apostas.

Alguns pontos importantes devem ser levados em consideração no aprofundamento desse estudo.

6.1.1 Comparar com mais métodos

Foram escolhidos métodos bastante modernos e avançados com alto índice de assertividade, porém existem outros métodos tão bons quanto que merecem atenção em uma possível evolução desse estudo, tais como HFSVM e SVM. Talvez a melhora seja marginal, mas ainda sim pode ser bastante interessante.

6.1.2 Incluir outras variáveis

O basquete é um esporte com muitas estatísticas envolvidas e certamente outras variáveis podem ser incluídas nesse estudo a fim de enriquecer ainda mais as análises. Um exemplo disso seria a inclusão do fator de alguns jogadores chaves para cada time. Por exemplo, se um jogador de alto nível como LeBron James ou Stephen Curry joga ou deixa de jogar em um determinado jogo, certamente a dinâmica da partida muda e consequentemente poderia estar refletido na previsão da mesma.

6.1.3 Aplicação em outros esportes

Os métodos utilizados nesse estudo são bastante versáteis e uma reprodução desse estudo com outros esportes também seria bastante interessante.

6.1.4 Previsão das estatísticas das partidas

Provavelmente esse tópico é o mais relevante para o aprofundamento desse estudo. Como mencionado anteriormente, a lacuna apresentada entre a previsão utilizando a média dos últimos 10 jogos e as estatísticas reais é bastante grande. Buscar uma alternativa à média dos últimos 10 jogos que se aproxime bastante das estatísticas reais é fundamental para o sucesso da ferramenta de predição de resultados.

6.1.5 Uso de dados em tempo real

Todas as análises feitas aqui nesse estudo se baseavam em dados das partidas completas, seja a média ou os dados reais. Olhando pela perspectiva dos dados reais, um caminho possível a se desenvolver seria executar essas análises com os dados em tempo real. Por exemplo, ao final do primeiro quarto da partida, utilizar as estatísticas reais para tentar prever qual equipe iria sair vitoriosa da partida.

6.1.6 Comitê de escolha pelo histórico

Ao melhorar a previsão das estatísticas provavelmente teríamos alguns métodos de *machine learning* prevendo muito bem os resultados das partidas. Porém qual desses escolher? A tentativa de criar um comitê de escolha pelo histórico é uma boa ideia porém precisaria ser mais aprofundado para ter um desempenho superior.

Referências

Basketball Reference, 2022 (2022). Web page. `https://www.basketball-reference.com/`. Acessado: 2022-06-22.

Cao, C. (2012). Sports data mining technology used in basketball outcome prediction.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754.

Datageeks, 2022 (2022). Web page. `https://www.datageeks.com.br/xgboost/`. Acessado: 2022-07-10.

El pais, 2022 (2022). Web page. `encurtador.com.br/bgwQ5`. Acessado: 2022-06-27.

Fadi Thabtah1, L. Z. e. N. A. (2019). Nba game result prediction using feature analysis and machine learning. *Annals of Data Science*.

Huang, M.-L. and Lin, Y.-J. (2020). Regression tree model for predicting game scores for the golden state warriors in the national basketball association. *Symmetry*, 12(5).

Jain, S. and Kaur, H. (2017). Machine learning approaches to predict basketball game outcome. In *2017 3rd International Conference on Advances in Computing,Communication & Automation (ICACCA) (Fall)*, pages 1–7.

Lieder, N. (2018). Can machine-learning methods predict the outcome of an nba game?

Mercadante, L. A. (2021). *Basquetebol por números: do jogo livre ao alto rendimento*. CRV.

NBA Stuffer, 2022 (2022). Web page. `https://www.nbastuffer.com/analytics101/nba-teams-that-have-analytics-department/`. Acessado: 2022-06-27.

Odds Portal, 2022 (2022). Web page. `https://www.oddsportal.com/basketball/`. Acessado: 2022-07-09.

Oliver, D. (2004). *Basketball on paper: rules and tools for performance analysis*. Potomac books.

Pai, P.-F., ChangLiao, L.-H., and Lin, K.-P. (2017). Analyzing basketball games by a support vector machines with decision tree model. *Neural Computing and Applications*, 28.

Puranmalka, K. (2013). Modelling the nba to make better predictions.

Quinlan, J. (1992). Learning with continuous classes.

Soliman, G., El-Nabawy, A., Misbah, A., and Eldawlatly, S. (2017). Predicting all star player in the national basketball association using random forest. In *2017 Intelligent Systems Conference (IntelliSys)*, pages 706–713.

Wang, Y. and Witten, I. (1997). Induction of model trees for predicting continuous classes. *Induction of Model Trees for Predicting Continuous Classes*.