UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

BÁRBARA OLIVEIRA NEVES
Orientador: Prof. Dr. Anderson Almeida Ferreira

MINERAÇÃO DE DADOS APLICADA À PREDIÇÃO DE RESULTADOS DE JOGOS DE BASQUETE

Ouro Preto, MG
2022

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

BÁRBARA OLIVEIRA NEVES

MINERAÇÃO DE DADOS APLICADA À PREDIÇÃO DE RESULTADOS DE JOGOS DE BASQUETE

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Anderson Almeida Ferreira

Ouro Preto, MG
2022

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Bárbara Oliveira Neves

Mineração de Dados Aplicada à Predição de Resultados de Jogos de Basquete

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 17 de Junho de 2022.

Membros da banca

Anderson Almeida Ferreira (Orientador) - Doutor - Universidade Federal de Ouro Preto
Elton José da Silva (Examinador) - Mestre - Universidade Federal de Ouro Preto
Marcelo Luiz Silva (Examinador) - Mestre - Universidade Federal de Ouro Preto

Anderson Almeida Ferreira, Orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 17/06/2022.

Documento assinado eletronicamente por **Anderson Almeida Ferreira**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 18/06/2022, às 10:43, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0343246** e o código CRC **D6845322**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35400-000
Telefone: 3135591692 - www.ufop.br

Dedico este trabalho à minha mãe Rosy

pela confiança e pelo amor incondicional

e ao meu avô Rui (in memoriam) por me ensinar

que é sempre tempo de recomeçar.

Agradecimentos

Ao meu orientador Prof. Dr. Anderson, a quem tenho uma grande admiração e respeito. Agradeço por todos os ensinamentos, pela oportunidade de trabalharmos juntos nesta pesquisa. Mas, acima de tudo, gratidão pela paciência e pela disponibilidade.

À minha mãe, registro a minha eterna e imensa gratidão pela confiança e pelos investimentos na minha educação ao longos de todos esses anos. Hoje, podemos dizer em meio a tantos desafios que nós conseguimos, juntas, finalizar mais uma etapa. Embora a responsabilidade por concluir esse trabalho seja minha, ele não teria se concretizado sem o seu amor incondicional.

Ao meu padrasto Agnelo pelo carinho, respeito e apoio durante toda a minha graduação. Muito obrigada!

A minha eterna gratidão aos meus amigos Maycon, Mateus, Douglas, Evair e Luiza pela parceria, pelos ensinamentos e pela amizade. Sem vocês eu não teria conseguido concluir essa graduação. Amo vocês!

À minha segunda família República Menina dos Olhos: Luiza, Gabrielle, Maria Elisa e a todas as ex-alunas que contribuíram com a existência da república durante todos esses anos. Muito obrigada pelas longas conversas, pela empatia e pela amizade!

"Porque se chamava moço
Também se chamava estrada
Viagem de ventania
Nem lembra se olhou pra trás
Ao primeiro passo, aço, aço…
Porque se chamava homem
Também se chamavam sonhos
E sonhos não envelhecem"
(Milton Nascimento — Clube da esquina II)

Resumo

A mineração de dados permite extrair informações de grandes volumes de dados, em que não seria possível utilizar técnicas tradicionais. Uma das aplicações de sucesso da mineração concerne a análise de partidas esportivas para predição de resultados, com especial interesse do mercado de apostas nessa aplicação. A liga nacional de basquete dos Estados Unidos, a NBA, foi a primeira liga de esporte profissional estadunidense a ter parceria com duas companhias de apostas. Apesar da agitação do mercado de apostas desde a oficialização dessa parceria, nas últimas temporadas a NBA sofreu com a redução dos investimentos e com a alteração na dinâmica dos jogos devido a pandemia do coronavírus. Este trabalho se propõe a aplicar técnicas de mineração de dados para a construção de um modelo capaz de predizer os resultados dos jogos da NBA e analisar os impactos da pandemia na predição de resultado desses jogos. Para isso, são utilizados dados estatísticos da NBA referentes às últimas nove temporadas. Após o treinamento dos modelos e análise dos resultados obtidos, percebeu-se que houve alterações nos padrões de vitórias e derrotas da NBA durante a pandemia. Essas alterações foram percebidas devido à perda de generalização do modelo para o experimento que Utilizou como conjunto de teste a temporada 2019-2020, disputada durante a pandemia, e como conjunto de treino as temporadas de 2012-2018, disputadas antes da pandemia. A perda de generalização maior foi observada para o algoritmo classificador Random Forest que obteve 73,23% e 61,75% de acurácia no treino e no teste, respectivamente. O algoritmo Naïve Bayes obteve o melhor resultado em todos os experimentos.

Palavras-chave: mineração de dados. predição de partidas esportivas. NBA.

Abstract

Data mining allows extracting information from large volumes of data, where it would not be possible to use traditional techniques. One of the successful applications of mining concerns the analysis of sports matches for predicting results, with special interest of the betting market in this application. The United States' national basketball league, the NBA, was the first American professional sports league to partner with two betting companies. Despite the upheaval in the betting market since this partnership was made official, in recent seasons the NBA has suffered from reduced investments and the change in game dynamics due to the coronavirus pandemic. This work proposes to apply data mining techniques to build a model capable of predicting the results of NBA games and analyzing the impacts of the pandemic on the prediction of the outcome of these games. For this, statistical data from the NBA for the last nine seasons are used. After training the models and analyzing the results obtained, it was noticed that there were changes in the winning and losing patterns of the NBA during the pandemic. These changes were noticed due to the loss of generalization of the model for the experiment that used the 2019-2020 season, played during the pandemic, as a test set, and the 2012-2018 seasons, played before the pandemic, as a training set. The greatest loss of generalization was observed for the Random Forest classifier algorithm, which obtained 73.23% and 61.75% of accuracy in the training and in the test, respectively. The Naïve Bayes algorithm obtained the best result in all experiments.

Keywords: datamining, prediction of sport matches, NBA.

Lista de Ilustrações

Lista de Tabelas

Lista de Abreviaturas e Siglas

CSV	Comma Separated Values
EFG	Effective Field Goal
FN	Falso Negativo
FP	Falso Positivo
FTR	Free Throw Rate
KDD	Knowledge Discovery in Databases
NBA	National Basketball Association
ORB	Offensive Rebound
RBC	Raciocínio Baseado em Casos
RFE	Recursive Feature Elimination
ROI	Return On Investment
SVM	Support Vector Machine
TN	Verdadeiro Negativo
TOV	Turnover
TP	Verdadeiro Positivo
TS	True Shooting
XLSX	Microsoft Excel Open XML Spreadsheet

Sumário

1 Introdução

O avanço tecnológico na coleta e armazenamento de dados permitiu às organizações acumularem uma enorme quantidade de dados. Entretanto, extrair informação desse grande volume de dados tem se mostrado uma tarefa desafiadora. A enorme quantidade de dados e a natureza não tradicional dos dados tornam, frequentemente, técnicas e ferramentas tradicionais de extração da informação obsoletas, surgindo a necessidade de desenvolvimento de novos métodos (TAN; STEINBACH; KUMAR, 2009).

De acordo com Tan, Steinbach e Kumar (2009), um dos campos de pesquisa da mineração de dados é o aprendizado de máquina. O aprendizado de máquina tem como finalidade criar sistemas computacionais capazes de aprender de forma automática e reconhecer complexos padrões de dados. Um problema típico de aprendizado de máquina, por exemplo, é codificar um sistema computacional para que seja possível o reconhecimento de códigos postais escritos à mão no correio, após aprender a partir de uma amostra de exemplos.

Segundo Cao (2012), antes do surgimento da mineração de dados, as organizações esportivas dependiam apenas da experiência de treinadores, jogadores e olheiros para a coleta de dados e extração da informação. Acreditava-se que esses especialistas eram capazes de converter todo registro histórico de jogos e jogadores em conhecimento útil. Entretanto, houve a necessidade de um melhor aproveitamento desses dados, sobretudo, devido aos investimentos feitos por grandes empresas e a consequente cobrança por melhores resultados.

Atualmente, a mineração de dados tem sido usada com sucesso em muitos esportes, tais como, beisebol, futebol de críquete, hóquei e basquete. Uma das aplicações mais famosas é mostrada no filme Moneyball[1] que conta a história de um técnico de baseball, o famoso Billy Beane que rompeu com os métodos convencionais e utilizou técnicas de mineração de dados para construir um modelo de avaliação de seus jogadores, através do percentual de vezes que os mesmos chegavam à base. E, por fim, foram capazes de formar times competitivos e baratos com a capacidade de lutar contra o famoso *New York Yankees* (CAO, 2012).

Na National Basketball Association (NBA), liga nacional de basquete dos Estados Unidos, devido a uma enorme quantidade e detalhamento de dados, é possível fazer previsões. Hoje, técnicos e auxiliares dispõem de ferramentas que, baseadas em *scout* (informações relacionadas a aspectos gerais envolvidos em um jogo), são capazes de fazer predições de resultados.

Apesar da existência de todas essas ferramentas, durante a pandemia do coronavírus diversas foram as alterações nas dinâmicas dos jogos da NBA que influenciaram na predição dos resultados dos jogos. O *Toronto Raptors*, por exemplo, time canadense que disputa a NBA jogou mais de 100 jogos fora de casa. De acordo com a nba.com, durante as temporadas de 2019-2020

[1] Mais informações em <https://www.imdb.com/title/tt1210166/>

e 2020-2021, o governo canadense proibiu que os jogos da NBA fossem disputados na cidade de Toronto para conter o avanço do coronavírus e, portanto, a equipe dos *Raptors* passou a jogar na cidade de Tampa na Flórida. Um dado que reflete essa mudança nos locais das partidas de várias equipes é porcentagem de vitórias dos times mandantes.Em média 58% do times ganhavam a partida quando jogavam em sua própria arena. Entretanto, durante a pandemia essa porcentagem caiu para 52%, principalmente, devido ao número de times que não puderam jogar em suas próprias arenas

Nesse contexto, este trabalho visa avaliar os impactos das alterações nas dinâmicas dos jogos da NBA, disputados durante a pandemia, na predição dos resultados das partida através de um estudo comparativo com temporadas anteriores e à partir da utilização de algoritmos de aprendizado de máquina.

1.1 Justificativa

A mineração de dados é um campo de estudo em constante crescimento, principalmente porque a extração do conhecimento em um grande volume de dados não é possível a partir de técnicas tradicionais de análise de dados.

Em 2018, a NBA tornou-se a primeira liga de esporte profissional dos Estados Unidos a ter uma parceria com duas companhias de apostas esportivas que passaram a ser distribuidoras oficiais dos dados fornecidos pela liga de basquete. Construir um modelo capaz de reconhecer padrões de performance dos times frente aos adversários e de prever os resultados dos jogos da NBA para fins de apostas esportivas se tornou extremamente atraente devido ao crescimento do mercado de apostas após a concretização de tal parceria.

Além de ter se tornado uma ferramenta interessante para o mercado de apostas, muitos entusiastas e especialistas do basquetebol utilizam as ferramentas e as técnicas de mineração de dados para medir o desempenho dos jogadores e ajustar as estratégias de jogo da equipe.

Entretanto, durante as temporadas de 2019-2020 e 2020-2021 a NBA precisou realizar diversas adaptações nas dinâmicas dos jogos com o intuito de conter o avanço do coronavírus. Dentre essas adaptações estão a diminuição na capacidade das arenas, a remarcação de jogos e a redução do número de partidas disputadas na temporada regular. Todas essas mudanças alteraram estatísticas muito conhecidas como , por exemplo, a probabilidade de o time mandante da partida vencer o jogo.

Avaliar esses dados e seus impactos na predição dos resultados das partidas através de um estudo comparativo com temporadas anteriores, é uma oportunidade de descobrir novas variáveis e , também, validar fatores de influência já conhecidos na predição de resultados das partidas.

Apesar do escopo deste projeto está limitado aos jogos da NBA, o modelo também pode ser aplicado na predição de resultados de partidas de outros esportes. Entretanto, pode ser

necessário a realização de algumas alterações do modelo criado de acordo com as características do esporte escolhido.

1.2 Objetivos

Nesta seção, são apresentados o objetivo geral e os objetivos específicos propostos neste trabalho.

1.2.1 Geral

Comparar algoritmos de aprendizado de máquina de modo a avaliar a performance dos mesmos na previsão de resultados de jogos da NBA, considerando as temporadas de 2012-2021. As temporadas anteriores à 2012 não foram selecionadas para assegurar que tendências atuais como, por exemplo, o aumento do número de arremessos de três pontos, não tenham impacto na previsão do modelo.

1.2.2 Específicos

- Estudar e descrever aplicações de técnicas de mineração de dados na área esportiva;
- Avaliar se a alteração no atributo porcentagem de vitórias em casa impacta na predição de resultados de jogos da NBA;
- Avaliar se as alterações nas dinâmicas dos jogos disputados durante a pandemia do coronavírus impactam na predição dos resultados de jogos da NBA;

1.3 Contribuições

São contribuições deste trabalho:

- Uma análise atualizada de algumas técnicas de mineração de dados disponíveis e os trabalhos que as utilizaram.
- Um método de trabalho que pode ser usado como apoio à decisão por treinadores e gestores esportivos.
- Uma análise dos impactos da pandemia do coronavírus na predição de resultados da NBA através de um estudo comparativo com temporadas anteriores.

1.4 Organização do Trabalho

O restante deste trabalho está estruturado da seguinte forma: Capítulo 2 apresenta a fundamentação teórica relacionada à área de Mineração de Dados e às estatística de jogos da NBA. Além disto, são também discutidos alguns trabalhos relacionados selecionados na literatura ; Capítulo 3 descreve o conjunto de dados utilizados; Capítulo 4 apresenta a método de trabalho proposto; Capítulo 5 descreve os experimentos realizados e os resultados obtidos à partir desses experimentos; Capítulo 6 conclui o trabalho e apresenta propostas de trabalhos futuros.

2 Revisão Bibliográfica

Este capítulo descreve o referencial teórico, com destaque para autores e conceitos que fundamentam a investigação sobre o objeto de pesquisa, que diz respeito à avaliação de algoritmos de aprendizado de máquina na previsão de resultados de jogos, no contexto dos jogos de basquete da NBA, descrevendo também a NBA e seus principais atributos. Para isso, apresentam-se as bases teóricas procedentes do campo da Ciência da Computação no que se refere às tarefas e técnicas de mineração de dados. São apontados os fundamentos teóricos sobre aquisição e processamento de dados, bem como os aspectos específicos concernentes à avaliação das tarefas e técnicas de mineração.

2.1 Fundamentação Teórica

2.1.1 Processo de descoberta de conhecimento

O avanço da tecnologia tem gerado uma enorme quantidade de dados, em que a capacidade de coletar e armazenar tais dados não é equivalente à capacidade de analisar e extrair conhecimento destes. É nesse cenário de superabundância de dados que surge a Mineração de Dados para encontrar anomalias, padrões e correlações em grandes conjuntos de dados e auxiliar na descoberta do conhecimento.

Segundo Castro e Ferrari (2016), a mineração de dados é parte de um processo mais extenso conhecido como Descoberta de Conhecimento em Bases de Dados/*Knowledge Discovery in Databases* (KDD), que objetiva explorar grandes quantidades de dados de maneira automatizada reconhecendo padrões existentes através da modelagem. Conforme mostra a Figura 2.1, o KDD pode ser dividido em quatro em etapas:

- **Aquisição de Dados**: Coleta e seleção dos dados de maneira a permitir a recuperação da informação;
- **Preparação ou pré-processamento de dados**: São estágios anteriores à mineração de dados que tem como finalidade preparar os dados para a análise. Tratamento de dados inconsistentes e ruídos, integração de dados obtidos de diferentes fontes, seleção e redução de dados, que consiste na definição daqueles dados relevantes para a análise, e transformação que consiste na estruturação dos dados que posteriormente serão utilizados na Mineração de Dados;
- **Mineração de Dados**: Aplicação de técnicas estatística, inteligência artificial e machine learning capazes de extrair conhecimento;

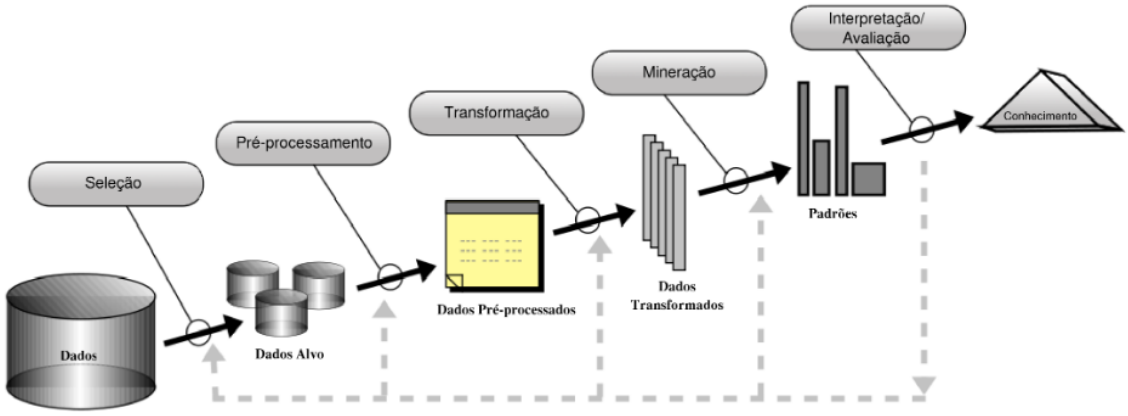


Figura 2.1 – Etapas do processo KDD

Fonte: Fayyad, Piatetsky-Shapiro e Smyth (1996).

- **Interpretação e avaliação**: Estágio de avaliação dos resultados obtidos com a Mineração de dados com a finalidade de avaliar conhecimentos úteis e não triviais.

2.1.1.1 Aquisição de Dados

Tan, Steinbach e Kumar (2009) afirmam que um conjunto de dados pode ser visto como uma coleção de objetos de dados que são descritos por uma série de atributos. Estes objetos também podem ser chamados de registros ou entidades e podem vir de uma única fonte de dados ou de várias fontes. Neste último caso, em que os dados são provenientes de várias fontes é necessário realizar a *integração* dos dados que será detalhada mais à frente. Um atributo é uma propriedade ou uma característica de um objeto que pode variar. Por exemplo, a cor dos olhos varia de pessoa para pessoa ou a temperatura é um conjunto ilimitado de valores. Um atributo pode ter valores do tipo quantitativos (Discretos e Contínuos) ou qualitativos (Nominais e Ordinais), em que:

- **Discretos**: assumem apenas valores inteiros. Ex.: número de irmãos, número de passageiros;
- **Contínuos**: assume qualquer valor no intervalo dos números reais. Ex.: peso, altura;
- **Nominais**: quando as categorias não possuem uma ordem natural. Ex.: cores;
- **Ordinais**: quando as categorias podem ser ordenadas. Ex.: classe social (baixa, média, alta).

2.1.1.2 Pré-Processamento de Dados

Schmitt (2005) relata que uma etapa importante para a extração da informação de maneira eficiente e eficaz é o pré-processamento de dados. O pré-processamento pode envolver limpeza de

dados, integração, redução, transformação e discretização. O objetivo do pré-processamento de dados é, portanto, preparar os dados brutos para que sejam analisados sem erros de incompletude, inconsistências e ruídos.

Limpeza de Dados

De acordo com Silva, Peres e Boscarioli (2017) a limpeza de dados (também conhecidos como *data cleaning*) tem a finalidade de resolver dois problemas provenientes da obtenção de dados imprecisos: a presença de valores ausentes (*missing values*) e a presença de valores ruidosos (*noise values*) ou inconsistências.

- **Valores Ausentes:** São identificados quando um conjunto de dados não apresenta valores para determinados exemplares ou quando um conjunto de dados não apresenta valores para um atributo de interesse. Algumas soluções para o problema dos valores ausentes envolvem a remoção do exemplar em que ocorre a falta de valor, preenchimento manual dos valores se houver a possibilidade de fazê-lo ou ainda preenchimento automático com um valor que pode ser obtido através da média ou mediana dos demais valores presentes no atributo em questão, ou inferidos a partir dos demais.
- **Valores Aberrantes e Ruidosos:** É identificado em valores consideravelmente diferentes da maioria dos outros valores da base dados, comumente chamados *outliers* ou valores aberrantes. Usualmente, são identificados quando já se conhece o valor esperado para o atributo ou a distribuição esperada para os valores de um atributo. Por exemplo, quando o valor de um atributo naturalmente deveria ser positivo, porém encontra-se negativo na base de dados. Uma possibilidade para tratar esses ruídos é a suavização dos dados através de uma função de regressão que pode ser linear (com uma variável independente) ou múltipla(com variáveis independente). Outra solução possível é a clusterização em que os dados semelhantes são agrupados em um *cluster* de maneira que os *outliers* sejam tratados separadamente ou deixados de fora dos *clusters*.

Integração de Dados

Nos relatos de Silva, Peres e Boscarioli (2017), a integração de dados tem o objetivo de unificar diferentes fontes de dados em apenas uma base de dados. Os principais desafios para a aplicação de integração de dados são: a presença de valores inconsistentes e a presença de valores redundantes.

- **Valores Inconsistentes:** É observado quando, para um mesmo atributo, ou para atributos equivalentes, são encontrados valores com discrepâncias em termos de tipo ou de domínio. Por exemplo, em uma pesquisa de satisfação do cliente pede-se para atribuir um conceito de qualidade, e o cliente atribui uma nota numérica, enquanto outro atribui um conceito por

meio de letra. Normalmente, a solução para este tipo de inconsistência inclui a remoção dos dados e ou a correção manual e o uso dos metadados;

- **Redundância de Dados:** É observada quando dois registros possuem nomenclaturas diferentes para registros equivalentes. Normalmente, neste caso, é realizada a redução do conjunto que pode ocorrer tanto de forma horizontal (eliminação de exemplares) quanto na forma vertical (eliminação de atributos).

Redução dos Dados

Um conjunto de dados pode ter um grande número de recursos. Existem inúmeros benefícios na redução do número de atributos e valores no processo de mineração de dados. A redução da dimensionalidade pode eliminar recursos irrelevantes e eliminar possíveis ruídos, melhorando o funcionamento dos algoritmos de mineração de dados. Outro benefício, é que um menor número de atributos pode levar a uma maior compreensão do modelo de mineração de dados criado (TAN; STEINBACH; KUMAR, 2009).

Apesar de o conhecimento do domínio permitir a eliminação de atributos irrelevantes e ruidosos, selecionar o melhor subconjunto de recursos requer uma análise sistemática. Para essa seleção de subconjuntos deve ser realizada a aplicação de técnicas e conceitos para que se possa encontrar um conjunto mínimo de atributos de forma que a distribuição de probabilidade resultante das classes de dados seja o mais próximo possível da distribuição original obtida usando todos os atributos. Isso pode ser alcançado através de abordagens em que o próprio algoritmo de mineração vai escolher qual atributo será utilizado e qual será ignorado ou ainda através do método de pesagem de características nas quais as características mais importantes recebem um peso maior, podendo ser realizado com base no conhecimento de domínio relativo às características ou de forma automática (máquinas de vetor de suporte) (HAN; PEI; KAMBER, 2011).

Transformação dos Dados

Segundo Castro e Ferrari (2016), um problema comum encontrado em base de dados é a não uniformidade dos atributos. Isso significa que a base de dados pode ser formada tanto por atributos categóricos quanto por atributos numéricos. Esse tipo de característica impacta o funcionamento dos algoritmos de mineração de dados e, portanto, precisa ser tratada.

A transformação de dados tem como finalidade modificar esses dados de modo a adequá-los ao processo de mineração de dados. Por exemplo, a integração de diferentes bases de dados pode ocasionar em dados com diferentes unidades de medida como centímetros ou metros. Neste caso, é usual padronizar os dados em uma mesma unidade de medida. Essa padronização pode envolver a normalização ou discretização desses dados.

Métricas de Avaliação

O processo de avaliação é realizado após o processo de treinamento com o objetivo de verificar qual algoritmo possui melhor acuracidade na classificação do vencedor de cada partida.

De acordo com (HAN; PEI; KAMBER, 2011), a avaliação de um classificador fundamenta-se no número de exemplos de teste corretamente e incorretamente previstos. A matriz de confusão possibilita uma clara visualização destes indicadores, sendo uma ferramenta interessante na análise da qualidade do classificador na identificação de exemplos de diferentes classes. As colunas representam as classes de previsão e as linhas as classes atuais dos dados, conforme a Tabela 2.1.

Tabela 2.1 – Matriz de Confusão

		Classe prevista	
		0	1
Classe correta	0	Verdadeiro Negativo	Falso Positivo
	1	Falso Negativo	Verdadeiro Positivo

Fonte: da autora, adaptado de Tan, Steinbach e Kumar (2009).

Quando os dados são divididos em apenas duas classes, é define-se uma como “positiva” e a outra como “negativa”. Desse modo, as entradas da matriz de confusão são definidas como:

- Verdadeiros positivos (*true positives* – TP) refere-se ao número de exemplos da classe “positiva” corretamente previstos como classe “positiva”;
- Falsos positivos (*false positives* – FP) refere-se ao número de exemplos da classe “negativa” incorretamente previstos como classe “positiva”;
- Verdadeiros negativos (*true negatives* – TN) representa o número de exemplos da classe “negativa” corretamente previstos como classe “negativa”;
- Falsos negativos (*false negatives* - FN) representa o número de exemplos da classe “positiva” incorretamente previstos como classe “negativa”;

Outras métricas podem ser definidas usando a matriz de confusão para avaliar o desempenho de classificadores como a Precisão, a Taxa de erro, a Revocação, a Especificidade, a exatidão e a Medida F. A precisão é a porcentagem de casos corretamente classificados.A taxa de erro é basicamente a porcentagem de casos incorretamente classificados. Em problemas em que o conjunto de dados reflete majoritariamente a classe negativa, deve-se usar também outras formas de medição como a Sensibilidade e a Especificidade. A sensibilidade é a proporção de exemplos positivos que são corretamente identificados e a especificidade a proporção de exemplos negativos que são corretamente identificados. O percentual de positivos corretamente previstos sobre o total

de positivos previstos é a Exatidão. E por fim, a medida F é a média harmônica entre a exatidão e a revocação (LUNELLI, 2020). A Tabela 2.2 mostra como essas métricas são calculadas em um problema de classificação.

Tabela 2.2 – Métricas para problemas de classificação

Medida	**Fórmula**
acurácia	$\frac{TP+TN}{TP+TN+FP+FN}$
taxa de erro	$\frac{FP+FN}{TP+TN+FP+FN}$
revocação	$\frac{TP}{TP+FN}$
especificidade	$\frac{TN}{TN+FP}$
exatidão	$\frac{TP}{TP+FP}$
medida F	$\frac{2 * \text{exatidão} * \text{recall}}{\text{exatidão} + \text{recall}}$

2.1.1.3 Tarefas de Mineração de Dados

As tarefas de mineração de dados normalmente são divididas em duas categorias principais: tarefas preditivas e tarefas descritivas. As tarefas preditivas inferem um modelo a partir do conjunto de dados disponível, que é útil na previsão de valores desconhecidos ou futuros de outro conjunto de dados de interesse. As tarefas descritivas geralmente encontram dados que descrevem padrões e apresentam informações novas e significativas do conjunto de dados disponível.

- **Análise descritiva de dados:** etapa inicial do processo de mineração de dados em que usa ferramentas capazes de medir, explorar e descrever características inerentes aos dados. A análise descritiva de dados consiste em sumarizar e compreender os objetos da base e seus atributos como, por exemplo, qual o salário médio dos professores universitários no Brasil;
- **Análise preditiva:** construção e utilização de um modelo capaz de classificar um objeto não rotulado ou para estimar um atributo específico de um determinado objeto. Classificação e regressão formam, portanto, os dois principais problemas da análise preditiva. Para exemplificar uma situação na qual a utilização da análise preditiva pode ser adequada, considere um contexto de um banco. A classificação de objetos pode ser útil para classificar um indivíduo como potencial devedor e a estimação para determinar qual valor deve ser

concedido como crédito a fim de evitar o não pagamento do empréstimo concedido pelo banco;

- **Agrupamento de dados:** tarefa descritiva que consiste na análise de conjuntos de dados em que há apenas as descrições dos dados. A informação sobre a classe à qual o objeto pertence não é relevante no agrupamento de dados. O objetivo é oferecer, através das similaridades dos objetos, um modelo de agrupamento ou perfis para grupos de dados. Para exemplificar a tarefa de agrupamento, considere um restaurante que recebe clientes com diferentes perfis e que a disposição das pessoas no ambiente seja feita através de um algoritmo que utiliza algumas informações prévias fornecidas pelos clientes. É interessante, portanto, que clientes com perfis distintos fiquem em ambientes distintos. Por exemplo, clientes jovens podem preferir um ambiente mais agitado, enquanto famílias com crianças podem preferir um local perto do playground. Para isso, poderia ser utilizado um algoritmo de agrupamento para a resolução desta tarefa;

- **Obtenção de padrões frequentes:** tarefa descritiva que busca por ocorrências frequentes e simultâneas entre elementos de um contexto. Nesta tarefa é comum identificar padrões triviais, entretanto, é esperado que padrões inesperados sejam descobertos. Ainda no contexto do restaurante, uma análise feita sobre a base de dados do restaurante (receita de cada prato) pode revelar que todo prato em que se usa alho, também se usa cebola. Mas é possível também que se descubra que "quanto menor o teor de álcool na cerveja usada em molhos, menor a necessidade do uso de açúcar", podendo representar um novo conhecimento.

2.1.1.4 Paradigmas de Aprendizado de Máquina

De acordo com Monard e Baranauskas (2003), alguns paradigmas de Aprendizado de Máquina estão sendo estudados constantemente, tais como: paradigma simbólico, estatístico/probabilístico, instance-based, conexionista e evolucionista.

- **Paradigma Simbólico:** Os sistemas de aprendizado simbólico procuram aprender a partir de construções simbólicas de um conceito a partir da análise de exemplos e contra-exemplos desse conceito. As representações simbólicas estão particularmente na forma de expressão lógica, árvore de decisão, regras ou rede semântica.

- **Paradigma Estatístico:** Diversos métodos de classificação criados por pesquisadores em estatística são semelhantes aos métodos posteriormente desenvolvidos pelos pesquisadores em aprendizado de máquina. A ideia geral é utilizar modelos estatísticos para encontrar uma boa aproximação de um conceito induzido. Um dos métodos estatísticos é o de aprendizado Bayesiano, em que utiliza-se um modelo probabilístico baseado no conhecimento prévio do problema de tal forma que combinado com os exemplos de treinamento torna-se possível determinar a probabilidade final de uma hipótese.

- **Baseado em Exemplos:** Uma forma de classificar um exemplo é lembrar de um outro cuja classe é conhecida e assumir que o novo exemplo terá a mesma classe. Esse tipo de aprendizado é conhecido como *lazy*. Esse tipo de sistema armazena os exemplos na memória para classificar novos exemplos. Portanto, é muito importante que apenas os exemplos (casos) de treinamento mais representativos sejam memorizados. *Nearest Neighbours* e Raciocínio Baseado em Casos (RBC) são os algoritmos mais conhecidos neste paradigma.

- **Conexionistas:** Redes Neurais são técnicas computacionais que apresentam um modelo matemático inspirado na estrutura neural de um ser inteligente. A representação da Rede Neural possui unidades extremamente interconectadas e, por essa razão, o nome conexionismo é usado para descrever a área de estudo. Alguns pesquisadores consideram redes neurais como métodos estatísticos paramétricos já que treinar uma rede neural normalmente é encontrar valores adequados para pesos (parâmetros) e viés pré-determinado.

2.1.2 National Basketball Association

A Associação Nacional de Basquetebol (National Basketball Association) foi fundada em 1946 na cidade de Nova York. A NBA só adotou o nome que possui hoje em 1949. Trata-se de uma liga profissional composta por 29 times dos Estados Unidos e um do Canadá. A liga é considerada a principal liga de basketball do mundo, e seus jogadores são os atletas mais bem pagos dentre todas as modalidades esportivas do mundo. Estas e outras informações são encontradas no site oficial da NBA[1].

É importante entender que a NBA possui duas conferências: Leste e Oeste, com 15 times cada, que se enfrentam para definir as 16 equipes que avançam para o *playoff*. Cada uma das conferências tem 3 divisões e cada uma das divisões possuem 5 times.

A temporada regular da NBA começa em meados de outubro e segue até meados de abril do ano seguinte em que cada time joga o total de 82 jogos. Os times da mesma divisão se enfrentam sempre 4 vezes. Entretanto, os times de outras divisões podem se enfrentar 3 ou 4 vezes. Essa variação no número de jogos se deve a utilização de um sistema em que a definição do número de jogos entre times de outras divisões irá depender de quantas vezes esses times se enfrentaram em temporadas passadas.

Os *playoffs* ou mata-mata começam em abril e terminam em maio. São séries compostas por 8 times de cada conferência em que primeiro colocado de cada conferência enfrenta o último, o segundo o penúltimo, o terceiro o antepenúltimo e assim sucessivamente.

As finais da NBA são realizadas em junho e é disputada pelo campeão da Conferência Leste *versus* o campeão da Conferência Oeste.

[1] Disponível em: <https://careers.nba.com/history/>

2.1.3 Estatísticas de jogos de basquete

Nesta seção, serão apresentadas as estatísticas registradas em partidas da NBA e suas respectivas descrições, disponíveis em nba.com, e que serão utilizadas para a aplicação dos algoritmos.

As informações gerais das partidas não abrangem estatísticas do jogo em questão. A Tabela 2.3 lista os dados informativos e suas respectivas descrições.

Tabela 2.3 – Informações Gerais das Partidas

Informação	**Descrição**
Equipe	Equipe mandante da partida
Adversário	Equipe visitante da partida
Temporada	Temporada em que a partida foi realizada
Dias de descanso time	Número de dias sem jogos
Porcentagem de vitórias	Percentual de jogos vencidos na temporada
Point Spread	Total de pontos favorito ou azarão

Fonte: da autora, adaptado de (NBA, 2022).

O *point spread* ou "vantagem de pontos" nada mais é do que uma maneira que as casas de apostas encontraram para investir na quantidade de pontos que um time vai vencer ou perder. Todo o conceito é baseado na ideia de *underdogs* (azarões) e favoritos. O *Handicap* (ou Point Spread) no basquete foi criado para tornar o jogo de basquete mais interessante, já que em confrontos entre uma equipe muito forte contra uma equipe muito fraca, as apostas simples ou *Money Line* (aposta na vitória do favorito) não são lucrativas para as casas de apostas e apostadores. Para exemplificar o spread, vamos considerar o jogo da temporada de 2018–2019 entre Golden State Warriors $\times$ Toronto Raptors. Na oportunidade, as linhas do Handicap apresentadas pela Bet365 eram: *Golden State Warriors* -2.5 e *Toronto Raptors* $+2.5$. Em suma, essas linhas representam que as casas de apostas atribuíram uma desvantagem de 2.5 pontos para o Golden State, ou seja, "começava" o placar com 2.5 pontos a menos. Sendo assim, o Golden State precisava vencer por pelo menos 3 pontos para essa aposta ser vencedora. Portanto, o spread incentivou apostas em ambas as equipes, diferentemente do sistema em que apenas o vencedor da partida era escolhido. A fonte dessas informações é o artigo[2] de Vinícius Duarte para o Clube da Aposta, uma plataforma de educação em apostas.

A cada partida da NBA, são registradas diversas estatísticas. A Tabela 2.4 descreve as principais delas.

A NBA reúne estatísticas individuais simples, como pontos, rebotes e assistências. Com o passar do tempo, a análise quantitativa das partidas cresceu e novas estatísticas baseadas no conceito de posse de bola surgiram e são conhecidas como estatísticas avançadas. Uma das estatísticas mais comuns para essa avaliação é o *rating*, em que os pontos marcados e os pontos

[2] Disponível online em https://clubedaposta.com/apostas-basquete/como-funciona-handicap-basquete/. Acesso em fevereiro de 2022.

Tabela 2.4 – Estatísticas registradas por partida

Estatística	**Descrição**
PTS	Pontos Marcados
FGA	Tentativa de arremeso
FGM	Arremeso Convertido
2PA	Tentativa de arremesso de dentro do garrafão com valor de 2 pts
2PM	Arremesso convertido de dentro do garrafão com valor de 2 pts
3PA	Tentativa de arremesso de fora do garrafão com valor de 3 pts
3PM	Arremesso convertido de fora do garrafão com valor de 3 pts
FTA	Tentativa de arremesso de fora do garrafão com valor de 3 pts
FTM	Arremesso convertido de fora do garrafão com valor de 3 pts
AST	Passe que resulta na conversão de um arremesso do jogador recebedor
TOV	Perda da posse de bola
REB	Recuperar a bola após um arremesso não convertido
OREB	Recuperar a bola após um arremesso não convertido do próprio time
DREB	Recuperar a bola após um arremesso não convertido do time adversário
BLK	Bloquear o arremesso do adversário de chegar a cesta
STL	Roubar a bola do adversário

Fonte: da autora, adaptado de (NBA, 2022).

sofridos são avaliados a cada 100 posses de bola conforme as Equações 2.1 e 2.2, em que *PTS* representa os pontos marcados e *Oppt PTS* os pontos sofridos pela equipe. Para a avaliação geral da eficiência de uma equipe foi criado o *NET rating*, conforme a Equação 2.3.

$$\text{ORTG} = \text{PTS} \div \text{POSS} \times 100 \tag{2.1}$$

$$\text{DRTG} = \text{Oppt PTS} \div \text{POSS} \times 100 \tag{2.2}$$

$$\text{NET RTG} = \text{ORTG} - \text{DRTG} \tag{2.3}$$

O *True Shooting (TS%)* é um forma de buscar mais detalhes sobre a eficiências de um jogador do time no que se refere aos arremessos. Uma diferença importante em relação ao *(EFG%)*, que será explicado com mais detalhes a seguir é que o *True Shooting* considera também os lances livres que ocorrem quando o jogador sofre falta no ato de arremessar, conforme a Equação 2.4, em que 0.44 é o coeficiente de lances livres que resultam no fim de uma posse de bola.

$$\text{POSS} = \text{FGA} + 0.44 \times \text{FTA} - \text{OREB} + \text{TOV} \tag{2.4}$$

Oliver (2004) identificou quatro medidas de aproveitamento nas partidas de basquete que aumentam as chances de um time vencer a partida. Uma das medidas é o *Effective Field Goal Percentage (EFG%)* que mede o aproveitamento de um time ou de um jogador nos arremessos. Neste caso, as bolas de três pontos ganham um peso um pouco maior em relação a de dois pontos, conforme a Equação 2.5.

$$\text{EFG\%} = (\text{FGM} + 0.5 \times \text{3PM}) \div \text{FGA} \tag{2.5}$$

O *Turnover Percentage (TOV%)* é a segunda medida em que é considerado o número de erros ofensivos cometidos por uma equipe ou de um jogador a cada 100 posses de bola, conforme a Equação 2.6.

$$\text{TOV\%} = (\text{TOV} \div \text{POSS}) \times 100 \tag{2.6}$$

A terceira medida é a *Offensive Rebound Percentage (ORB%)*, em que é considerado o número de rebotes disponiveis para o ataque que o mesmo conseguiu pegar, aumentando as chances de marcar pontos, conforme a Equação 2.7.

$$\text{ORB\%} = \text{ORB} \div (\text{ORB} + \text{Oppt DREB}) \tag{2.7}$$

A última medida *Free Throw Rate (FTR)* é a proporção entre o número de lances livres cobrados em relação a quantidade de arremessos tentados na partida, podendo ser aplicado para um time ou para jogador, conforme a Equação 2.8.

$$\text{FTR} = \text{FTA} \div \text{FGA} \tag{2.8}$$

2.2 Trabalhos Relacionados

Nesta seção são descritos quatro trabalhos relacionados à temática de predição de resultados de jogos, sobretudo, na NBA. Nesses trabalhos são analisadas as técnicas de mineração de dados e de aprendizado de máquina com o objetivo de avaliar o que foi feito de importante e os resultados obtidos sobre o problema em questão para fins de comparação.

Bogoni (2019) propõe usar dados estatísticos e informações de jogos da NBA de cinco temporadas no período entre 2013 e 2018. A coleta dos dados foi realizada no site bigdataball e kaggle que possuem as estatísticas de cada time por partida realizada. Na primeira etapa do pré-processamento foi realizada a transformação de algumas variáveis e a eliminação de variáveis redundantes resultantes da integração das duas bases de dados. Em seguida, foram criadas funções no banco de dados de modo que as estatísticas fossem acumuladas e calculadas para cada partida. A informação referente ao percentual de vitórias na temporada, por exemplo, não estava presente nos dados anteriormente coletados e para isso foi criada uma função no banco de dados para obter esse valor. Após o pré-processamento, foi realizada a seleção dos atributos de maior influência. Para isso, foram aplicados três algoritmos. O primeiro algoritmo aplicado foi o *ExtraTreesClassifier*, em que foram parametrizadas 250 árvores e os atributos de maior influência foram o *spread* de pontos e o percentual de vitória. O segundo algoritmo aplicado foi o *Ridge* que

gerou resultados diferentes do *ExtraTreesClassifier*, atribuindo maior influência para os rebotes ofensivos e percentual de vitórias. O último algoritmo aplicado foi o *RandomForestRegressor* que também gerou 250 árvores e que, assim como o algoritmo *ExtraTreesClassifier*, definiu o spread como atributo de maior influência.

Os dez atributos com melhor média foram selecionados para a aplicação dos algoritmos de classificação. O conjunto de dados de entrada foi dividido em conjunto de treinamento e conjunto de teste, em que o conjunto de treinamento representou dois terços do total de dados. De acordo com Bogoni (2019), os algoritmos de indução que levaram aos melhores resultados foram *Logistic Regression* e *Perceptron Multi-Camadas (MLP)*, com acurácias de 67.94% e 68.04%, respectivamente. Em relação aos atributos, os mais influentes para a classificação do vencedor foram: o sistema de pontos de favoritismo e a campanha da equipe naquela temporada.

Cao (2012) coletou os dados estatísticos de jogos referentes a seis temporadas da NBA no período de 2005 a 2011. Para facilitar as tarefas de mineração de dados, um datamart foi construído para armazenar dados de estatísticas da NBA limpos e bem gerenciados. O processo de extração de recursos, que extrai recursos representativos do datamart, prepara as amostras de dados que podem ser consumidas diretamente pelas ferramentas de treinamento de modelo. O processo de extração assume que o jogo a ser previsto é o próximo jogo entre dois times. E todos os dados disponíveis são anteriores a data desse jogo. Alguns recursos extraídos por Cao (2012) e incorporados as demais estatísticas foram: o número de jogos nos últimos cinco dias, o número de dias de descanso antes do próximo jogos, vitórias ou derrotas nas ultimas dez partidas da NBA e estatísticas referentes aos últimos jogos realizados entre os dois times em temporadas anteriores. As estatísticas analisadas são relativas as temporadas regulares de 2005-2011 e é dividido em conjunto de treinamento, teste e validação. Para isso, o autor optou por utilizar Validação Cruzada com 10 folds por considerar que 6.000 dados não eram suficientes para aplicar *handout*, abordagem mais comum que divide os dados em dois conjuntos complementares, conjunto de treinamento e conjunto de testes. Assim, as temporadas 2006-2010 foram utilizadas para treinamento e teste.E a temporada de 2011 foi utilizada para avaliar modelo. Para esse experimento, quatro modelos foram utilizados: *Simple Logistics*, *Naïve Bayes*, *Support Vector Machine* (SVM) e Redes Neurais Artificiais. Segundo Cao (2012), os algoritmos de indução apresentaram os seguintes resultados *Simple Logistics* com 67.82% de acurácia, *Naïve Bayes* com 65.82%, *Support Vector Machine* com 67.22% e Redes Neurais Artificiais com 66.67%.

Praet (2017) escolheu o tênis para a construção de um modelo de machine learning capaz de prever os resultados dos jogos. O fato de o tênis ser um esporte individual e com estatísticas de apenas um jogador e a influência limitada que o dinheiro exerce sobre o esporte quando comparado com outros esportes, fez com que o autor considerasse o tênis um candidato viável para se explorar a predicabilidade resultados de jogos. Os dados foram coletados de diversos websites através de um script desenvolvido em Python e são referentes a temporada de 2006 a 2016, totalizando 66.113 partidas. Os 500 melhores jogadores, segundo o ranking do ATP

World Tour, foram selecionados para fazerem parte da pesquisa. Para armazenamento dos dados, colocou-se uma partida por linha em uma tabela de um banco de dados MySQL devido à sua interoperabilidade e segurança. Para a limpeza dos dados, um script em python foi desenvolvido para eliminar ruídos mais evidentes como jogos que terminam em menos de 15 minutos devido a lesão de um dos jogadores. Segundo o autor, ATPWorldTour é uma das bases de dados mais utilizadas quando se deseja obter estatísticas relacionadas a jogos de tênis. No entanto, essa classificação tem algumas limitações e, por isso, os sistemas de classificação foram testados. Um desses sistemas foi o *Content-Based Filtering* ou filtragem baseada em conteúdo. De acordo com Praet (2017), a filtragem baseada em conteúdo é um sistema de recomendação em que palavras-chave são usadas para descrever os itens e um perfil de usuário é criado para indicar o tipo de item que esse usuário gosta. Em outras palavras, esses algoritmos tentam recomendar itens semelhantes àqueles que um usuário gostou no passado. Para essa pesquisa, os itens representam as partidas esportivas e as palavras-chave são as estatísticas das partidas e jogadores. Outro método estatístico, o *Elo Rating*, foi utilizado para tentar amenizar a contabilização dos pontos de cada jogador que não considera algumas variáveis como a fase atual do adversário. E por fim, foi utilizado o método *collaborative filtering systems* ou sistema de filtragem colaborativa em que um número de usuários é selecionado com base em sua similaridade com o usuário ativo. Assim, ao invés de considerar todas as partidas anteriores de um determinado jogador, apenas as partidas que são feitas com adversários contra os quais ambos os jogadores competiram são consideradas. Apesar de o número de jogos serem relativos a 10 anos, o autor considerou que a quantidade de dados era limitada e, por isso, optou por utilizar o algoritmo *cross-validation* para teste e treinamento. Os métodos de avaliação utilizados foram focados na probabilidade de predição, sendo eles: *accuracy* A_i, *Logarithmic loss* (ou *logloss*) e *Return-On-Investment* (*ROI*). Os resultados obtidos para os algoritmos classificadores aplicados são apresentados na Tabela 2.5.

Tabela 2.5 – Comparison of the best obtained models regarding prediction accuracy

Classifier	**Parameters**	**Accuracy (%)**
Bagging + logistic regression	38 Iterations & Bag Size % 44	69.3193
Logistic regression	Ridge Estimator 10^{-8}	69.3072
AdaboostMl + Logistic regression	Standard	69.1958
Random Forest	Standard	67.9729

Fonte: Praet (2017).

Marveldoss (2018), em sua tese, utilizou uma abordagem Elo-based em que foi empregada para modelar a força individual de cada jogador de basquete com base na pontuação plus-minus desse jogador. Segundo o autor, a pontuação plus-minus é uma métrica poderosa por quantificar a contribuição de cada jogador como, por exemplo, a capacidade de defesa. A classificação dos jogadores é combinada para obter a classificação da equipe e as classificações das equipes são comparadas aos pares para obter a probabilidade de vitória de cada uma das equipes durante um confronto. Esse método não apenas prevê vitórias e derrotas, mas também oferece mais

informações do que o sistema de classificação Elo, já que as classificações são atribuídas a cada jogador ao invés de considerar apenas as equipes. Essas informações incluem, por exemplo, o efeito de transferências no meio da temporada ou o impacto de lesões. Como não há informações prévias relativas a força de cada jogador de basquete, os jogadores recebem um valor padrão de 1000 em seu primeiro jogo. As estatísticas analisadas são relativas as temporadas regulares de 2015-2018 e é dividido em conjunto de treinamento, validação e teste, em que 75% de conjunto de treinamento e 25% de conjunto de validação cruzada. O conjunto de treinamento é usado para atualizar a classificação dos jogadores sem nenhuma métrica de desempenho. O conjunto de validação é usado para escolher os parâmetros ótimos do modelo. O desempenho do algoritmo proposto é comparado com o algoritmo Elo padrão medindo a média da discrepância preditiva ao longo da temporada e a taxa de previsão. Nos resultados obtidos por Marveldoss (2018), percebe-se que o algoritmo Elo padrão supera o algoritmo Elo-based. Isso se deve ao fato de que ao considerar o rating individual dos jogadores aumenta-se a complexidade do algoritmo proposto quando comparado com a complexidade do algoritmo Elo padrão. Entretanto, esse algoritmo fornece informações valiosas para a gestão da equipe como a possibilidade de quantificar o valor de um jogador para equipe e viabilidade e valor de venda desse atleta.

Com base na leitura e no aprendizado adquirido à partir das pesquisas relacionadas, este trabalho propõe uma ferramenta capaz de predizer os resultados dos jogos da NBA através da mesclagem de algumas abordagens apresentadas pelos autores. Alguns recursos incorporados às estatísticas utilizadas neste trabalho como, por exemplo, o número de vitórias nos últimos jogos, baseou-se nos bons resultados obtidos por Cao (2012), ao considerar o número de vitórias no últimos 5 jogos. A base de dados kaggle utilizada por Bogoni (2019) foi também importante neste trabalho, para obtenção dos resultados apresentados, por conter estatísticas organizadas e sem valores ausentes. Entretanto, um diferencial desta pesquisa é o conjunto de teste escolhido que considera as temporadas disputadas durante a pandemia do coronavírus com o objetivo de avaliar os impactos das alterações das dinâmicas dos jogos na predição de resultados da NBA.

3 Descrição dos Conjuntos de Dados

Neste capítulo, são apresentados os dados de entrada usados na construção do método de trabalho descrito no capítulo 4, visando alcançar os objetivos propostos nesse trabalho. Nas próximas seções, são descritas as fontes dos dados, as variáveis utilizadas, o embasamento teórico para a sumarização e criação de algumas dessas variáveis e a redução do conjunto de dados.

3.1 Escolha dos Dados

Os dados obtidos são referentes às temporadas regulares da NBA realizadas entre 2003-2022 e incluíram 52 variáveis que caracterizaram a eficácia ofensiva e defensiva de 30 equipes. Para decidir a fonte de dados, muitos sites foram visitados. A maioria dos sites tem uma enorme variedade de dados que vão desde a pontuação obtida por cada equipe até o salário dos jogadores. Alguns pontos foram essenciais para decidir a fonte de dados: o número de temporadas fornecidas e a facilidade em extraí-los do site. Os site nba.com e kaggle.com apresentaram essas características e por esse motivo foram escolhidos como as duas fontes principais. Alguns exemplos de características obtidas da primeira fonte são mostradas na Figura A.1 no apêndices.

3.2 Descrição das Variáveis

A enorme quantidade de variáveis disponíveis e a natureza dinâmica do basquete faz com que a avaliação da influência de cada característica dos dados de entrada extraídos seja essencial para a obtenção de bons resultados na construção e aplicação de um modelo capaz de atuar na predicabilidade dos resultados de jogos da NBA.

Um das formas de analisar um time de basquete de modo que sua vitória possa ser prevista é através do total de pontos marcados, dos arremessos em quadra, dos lances livres e dos rebotes ofensivos e defensivos nos últimos e jogos e nas últimas temporadas. Para um bom aproveitamento das estatísticas, descritas anteriormente na Subseção 2.1.3, é necessário transformar e relacionar essas métricas para que se possa capturar o desempenho da equipe em um período de tempo. Além disso, é importante associar essas estatísticas com outras informações que podem influenciar o resultado da partida como, por exemplo, se o time é o mandante do jogo em que se pretende prever o vencedor.

De acordo com os dados divulgados em nba.com, dos quatro principais esportes americanos, a vantagem de jogar em casa é mais significativa na NBA, com as equipes ganhando consistentemente cerca de 60% de seus jogos da temporada regular em suas arenas. Para verificar tal comportamento nos dados coletados, criou-se uma nova coluna `HOME_TEAM_WINS` no

Dataframe com o intuito de identificar as equipes que haviam vencido em casa. Essa informação foi utilizada para concatenar as médias e as porcentagens das estatísticas em `home` e `away`.

Além disso, segundo (ARKES; MARTINEZ, 2011), o número de vitórias e derrotas nos últimos jogos resulta em uma maior probabilidade de ganhar ou perder as próximas partidas. Portanto, além de transformar e gerar novos recursos relacionados às temporadas anteriores por equipe, foram criadas variáveis como a porcentagem de vitórias como visitante e como mandante nos últimos 11 e 100 jogos dentre outras variáveis que são brevemente explicadas na Tabela B.1 nos apêndices.

3.3 Redução dos Dados

Para a construção do modelo proposto neste trabalho o conjunto de dados foi reduzido às partidas das últimas oito temporadas disponíveis (de 2011-12 até 2021-22).

As temporadas anteriores a 2011-2012 foram descartadas para assegurar que tendências atuais como, por exemplo, o aumento do número de arremessos de três pontos, não tenham impacto na previsão do modelo. A temporada de 2011-2012 também foi eliminada do conjunto de dados por ter sido reduzida, dos 82 jogos comuns por equipe para 66, devido a quase dois meses de inatividade.

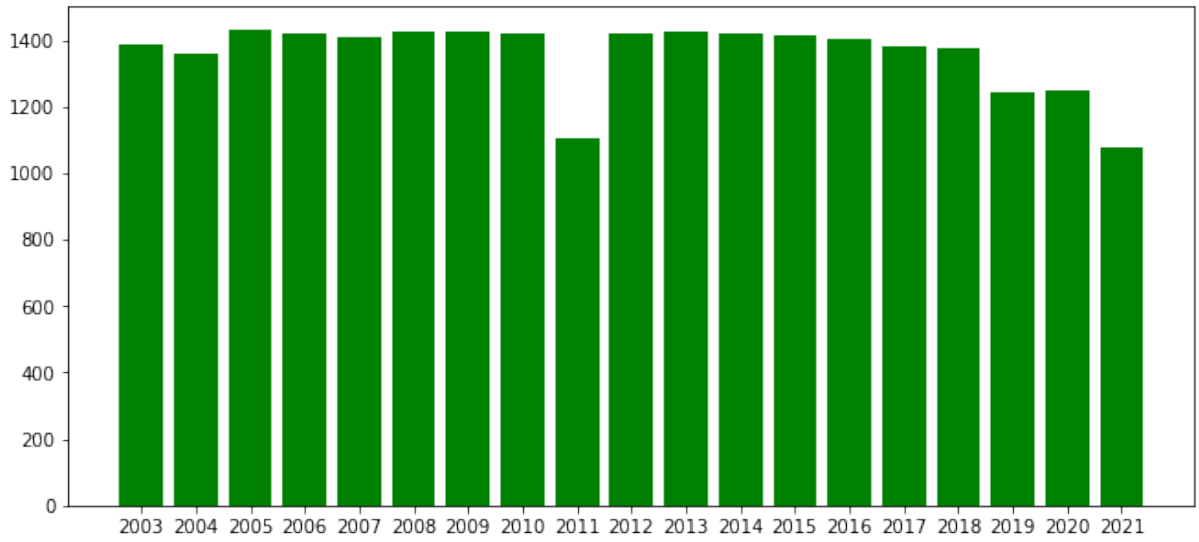


Figura 3.1 – Quantidade de jogos por temporada

Fonte: da autora.

Durante a pandemia do coronavírus, a NBA teve novamente sua temporada reduzida , conforme o gráfico 3.1. Além disso, a dinâmica dos jogos foi alterada com o intuito de diminuir circulação do vírus. O número total de partidas, por exemplo, disputado por cada equipe foi reduzido de 82 para 72 jogos. Esses fatores sem dúvida impactaram nos resultados dos jogos e consequentemente no campeão da liga. Entretanto, as temporadas 2019-2020 e 2020-2021 não

foram excluídas por ser escopo dessa pesquisa analisar o impacto dessas alterações na predição dos resultados dos jogos da NBA.

4 Método de Trabalho

Este capítulo descreve as etapas realizadas durante aplicação do método de trabalho utilizado. Nas próximas seções são descritos visão geral do método, o pré-processamento dos dados, o processo de transformação dos dados e o processo de seleção de *features* na construção do modelo.

4.1 Visão Geral

O método de trabalho proposto nesta pesquisa, primeiramente, compreende os aspectos conhecidos da NBA que mais influenciam nos resultados das partidas a partir da compreensão do assunto e dos dados. Essa etapa garante que as variáveis *(features)* já conhecidas pela sua relevância na predição de sucesso de uma equipe não sejam ignoradas.

Para garantir que novos conhecimentos sejam alcançados, o próximo passo foi a preparação dos dados em que analisou-se a granularidade ou detalhamento das informações para que se pudesse gerar diferentes variáveis *(features)* a partir de diferentes agrupamentos e sumarizações.

Para a implementação do modelo, realizou-se a seleção dos classificadores através de inúmeros testes com diferentes combinações de variáveis e hiperparâmetros e, em seguida, avaliou-se os resultados obtidos a partir de um conjunto de entradas.

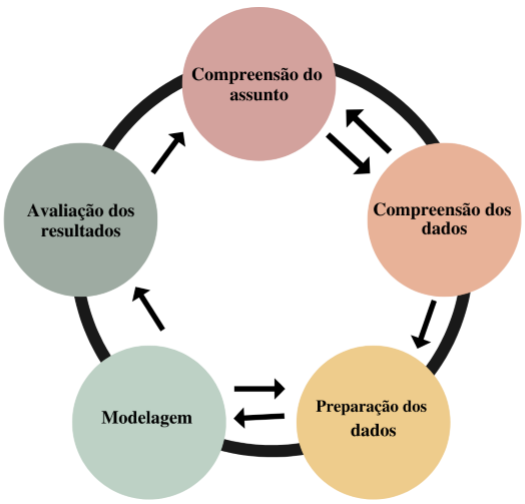


Figura 4.1 – Metodologia utilizada na construção da Ferramenta

Fonte: da autora.

4.2 Pré-Processamento

Neste processo, a primeira etapa consistiu em transformar as estatísticas coletadas e gerar novas variáveis que pudessem ser úteis na previsão de vitória de uma equipe, conforme descrito no Capítulo 3.

4.2.1 Normalização

Considerando a existência de diferentes variáveis com diversas escalas de avaliação, o próximo passo a ser realizado antes da seleção das variáveis foi a normalização dos dados. Por exemplo, um jogador pode marcar mais 30 de pontos em um jogo, mas não pode cometer 6 faltas. Para isso, utilizou-se a abordagem Min-Max por resultar em dados com um menor desvio padrão quando comparado à abordagem de padronização por *z-score*.

A normalização desses dados, mostrada na Figura 4.2, evita que uma variável tenha sua importância reduzida por ter uma escala mais baixa em relação à aquelas variáveis que têm valores com uma escala maior. Além disso, com os dados normalizados o tempo de treinamento necessário para o aprendizado de máquina é menor, aumentando a eficiência da aplicação do modelo.

Após a normalização das variáveis, `PTS_h_11g`, `AST_h_11g`, `REB_h_11g`, `PTS_a_11g`, `AST_a_11g`, `REB_a_11g`, `PTS_h_100g`, `AST_h_100g`,`REB_h_100g`, `PTS_a_100g`, `AST_a_100g`, `REB_a_100g`, descritas na Tabela B.1, todos os valores pertenciam ao intervalo [0,1].

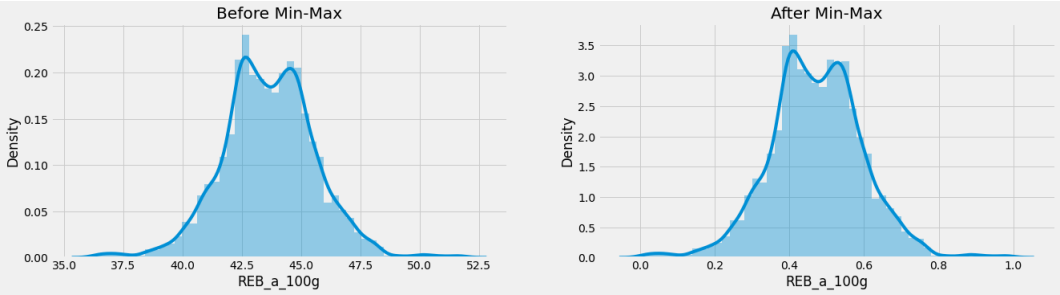


Figura 4.2 – Gráfico Min-Max

Fonte: da autora.

4.2.2 Seleção de Feature

A próxima etapa do pré-processamento é entender melhor as relações entre as variáveis e também identificar quais são as mais relacionadas à vitória. Para isso, foi calculado a correlação de Pearson entre as variáveis. A Figura 4.3 mostra as 42 variáveis mais relacionadas com a vitória.

A partir da matriz de correlação é possível notar que as variáveis mais relacionadas com a vitória do time mandante são aquelas relacionadas à porcentagem de vitória nos últimos jogos

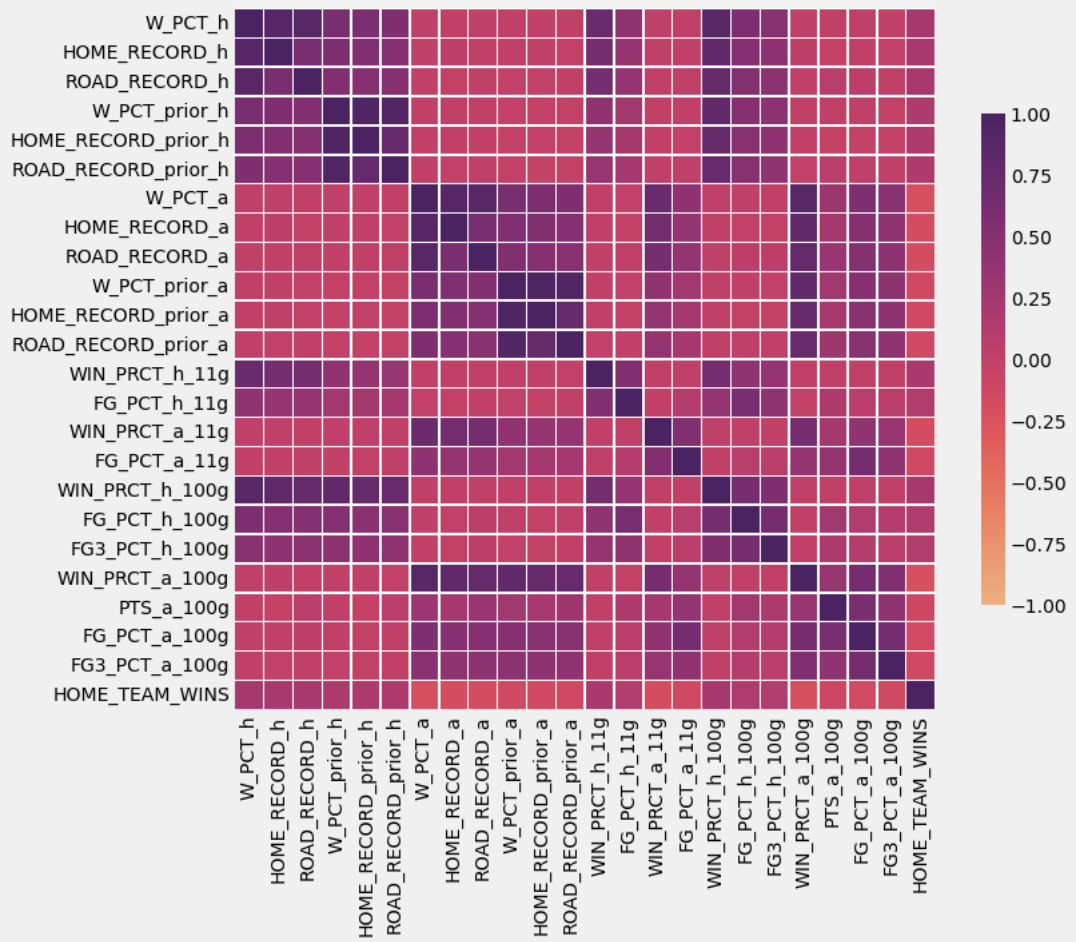


Figura 4.3 – Matriz de Correlação

Fonte: da autora.

e nas últimas temporadas. Apesar de a matriz de correlação fornecer informações importantes sobre o conjunto de dados, essa grande quantidade de variáveis dificulta a criação dos modelos. Um dos maiores desafios, por exemplo, é o tempo necessário para o modelo ser aprendido.

Por isso,o uso de técnicas de seleção de variáveis/atributos é essencial para evitar que os dados possuam variáveis redundantes e insignificantes. A utilização de um algoritmo capaz de encontrar um subconjunto de variáveis que funcione melhor com os algoritmos de aprendizagem permite a remoção sem perda de informação de algumas variáveis que não contribuem para o resultado.

Para a seleção de atributos utilizou-se o algoritmo de Eliminação Recursiva de Variáveis RFE (Recursive Feature Elimination), que seleciona de maneira automática as variáveis de maior influência no resultado e exclui aquelas de menor importância até que um número determinado de variáveis seja atingido. Essa influência das variáveis é medida através da colinearidade, ou seja, aquelas variáveis de maior colinearidade são consideradas menos importantes (LUNELLI, 2020).

Para implementar o RFE, utilizou-se o pacote Scikit-learn que implementa algoritmo

RFECV que combina eliminação recursiva com validação cruzada.

As variáveis selecionadas e o número ótimo de variáveis depende do algoritmo de classificação considerado. Apesar disso, as variáveis `W_PCT_h, W_PCT_a, WIN_PRCT_h_11g,WIN_PRCT_h_100g, WIN_PRCT_a_100g` foram selecionadas para todos os algoritmos de classificação.

4.3 Indução do Classificador

A seleção dos modelos tratou de experimentar vários classificadores e testá-los com base na acurácia obtida e mantendo os dados de entrada anteriormente selecionados constantes. Os classificadores induzidos pelas técnicas RandomForest, Naïve Bayes, Regressão logística, Árvore de Decisão e XGBoost Classifier, baseados no paradigma simbólico e probabilístico descrito na subseção 2.1.1.4, foram os que apresentaram os melhores resultados e, por isso, foram os escolhidos para aplicação do módulo GridSearchCV, que otimiza os valores dos hiperparâmetros de cada algoritmo.

5 Avaliação Experimental

Este capítulo apresenta os experimentos em que realizou-se diversos testes com o intuito de analisar e avaliar os resultados obtidos.

5.1 Configuração dos Experimentos

Todas as etapas de desenvolvimento da ferramenta proposta nesta pesquisa foram realizadas utilizando a linguagem de programação Python. As bibliotecas Pandas e Numpy, que fornecem estruturas de dados e ferramentas de análise de dados de alto desempenho, foram constantemente utilizadas.

Em relação à configuração dos dados de entrada, o conjunto de treinamento foi formado pelas temporadas de 2012-2013 a 2017-2018, totalizando 8.610 partidas. Com o intuito de analisar o impacto da pandemia do coronavírus na predição dos resultados de jogos da NBA, considerando as alterações nas dinâmicas dos jogos, o primeiro conjunto de testes foi formado pela temporada de 2019-2020, disputada durante a pandemia, em que foram realizadas 1.080 partidas. Para que se possa comparar os resultados, o segundo conjunto de teste foi formado pela temporada 2018-2019, disputada antes da pandemia, totalizando 1.230 partidas.

Para aplicação dos algoritmos de classificação, descritos na Seção 4.3, utilizou-se uma abordagem de aprendizagem supervisionada em que os classificadores indicam se o vencedor da partida é o mandante ou não.

Os dados foram, portanto, primeiramente divididos em duas variáveis (x e y) em que x é o conjunto das 23 variáveis mais relacionadas com a vitória, obtidas a partir da matriz de correlação, e y a variável que define a classe a qual o dado pertence. Em Python, essa abordagem foi representada da seguinte forma:
`x_train, y_train = train[select_columns], train['HOME_WINS']` e
`x_test, y_test = test[select_columns], test['HOME_WINS']`.

Durante o desenvolvimento de cada modelo foi utilizado o algoritmo de eliminação recursiva de variáveis (RFECV) para a seleção de variáveis, disponível no módulo *feature_selection* da biblioteca *sklearn*, conforme detalhado na seção 4.2.2, em que definiu-se o número de *folds* igual a 10 . Esse número foi definido através de diversos testes realizados em que alterava-se os parâmetros desse algoritmo de seleção e avaliava-se o resultado a partir da acurácia do modelo.

Considerando que a base de dados utilizada para esta pesquisa possui uma boa variedade de resultados, sendo eles, as vitórias e as derrotas de cada equipe, optou-se por utilizar a acurácia através do módulo *accuracy_score* da biblioteca *sklearn* como métrica de avaliação da ferramenta proposta. Assim, após a aplicação dos algoritmos de classificação comparou-se as classificações

previstas com as originais associadas ao grupo de teste para o cálculo da acurácia.

O ajuste de hiperparâmetros para cada um desses algoritmos de classificação dependeu mais dos resultados experimentais e, portanto, o melhor método para determinar as configurações ideais foi experimentar várias combinações diferentes e avaliar o desempenho de cada modelo.

No entanto, avaliar a acurácia de cada modelo apenas pelo conjunto de treinamento pode levar a um dos problemas mais fundamentais em aprendizado de máquina: o *overfitting*. Por esse motivo, avaliou-se a acurácia tanto do conjunto de teste quanto do conjunto de treinamento.

Ainda em relação à avaliação dos modelos, para mostrar a frequência de classificação para cada classe do modelo foram utilizadas matrizes de confusão através do método *confusion_matrix*.

5.2 Resultados e Análises

No primeiro experimento mostrado na Tabela 5.1 considerou a temporada de 2019-2020, disputada durante a pandemia, como o conjunto de teste. Os resultados obtidos à partir desse experimento se mostraram muito semelhantes com destaque para o algoritmo Naïve Bayes que obteve a melhor acurácia. *overfitting*.

Tabela 5.1 – Acurácias dos modelos

Modelo	**Acurácia no Treino**	**Acurácia no Teste**	**Tempo de Treino (s)**
Random Forest	73,23%	61,75%	10
Logistic Regression	65,75%	61,11%	4
Árvore de Decisão	66,91%	60,22%	18
Naive Bayes	65,23%	60,84%	6
XGBClassifier	70,91%	60,75%	4

Fonte: da autora.

Os algoritmos Random Forest, Naive Bayes, XGBoost Classifier foram os que apresentaram melhores acurácia de teste. Como se pode observar na Tabela 5.1, a diferença da acurácia em relação ao conjunto de treino e de teste foi maior para os algoritmos Random Forest e XBClassifier.

Em relação ao algoritmo Random Forest, essa maior diferença pode estar relacionada com a quantidade de dados utilizados para o treino do modelo. Como esse algoritmo é mais robusto por combinar várias árvores de decisão, é provável que a quantidade de dados tenha sido insuficientes para construir um modelo com maior taxa de acerto em que a acurácia do teste esteja mais próxima da acurácia do treino.

Outro ponto que justifique essa diferença pode estar associado com a profundidade máxima da árvore definida nos hiperparâmetros dos algoritmos Random Forest e XBClassifier . Normalmente, as árvores com grande profundidade tende a responder muito bem a dados conhecidos, porém respondem mal a dados desconhecidos.

Os resultados mostrados na Tabela 5.2 referentes ao segundo experimento que considerou a temporada de 2018-2019, como conjunto de teste, também foram muito semelhantes para todos os algoritmos aplicados com destaque para o Naïve Bayes que obteve novamente o melhor resultado.

Tabela 5.2 – Acurácias dos modelos

Modelo	**Acurácia no Treino**	**Acurácia no Teste**	**Tempo de Treino (s)**
Random Forest	72,08%	65,45%	11
Logistic Regression	66,19%	65,52%	6
Árvore de Decisão	67,32%	64,57%	19
Naive Bayes	65,23%	67,82%	8
XGBClassifier	67,15%	63,13%	3

Fonte: da autora.

Essa semelhança nos resultados dos algoritmos, observada em ambos os experimentos, pode estar relacionada com o fato de que todos os algoritmos foram treinados com a mesma quantidade de dados. Adequar a quantidade de dados de treino de acordo com o algoritmo pode ajudar a obter uma acurácia melhor.

Uma hipótese para o algoritmo Naïve Bayes ter tido o melhor resultado nos dois experimentos é o fato de a baixa correlação entre as *features* selecionadas ter sido ignorada. Esse algoritmo assume que as features são independentes entre si e, portanto, considera todas igualmente importantes para o resultado.

A diferença no primeiro experimento entre a acurácia obtida no treino e no teste é consideravelmente maior para todos os algoritmos se comparada com os resultados obtidos no segundo experimento. Essa maior diferença no primeiro experimento indica que houve perda de generalização causada pela alteração de padrões de vitórias e derrotas durante a pandemia.

Essa menor capacidade de generalização observada no primeiro experimento é atribuída ao fato de que quando utilizamos modelos para prever dados não observados durante o treinamento, a performance diminui consideravelmente, pois os padrões aprendidos durante o treinamento não estão na amostra de teste.Por esse motivo, pode-se aformar que as alterações nas dinâmicas dos jogos realizados durante a pandemia alterou os padrões de vitórias e derrotas dos times mandantes da NBA.

Apesar de o modelo ter tido sua capacidade de generalização reduzida quando o conjunto de teste foi a temporada disputada durante a a pandemia, o resultado ainda é superior à escolha ao acaso em que considera-se a porcentagem bruta de equipes que venceram em casa as partidas da NBA.

Para as temporadas disputadas durante a pandemia essa porcentagem é igual a 53.8%. Entretanto, para as partidas disputadas antes da pandemia essa porcentagem foi um pouco maior, estando também alinhado com o resultado obtido nessa pesquisa. Para esse caso, 58% das equipes

venceram em suas arenas.

6 Considerações Finais

6.1 Conclusão

Este trabalho se propôs a aplicar técnicas de mineração de dados na predição de resultados de jogos da NBA e comparar algoritmos de aprendizado de máquina de modo a avaliar a performance dos mesmos na previsão de resultados de jogos da NBA.

O projeto foi construído a partir da coleta dos dados estatísticos, passando pelo pré-processamento, inferência de novas variáveis , treinamento do modelo preditivo e avaliação dos resultados, a fim de avaliar possíveis tarefas e técnicas de mineração que possam ser utilizadas nessas predições.

Em relação aos resultados obtidos, o algoritmo Naïve Bayes obteve o melhor resultado para os dois experimentos. Para o primeiro experimento que considerou a temporada de 2019-2020, disputada durante a pandemia como conjunto de teste, a acurácia no teste e no treino foram de 60.84% e 65.23%, respectivamente. Para o segundo experimento que considerou a temporada de 2018-2019, disputada antes pandemia como conjunto de teste, a acurácia no teste e no treino foram de 67.82% e 65.23%, respectivamente. A diferença no primeiro experimento entre a acurácia obtida no treino e no teste é consideravelmente maior para todos os algoritmos se comparada com os resultados obtidos no segundo experimento. Isso indica uma perda de generalização do modelo causada pela alteração de padrões de vitórias e derrotas durante a pandemia em relação aos padrões anteriores à pandemia. A maior perda de generalização foi observada para o algoritmo Random Forest que obteve 73,23% e 61,75% de acurácia no treino e no teste, respectivamente.

A partir dos estudos realizados e resultados obtidos, pode-se afirmar que as alterações nas dinâmicas dos jogos realizados durante a pandemia impactou na predição de resultados de jogos da NBA. Além disso, a porcentagem de vitórias em casa é um fator de extrema influência na NBA, considerando que a queda desse valor durante a pandemia para 53.8% foi um dos fatores que resultou em perda de generalização do modelo preditivo.

6.2 Trabalhos Futuros

Para melhorar os resultados e como sugestão para trabalhos futuros, são descritas possíveis maneiras de se obter melhores resultados na classificação de partidas da NBA:

- Explorar as variáveis que mais influenciaram diante de toda a mudança na dinâmica dos jogos durante a pandemia do COVID, nos resultados das partidas da NBA com a orientação

de um especialista em basquete;

- Avaliar a influência de algumas característica dos jogadores das equipes nos resultados das partidas como, por exemplo, a altura média e a idade;
- Incluir novas variáveis com informações relativas aos investimentos realizados pelos times da NBA nas temporada realizadas durante a pandemia. O salário médios dos jogadores de cada equipe da partida é uma das variáveis que pode melhorar o desempenho do modelo;

Referências

ARKES, J.; MARTINEZ, J. Finally, evidence for a momentum effect in the NBA. *Journal of Quantitative Analysis in Sports*, De Gruyter, v. 7, n. 3, 2011.

BOGONI, J. P. *Aplicação de técnicas de mineração de dados para previsão de jogos de basquete*. Monografia (Graduação em Sistemas de Informação) — Centro de Ciências Exatas e Tecnológicas da Universidade do Vale do Taquari, 2019.

CAO, C. *Sports Data Mining Technology Used in Basketball Outcome Prediction*. Dissertação (Mestrado) — Technological University Dublin, Dublin, 2012.

CASTRO, L. N. de; FERRARI, D. G. *Introdução à Mineração de Dados: conceitos básicos, algoritmos e aplicações*. São Paulo: Saraiva, 2016.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Knowledge discovery and data mining: Towards a unifying framework. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. Portland, Oregon: AAAI, 1996. (KDD), p. 82–88.

HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. Waltham: Elsevier, 2011.

LUNELLI, L. M. *Previsão de resultado de jogos da NBA com algoritmos de machine learning*. Dissertação (Mestrado) — Universidade Nova de Lisboa, Lisboa, 2020.

MARVELDOSS, R. E. D. *An Elo-Based Approach to Model Team Players and Predict the Outcome of Games*. Dissertação (Mestrado) — Texas A&M University, College Station, Texas, 2018.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes — Fundamentos e aplicações*, Manole, v. 1, n. 1, p. 32, 2003.

NBA. *The official site of the NBA for the latest NBA Stats & News*. 2022. <https://nba.com>.

OLIVER, D. *Basketball on Paper: Rules and Tools for Performance Analysis*. Washington: Potomac Books, 2004.

PRAET, R. *Predicting Sport Results by using Recommendation Techniques*. Dissertação (Mestrado) — Ghent University, Ghent, Belgium, 2017.

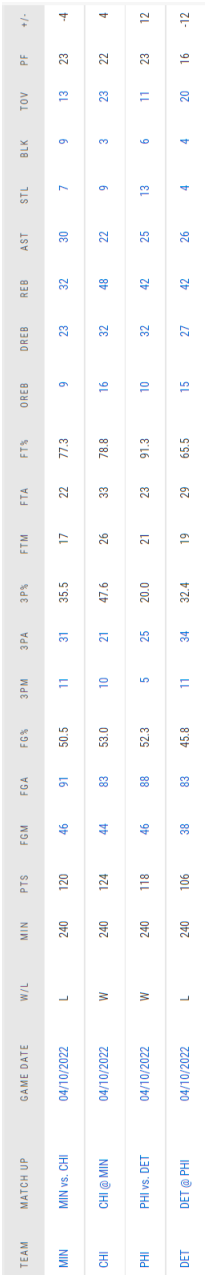
SCHMITT, J. *Pré-processamento para a mineração de dados: uso da análise de componentes principais com escalonamento ótimo*. Dissertação (Mestrado) — Universidade Federal de Santa Catarina, Florianópolis, 2005.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. *Introdução à Mineração de Dados com Aplicações em R*. Rio de Janeiro: Elsevier, 2017.

TAN, P. N.; STEINBACH, M.; KUMAR, V. *Introdução ao datamining: mineração de dados*. Rio de Janeiro: Ciência Moderna, 2009.

Apêndices

APÊNDICE A – Exemplos de Características



TEAM	MATCH UP	GAME DATE	W/L	MIN	PTS	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	STL	BLK	TOV	PF	+/-
MIN	MIN vs. CHI	04/10/2022	L	240	120	46	91	50.5	11	31	35.5	17	22	77.3	9	23	32	30	7	9	13	23	-4
CHI	CHI @ MIN	04/10/2022	W	240	124	44	83	53.0	10	21	47.6	26	33	78.8	16	32	48	22	9	3	23	22	4
PHI	PHI vs. DET	04/10/2022	W	240	118	46	88	52.3	5	25	20.0	21	23	91.3	10	32	42	25	13	6	11	23	12
DET	DET @ PHI	04/10/2022	L	240	106	38	83	45.8	11	34	32.4	19	29	65.5	15	27	42	26	4	4	20	16	-12

Figura A.1 – Exemplos de características do nba.com

Fonte:nba.com

APÊNDICE B – Significado de Atributos

Tabela B.1 – Atributos e seus significados — Dados do Time

Sigla	Significado
W_PCT_h	Porcentagem de vitória em casa nas últimas temporadas
W_PCT_a	Porcentagem de vitória fora de casa nas últimas temporadas
HOME_RECORD_h	Maior pontuação em jogos disputados em casa da equipe mandante na temporada
ROAD_RECORD_h	Maior pontuação em jogos disputados fora de casa da equipe mandante na temporada
W_PCT_prior	Porcentagem de vitória nas últimas temporadas
HOME_RECORD_prior_h	Maior pontuação em jogos disputados em casa da equipe mandante nas últimas temporadas
ROAD_RECORD_pior_h	Maior pontuação em jogos disputados fora de casa da equipe mandante nas últimas temporadas
HOME_RECORD_a	Maior pontuação em jogos disputados em casa da equipe não mandante na temporada
ROAD_RECORD_a	Maior pontuação em jogos disputados fora de casa da equipe não mandante na temporada
HOME_RECORD_prior_a	Maior pontuação em jogos disputados em casa da equipe não mandante nas últimas temporadas
ROAD_RECORD_prior_a	Maior pontuação em jogos disputados fora de casa da equipe não mandante nas últimas temporadas
WIN_PRCT_h_11g	Porcentagem de Vitória nos últimos 11 jogos do time mandante
WIN_PRCT_a_11g	Porcentagem de vitoria nos últimos 11 jogos do time mandante
FG_PCT_h_11g	Porcentagem de arremessos de 2 pontos convertidos nos últimos 11 jogos do time mandante
FG_PCT_a_11g	Porcentagem de arremessos de 2 pontos convertidos nos últimos 11 jogos do time não mandante
FG3_PCT_h_11g	Porcentagem de arremessos de 3 pontos convertidos nos últimos 11 jogos do time mandante
FG3_PCT_a_11g	Porcentagem de arremessos de 3 pontos convertidos nos últimos 11 jogos do time não mandante

AST_h_11g	Número de arremessos (perfeitos) convertidos em cestas nos últimos 11 jogos do time mandante
AST_a_11g	Número de arremessos (perfeitos) convertidos em cestas nos últimos 11 jogos do time não mandate
REB_h_11g	Número de rebotes nos últimos 11 jogos do time mandante
REB_a_11g	Número de rebotes nos últimos 11 jogos do time não mandante
PTS_a_11g	Total de pontos marcados no últimos 11 jogos do time não mandante
PTS_h_11g	Total de pontos marcados no últimos 11 jogos do time mandante
WIN_PRCT_h_100g	Porcentagem de vitória nos últimos 100 jogos do time mandante
WIN_PRCT_a_100g	Porcentagem de vitória nos últimos 100 jogos do time não mandante
FG_PCT_h_100g	Porcentagem de arremessos de 2 pontos convertidos nos últimos 100 jogos do time mandante
FG_PCT_a_100g	Porcentagem de arremessos de 2 pontos convertidos nos últimos 100 jogos do time não mandante
FG3_PCT_h_100g	Porcentagem de arremessos de 3 pontos convertidos nos últimos 100 jogos do time mandante
FG3_PCT_a_100g	Porcentagem de arremessos de 3 pontos convertidos nos últimos 100 jogos do time não mandante
AST_h_100g	Número de arremessos (perfeitos) convertidos em cestas nos últimos 100 jogos do time mandante
AST_a_100g	Número de arremessos (perfeitos) convertidos em cestas nos últimos 100 jogos do time não mandante
REB_h_100g	Número de rebotes nos últimos 100 jogos do time mandante
REB_a_100g	Número de rebotes nos últimos 100 jogos do time não mandante
PTS_a_100g	Total de pontos marcados no últimos 100 jogos do time não mandante
PTS_h_100g	Total de pontos marcados no últimos 100 jogos do time mandante
HOME_TEAM_WINS	Valor atribuido para cada equipe por partida. Em que 0 representa derrota e 1 vitória

Fonte: da autora