Universidade de Brasília
Departamento de Estatística

Modelos de Previsão da *National Basketball Association* (NBA)

Bruno Henrique Brandão de Souza

Brasília
2022

Bruno Henrique Brandão de Souza

Modelos de Previsão da *National Basketball Association* (NBA)

Orientador: Prof. Eduardo Monteiro de Castro Gomes

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2022

Agradecimentos

Aos meus pais, Cláudio e Márcia, por todo amor e dedicação que tiveram com o meu bem estar a vida inteira. Fizeram de tudo para eu estar em minha melhor versão, e não tenho palavras para agradecer pelo que fizeram por mim. Gratidão eterna.

Aos meus irmãos, Artur e Juliana, pelo companheirismo e ajuda em qualquer situação.

As minhas avós, Carmélia e Zélia, por ajudarem na minha formação e por todo carinho.

A minha namorada, Luísa, que além de ajudar e compartilhar momentos do curso e de UnB, me fez crescer como pessoa. Obrigado por tudo que passamos juntos.

Aos meus tios e primas pelo acolhimento familiar em todos os momentos.

Aos meus amigos, Bianchi, Dani, Franjinha, Lara, Luly e Xamps, por todos momentos de alegria e descontração que vocês me deram para eu conseguir chegar até aqui.

Aos meus amigos de curso, Amanda, Juliana, Matheus, Rafael e Ramon, por estarem sempre dispostos a ajudar nos desafios que o curso de estatística proporcionou.

Dedico este trabalho ao meu avô, Djalma, que infelizmente não está mais aqui, mas tenho certeza que ele está me protegendo onde estiver.

Resumo

Esse trabalho teve como objetivo principal elaborar um modelo que fosse capaz de realizar predições se um time classifica aos *playoffs* da NBA com base nas estatísticas da equipe de até 26 jogos disputados.

Os dados foram retirados de uma base disponibilizada no *Kaggle* e do site oficial da NBA via *Web scraping*. Esses dados eram referentes a cada partida disputada, com informações do time mandante e visitante. Depois foram criadas variáveis pelo processo de *feature engineering* para colaborar na predição dos modelos. Essas variáveis são relacionadas as estatísticas dos oponentes, média móvel das estatísticas dos time e dos seus adversários referente aos últimos jogos da quantidade de partidas estudadas e da força dos oponentes. Uma função foi criada para formar uma base de dados com as estatísticas para cada time de acordo com a quantidade de jogos que se deseja estudar. Depois uma outra função foi elaborada para realizar o ajuste dos modelos preditivos de acordo com a variação do número de jogos, número de jogos para a média móvel e a quantidade de temporadas utilizadas para a base de teste. Essa função elabora bancos de dados pelos parâmetros mencionados anteriormente, depois foram separadas as bases de treinamento e teste para a modelagem. As técnicas utilizadas foram de *Lasso*, *Ridge*, *Random Forest* e Árvore Binária. Por fim, uma grande tabela foi formada com as medidas que permitem a comparação dos modelos propostos.

Após análises, foi verificado que o modelo com melhor desempenho de acordo as intenções deste trabalho foi elaborado utilizando a técnica de *Random Forest*, estatísticas dos 15 primeiros jogos, 4 jogos para o cálculo da média móvel e lidando com os dados das 3 últimas temporadas na base de teste dos modelos. As variáveis com maior significância foram as de vitórias, derrotas, *plus-minus* e suas derivações.

Palavras-chave: NBA, Modelagem, Previsão, Regressão Logística, *Lasso*, *Ridge*, *Random Forest*, Árvore Binária, Curva ROC, Matriz de Confusão.

Abstract

This work had the objective of elaborating a model capable of predicting whether or not a team will achieve the NBA playoffs, based on up to 26 matches played.

The data was acquired from a database available on Kaggle and the official NBA website via Web scraping. The data is related to each match played, for both home and away team. Afterwards, variables were created based on feature engineering to collaborate on the model predictive capabilities. Such variables are related to opponent statistics, moving average of both teams statistics, with respect to the most recent matches played and opponent strength. A function was then created to construct a database with each teams statistics according to number of games to be studied. Next, another function was created to ajust the predictive models according to number of games played, number of games considered in the moving average and quantity of seasons used in the test database. This second function creates databases based on the aforementioned parameters. Afterwards, the test databases and training database were separated. The tecniques utilized were Lasso, Ridge, Random Forest and Binary Tree. Finally, a table was created with all the useful information in comparison of proposed models.

After analyses, it was verified that the model with best performance, according to this works objectives, was utilizing Random Forest, the first 15 games, 4 games for moving average and using the three most recent seasons as a test database. The variables with greatest significance were wins, losses, plus-minus and their derivations.

Key-words: NBA; Modelling; Prediction; Logistic Regression; Lasso; Ridge; Random Forest; Binary Tree; ROC Curve; Confusion Matrix.

Lista de Tabelas

1 TOP 20 modelos ordenados por F1 30
2 Frequência de modelos por qnt. de jogos de média móvel 31
3 Frequência de modelos por técnica 31
4 Melhores modelos de cada técnica utilizando 3 jogos de média móvel 32
5 Melhores modelos de cada técnica utilizando 4 jogos de média móvel 32

Lista de Quadros

1 Exemplo base de dados após primeira limpeza 26
2 Exemplo banco de dados gerado 27
3 Descrição das variáveis 40

Lista de Figuras

1 Exemplo árvore binária 21

2 Matriz de confusão 23

3 Correlograma das variáveis 29

4 Medidas dos modelos utilizando 3 jogos de média móvel e 3 temporadas para teste 32

5 Medidas dos modelos utilizando 4 jogos de média móvel e 3 temporadas para teste 33

6 Qnt. de FP, VP e nº de jogos do *Random Forest* 34

7 TOP 10 variáveis mais importantes do modelo com base na diminuição média da acurácia 35

8 Comportamento das variáveis mais significativas no modelo selecionado de acordo com a classificação ou não aos *playoffs* 36

Sumário

1 Introdução 17

2 Referencial Teórico 19

2.1 Regressão Logística. 19

2.1.1 Técnicas de penalização 19

2.1.1.1 *Lasso* 19

2.1.1.2 Regressão *Ridge* 20

2.2 Árvores Binárias 21

2.2.1 Medida de Entropia 22

2.2.2 Índice de Gini 22

2.2.3 *Random Forest* 22

2.3 Medidas 23

2.3.1 Matriz de confusão 23

2.3.2 Acurácia 24

2.3.3 Precisão 24

2.3.4 *Recall* e Sensibilidade 24

2.3.5 F1 ou F-score 24

2.4 Curva ROC 24

2.4.1 Método Índice de União (IU) 25

3 Metodologia 26

3.1 Banco de dados 26

3.2 Técnicas Modelagem 28

4 Resultados 29

4.1 Modelos 30

4.2 Modelo Escolhido. 35

5 Conclusão 38

6 Referências 39

7 Apêndice 40

1 Introdução

A *National Basketball Association* (NBA) é a maior liga de basquete no mundo e é disputada nos Estados Unidos da América (EUA) e no Canadá, onde as melhores equipes se enfrentam até que um time se consagre campeão. A criação da liga se deu a partir da união de duas organizações rivais, *Basketball Association of America* (BAA) e a *National Basketball League* (NBL), em 1949 (NELSON, 2009). No início da NBA, 17 franquias (equipes) participavam da liga, porém ao longo do tempo esse número foi se alterando, até que em 2004 chegou ao número de 30 equipes e isso se mantém até a atualidade. Desses 30 times, 29 estão localizados nos EUA e 1 encontra-se no Canadá. Estes são divididos em 2 grandes conferências (Leste e Oeste), sendo que em cada conferência há 15 franquias. Os maiores vencedores da NBA até os dias atuais são o *Los Angeles Lakers* e o *Boston Celtics*.

A competição é dividida em 3 etapas: temporada regular, *playoffs* e finais. A temporada regular, em sua normalidade, é composta por 82 jogos para cada equipe, com o equilíbrio de partidas dentro e fora de sua cidade de origem (casa). Após os 82 jogos disputados por cada equipe, as 6 primeiras classificadas de cada conferência avançam diretamente aos *playoffs*, enquanto as franquias classificadas entre 7º e 10º lugar de cada conferência se enfrentam em um esquema de repescagem, mais conhecido como "*play-in*", e depois mais 2 equipes de cada conferência se integram aos *playoffs*, totalizando 8 equipes de cada conferência nos *playoffs*. A importância de um time ir aos *playoffs*, além de ter a chance de disputar o título e ganhar uma grande premiação financeira, a equipe em geral (comissão técnica e jogadores) ganham mais experiência para os próximos anos e também aumenta o entrosamento da equipe (TARLOW, 2012). Os *playoffs* tem um modo de competição denominado "mata-a-mata", em que, separados em suas conferências, os times se enfrentam até que um time seja campeão de seu grupo. Por fim, os times campeões de suas conferências se enfrentam na final da NBA para decidir quem será o campeão geral.

A NBA dispõe de dados sobre as temporadas passadas que podem auxiliar na realização de diversos trabalhos. Pode-se citar como estudo de interesse a possibilidade de avaliar quem irá aos *playoffs* da NBA. Esse exemplo de pesquisa é capaz de orientar diversas pessoas e setores, como por exemplo as próprias equipes, pois com a informação de que já está classificada para ir aos *playoffs* ou estar perto de conseguir sua classificação ou que não há nenhuma chance de ir à próxima fase faz com que a equipe tome decisões em prol do time. Algumas dessas decisões podem ser: poupar jogadores importantes para que eles cheguem saudáveis à próxima fase e evitar possíveis lesões ou ainda perceber déficits em alguma área do time e fazer contratações de jogadores para suprir essas deficiências. As casas de apostas e seus apostadores também podem ser beneficiados com esse estudo,

no caso das casas de apostas terem uma base para calcular suas possíveis cotações e para os seus frequentadores terem um "norte"para apostas mais certeiras. Também os fãs tem interesse nessa previsão por proveito próprio e também por jogar jogos que simulam o torneio real. Com isso, neste trabalho, os dados adquiridos serão utilizados para obter modelos estatísticos com o objetivo de realizar predições referentes as chances de cada equipe avançar para a etapa de *playoffs*.

O objetivo geral desse trabalho é formular modelos a partir de algumas técnicas estatísticas visando prever se um time irá aos *playoffs* da NBA após ter jogado até 26 partidas. Deve-se levar em consideração que efetuar a previsão com estatísticas em menor tempo no campeonato é melhor, assim as equipes conseguem elaborar uma reconstrução do problema encontrado em um maior período e as apostas com as cotações mais elevadas podem render mais dinheiro aos apostadores. Para isso, será realizada a obtenção dos dados e as manutenções necessárias. Depois disso, os modelos preditivos serão ajustados e comparados pelos seus desempenhos. Também será analisado o comportamento dos times durante o período e estudado quais fatores com maiores influências.

2 Referencial Teórico

As seguintes técnicas estatísticas serão utilizadas nesse trabalho para a elaboração do modelo final.

2.1 Regressão Logística

Os modelos de regressão logística são utilizados para estudar os efeitos de variáveis preditoras em resultados categóricos e normalmente o resultado é binário (NICK; CAMPBELL, 2007), isto quer dizer que essa técnica relaciona respostas categóricas (binárias ou ordinais) com variáveis explicativas categóricas e/ou contínuas.

Assumindo C = 0,1, a regressão logística apresenta a seguinte forma paramétrica:

$$P(Y = 1|x) = \frac{e^{\beta_0+\sum_{i=1}^{d}\beta_i x_i}}{1 + e^{\beta_0+\sum_{i=1}^{d}\beta_i x_i}}, \tag{2.1.1}$$

A estimação dos coeficientes da regressão logística podem ser calculados pelo método de máxima verossimilhança. Com isso, dada uma amostra de i.i.d. $(X_1, Y_1), \ldots, (X_n, Y_n)$, a função de verossimilhança condicional nas variáveis, é

$$L(y; (x, \beta)) = \prod_{k=1}^{n} (P(Y_k = 1|x_k, \beta))^{y_k} (1 - P(Y_k = 1|x_k, \beta))^{1-y_k}$$

$$L(y; (x, \beta)) = \prod_{k=1}^{n} \left(\frac{e^{\beta_0+\sum_{i=1}^{d}\beta_i x_{k,i}}}{1 + e^{\beta_0+\sum_{i=1}^{d}\beta_i x_{k,i}}} \right)^{y_k} \left(\frac{1}{1 + e^{\beta_0+\sum_{i=1}^{d}\beta_i x_{k,i}}} \right)^{1-y_k} \tag{2.1.2}$$

O $L(y; (x, \beta))$ é maximizado para encontrar as estimativas dos coeficientes de β. Como na regressão linear, também é possível utilizar penalização para estimar os seus coeficientes. Diante disso, é esperado que quando reduzida a variância do estimador, há melhora no poder preditivo. (IZBICKI; SANTOS, 2020)

2.1.1 Técnicas de penalização

2.1.1.1 *Lasso*

A técnica *Lasso* busca achar um estimador de uma regressão que tem risco menor que o de mínimos quadrados, ou seja, reduzir a variância do estimador de mínimos

quadrados. A complexidade do modelo *Lasso* é medida pela norma do vetor $\sum_{j=1}^{d} |\beta_j|$, essa norma é menor para muitos coeficientes iguais a zero, também ela captura a ideia de que pequenas alterações nos valores de β não alterarão indevidamente a complexidade do modelo resultante (já que suas previsões são quase idênticas).

Com a técnica de *Lasso*, busca-se por:

$$\hat{\beta}_{L_1,\lambda} = arg\ min_{\beta} \sum_{k=1}^{n} \left(y_k - \beta_0 - \sum_{j=1}^{d} \beta_j x_{k,j} \right)^2 + \lambda \sum_{j=1}^{d} |\beta_j|, \tag{2.1.3}$$

em que L_1 indica que será usada a norma $\sum_{j=1}^{d} |\beta_j|$ para medir a esparsidade de um vetor β.

Obs: Quando todos os coeficientes são diferentes de zero, o *Lasso* torna-se idêntico ao estimador de mínimos quadrados.

Quando λ é grande,

$$\sum_{k=1}^{n} \left(y_k - \beta_0 - \sum_{j=1}^{d} \beta_j x_{k,j} \right)^2 + \lambda \sum_{j=1}^{d} |\beta_j| \approx \lambda \sum_{j=1}^{d} |\beta_j|, \tag{2.1.4}$$

assim tem-se que o estimador do *Lasso* tem variância próxima de zero, mas um viés muito alto. Diante disso, adicionando a penalização $\lambda \sum_{j=1}^{d} |\beta_j|$ leva a variância ser menor do que a estimativa dada pelo método dos mínimos quadrados. (IZBICKI; SANTOS, 2020)

A escolha de λ geralmente é realizada por validação cruzada. Ou seja, estimamos uma perda quadrática, $R(g\lambda)$, para cada λ de interesse e, então, escolhemos o λ que leva ao melhor modelo selecionado. Sendo g uma função de predição $\mathbb{R}^d \to \mathbb{R}$.

2.1.1.2 Regressão *Ridge*

A ideia da Regressão *Ridge* é semelhante ao *Lasso*, ela procura o modelo que minimize a soma do erro quadrático médio de g com uma medida de sua complexidade. O estimador pode ser encontrado da seguinte forma:

$$\hat{\beta}_{L_2,\lambda} = arg\ min_{\beta} \sum_{k=1}^{n} \left(y_k - \beta_0 - \sum_{j=1}^{d} \beta_j x_{k,j} \right)^2 + \lambda \sum_{j=1}^{d} |\beta_j^2|, \tag{2.1.5}$$

em que L_2 indica que será usada a norma $\sum_{j=1}^{d} \beta_j^2$ para medir a complexidade de um vetor β.

Diferentemente do *Lasso*, a Regressão *Ridge* utiliza uma solução analítica que é representada por:

$$\hat{\beta}_{L_2,\lambda} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_d) = (\mathbb{X}^T\mathbb{X} + \lambda\mathbb{I}_0)^{-1}\mathbb{X}^T\mathbb{Y}, \tag{2.1.6}$$

onde $\mathbb{I}_0$ representa uma matriz identidade $(d+1) \times (d+1)$ de maneira que $\mathbb{I}_0(1,1) = 0$.

A Regressão *Ridge* diminui a variância dos estimadores da regressão pois "escolhe"os coeficientes β estimados pela regressão, mesmo não incluindo soluções com zeros como no *Lasso*. Com isso, o seu viés é maior, mesmo a variância sendo menor (IZBICKI; SANTOS, 2020). O λ selecionado terá a função de controlar o balanço do viés-variância. Isso é possível ser realizado via validação cruzada.

2.2 Árvores Binárias

Método flexível e não paramétrico, permite entender quais variáveis são mais significativas. Não necessita possuir um conhecimento prévio de todos os elementos estudados. A técnica consiste em um diagrama no formato de árvore, onde suas ramificações (em 2 segmentos) possuem probabilidades de acordo com a frequência de cada evento ocorrer. A intenção desse formato é de, ao longo das escolhas no processo das divisões, proporcionar uma melhor previsão do que a divisão anterior (SOBRAL, 2014).

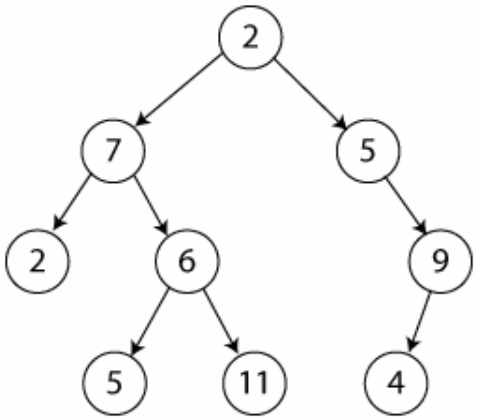

Figura 1: Exemplo árvore binária
Fonte: (FEOFILOFF, 2018)

As partições na árvore segue o critério do proveito da característica para a classificação. A partição é feita a partir do ganho de informação a cada ramo, então a característica escolhida será a que converte em mais informação, e a partir dai, um novo processo de partição é gerado.

Os critérios de avaliação para escolher o ramo com maior informação serão a Medida de Entropia e o Índice de Gini:

2.2.1 Medida de Entropia

A Entropia busca entender a desordem nos dados, e com isso descobrir quanto aquele dado é significativo para o que se busca. Ela é calculada da seguinte forma:

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \tag{2.2.1}$$

Sendo:

- S = conjunto de exemplos;
- i = decisões
- p_i = proporção de exemplos em S cuja decisão é i

2.2.2 Índice de Gini

O Índice de Gini tem como objetivo medir a heterogeneidade dos dados, e com isso medir a impureza de um nó. Sua forma tem como p_i a frequência relativa de cada classe em cada nó e c como a quantidade de classes.

$$G = 1 - \sum_{i=1}^{c} p_i^2 \tag{2.2.2}$$

O índice de Gini varia de 0 a 1. Quando o resultado é mais próximo de zero, representa igualdade entre as características dos ramos e que possui menos impurezas, e quando a resposta for mais próximo de 1, significa que há diferença das características e mais impurezas.

2.2.3 *Random Forest*

Random Forest, ou Floresta Aleatória, é uma técnica baseada em Árvores Binárias. Esse método constrói diversas árvores independentes, sorteando aleatoriamente variáveis e as obervações de análise. A partir dos resultados dessas árvores, e sabendo que esse trabalho tem como finalidade classificar as previsões, será realizada uma "votação"para decidir qual valor obteve uma maior frequência, e assim, ser selecionado.

O *overfitting* é um problema de estimação que consiste em modelos que são ajustados muito bem com um banco de dados específico, mas que não apresentam a mesma eficácia para novos resultados. Apesar da técnica de *Random Forest* ser mais demorada

do que uma árvore de classificação, ela tem a prevenção do erro de *overfitting*.

As previsões feitas por uma árvore binária são eficientes, porém limitadas, pois ela é "treinada"para um certo banco de dados. E, caso haja uma mudança, talvez ela não seja tão eficiente (BIAU, 2012). Ela faz os cortes para encontrar ótimos locais, isso quer dizer que ela faz uma previsão sem considerar contextos futuros. Já a técnica de *Random Forest* busca por ótimos globais e, com a visualização de diversos cenários, é capaz de remover uma variável que foi selecionada como significante no início do corte e permite selecionar uma outra variável entre as mais importantes, podendo ser uma solução para se obter melhores previsões.

2.3 Medidas

Segundo (SHUNG, 2018), a precisão não é a única métrica em que os cientistas de dados devem analisar na hora de escolher o modelo. Há outras métricas que também ajudam a escolher o melhor modelo de previsão de acordo com o objetivo do trabalho.

2.3.1 Matriz de confusão

A matriz de confusão apresenta as frequências de classificação para cada objeto do modelo.

		Valor Predito	
		Não	**Sim**
Real	**Não**	Verdadeiro Negativo (TN)	Falso Positivo (FP)
	Sim	Falso Negativo (FN)	Verdadeiro Positivo (TP)

Figura 2: Matriz de confusão

- **Verdadeiro Positivo (TP)**: Os valores são verdadeiros e foram classificados como verdadeiros;
- **Verdadeiro Negativo (TN)**: Os valores são falsos e foram classificados como falsos;
- **Falso Positivo (FP)**: Os valores são falsos e foram classificados como verdadeiros;
- **Falso Negativo (FN)**: Os valores são verdadeiros e foram classificados como falsos.

2.3.2 Acurácia

$$Acurácia = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.3.1}$$

A medida de Acurácia apresenta a proporção de predições corretas feitas pelo modelo.

2.3.3 Precisão

$$Precisão = \frac{TP}{TP + FP} \tag{2.3.2}$$

A medida de Precisão avalia a quantidade de verdadeiros positivos em relação a todos que foram classificados como positivos. (SHUNG, 2018)

2.3.4 *Recall* e Sensibilidade

$$Recall = \frac{TP}{TP + FN} \tag{2.3.3}$$

A medida de *Recall* apresenta a quantidade de sucessos positivos em relação a todos os acertos. (SOUZA, 2019)

2.3.5 F1 ou F-score

$$F1 = 2 \times \frac{Precisão \times Recall}{Precisão + Recall} \tag{2.3.4}$$

A medida de F1 simboliza o equilíbrio entre as medidas de Precisão e *Recall*.

2.4 Curva ROC

A curva ROC é uma técnica utilizada para medir a precisão diagnóstica do teste e fornece o ponto de ótimo (ponto de corte) ideal para o teste. Ela é apresentada a partir de um gráfico $sensibilidade \times 1 - especificidade$ para todos os valores de corte possíveis entre casos e controles. Um ponto de corte denominado como ótimo refere-se ao ponto que melhor classifica os indivíduos.

Além de usar as medidas de especificidade e de sensibilidade, também é utilizada a medida AUC (Área sob a curva). O valor de AUC varia entre 0 e 1. Quando o modelo apresenta um AUC igual a 0, isso mostra que 100% das previsões estão incorretas, e quando o AUC é igual a 1, mostra que 100% das previsões estão corretas.

2.4.1 Método Índice de União (IU)

Segundo o artigo *Defining an Optimal Cut-Point Value in ROC Analysis: An Alternative Approach* (UNAL, 2017), o método Índice de União tem como objetivo encontrar os valores máximos de sensibilidade e especificidade que sejam simultaneamente próximos ao valor de AUC. Isso pode ser dado pela seguinte expressão:

$$IU(c) = (|Se(c) - AUC| + |Sp(c) - AUC|), \tag{2.4.1}$$

O ponto de corte $\hat{c}_{IU}$ minimiza o $IU(c)$ da expressão apresentada em 2.4.1, será denominado o ponto de ótimo para aquele modelo.

3 Metodologia

O *software R* (versão 4.1.0 (2021-05-18)) e o *software Python* foram utilizados neste trabalho para manutenção no banco de dados e análises estatísticas.

3.1 Banco de dados

A primeira etapa do estudo ocorreu a extração dos dados da internet. O primeiro banco de dados foi retirado da tabela "*Game*"do site *Kaggle* (WALSH, 2021), esta possui informações de todos os jogos da NBA desde a temporada 1946-1947 até a temporada 2020-2021. Cada observação nessa base de dados representa um jogo e tem as informações e estatísticas do confronto, da equipe mandante e da equipe adversária. A base de dados original tem 62448 linhas e 149 colunas. Nesse primeiro banco de dados havia dados faltantes da temporada 2020-2021 e não possuía informações sobre a temporada 2021-2022. Com isso, foi necessário extrair um segundo banco de dados com as informações fornecidas pelo site oficial da NBA. A extração desses dados foi realizada a partir da técnica de *Web scraping*, que é uma coleta de dados a partir de uma "raspagem"de informações em *websites*, e esses dados apresentam as estatísticas gerais de cada time em um jogo, semelhante ao apresentado no primeiro banco de dados.

Após a unificação dos dois bancos de dados foi necessário realizar uma limpeza nessa base. Colunas e linhas desnecessárias foram removidas, foi filtrado apenas os jogos das temporadas de 2004-2005 até a temporada 2021-2022 e por fim executado algumas manutenções necessárias. O banco de dados ficou com 21578 linhas e 55 colunas. O Quadro 1 contém uma amostra de como ficou a base de dados após a limpeza:

Quadro 1: Exemplo base de dados após primeira limpeza

SEASON	GAME_DATE	TEAM_NAME_HOME	PTS_HOME	TEAM_NAME_AWAY	PTS_AWAY
2004-2005	2004-11-02	Detroit Pistons	87	Houston Rockets	79
2004-2005	2004-11-02	Dallas Mavericks	107	Sacramento Kings	98
2004-2005	2004-11-02	Los Angeles Lakers	89	Denver Nuggets	78
2004-2005	2004-11-03	San Antonio Spurs	101	Sacramento Kings	85
2004-2005	2004-11-03	Toronto Raptors	95	Houston Rockets	88

Uma função foi criada para que retornasse um banco de dados com as estatísticas de cada time referente à quantidade de jogos iniciais de cada temporada. Essa base contém 540 linhas e 32 colunas. O Quadro 2 apresenta um exemplo desse banco de dados referente aos 25 primeiros jogos da temporada 2004/2005:

Quadro 2: Exemplo banco de dados gerado

SEASON	TEAM_NAME	CONF	PLAYOFFS	WINS	LOSES
2004-2005	Atlanta Hawks	EAST	0	5	20
2004-2005	Boston Celtics	EAST	1	12	13
2004-2005	Charlotte Bobcats	EAST	0	7	18
2004-2005	Chicago Bulls	EAST	1	9	16
2004-2005	Dallas Mavericks	WEST	1	16	9

Para uma melhor predição e exploração dos dados, foram criadas variáveis pelo processo de *feature engineering*, que são variáveis ainda não explícitas e que serão obtidas através da combinação de outras covariáveis já existentes.

As primeiras variáveis elaboradas pela *feature engineering* foram em relação as estatísticas dos times oponentes enfrentados por um determinado time naquele período em que foi estabelecido. Essas variáveis apresentam como foi a dificuldade no trajeto que o time estudado enfrentou no tempo observado. Após a criação desses fatores, foram elaborados outros indicadores relacionados a média móvel dessas estatísticas já existentes. As variáveis de média móvel podem ser explicadas da seguinte maneira: caso o estudo seja de 25 jogos e média móvel de 5 jogos, então os resultados dos itens de média móvel resultam nas estatísticas dos últimos 5 jogos desses 25 jogos abordados. Esses indicadores de média móvel apresentam o desempenho do time nas últimas rodadas jogadas e fornece qual é a situação que ele está enfrentando naquele momento da competição, se o time está em ascensão, se começou uma decadência no desempenho ou se ele mantém um padrão de competitividade desde o início do campeonato.

Além dessas variáveis já apresentadas, também foi criada uma variável pela *feature engineering* para avaliar o nível dos times que uma determinada equipe enfrentou até aquele momento da competição, pois algumas equipes podem ter sido favorecidas de estar em uma sequência de jogos mais fáceis enquanto outras equipes em uma sequência de jogos mais difíceis. Com isso, essa variável é calculada como a média das proporções de vitórias das equipes adversárias que o time enfrentou até aquele momento da competição.

As principais variáveis do banco de dados com os indicadores de cada equipe referente ao número de jogos iniciais de cada temporada são:

- ***SEASON***: Indica a temporada estudada;
- ***TEAM_NAME***: Identifica a equipe analisada;
- ***PLAYOFFS***: Variável resposta que mostra se o time classificou aos *playoffs*;
- ***WINS***: Número de vitórias;

- ***LOSES***: Número de derrotas.

As outras variáveis e suas descrições encontram-se no Quadro 3 na seção Apêndice.

3.2 Técnicas Modelagem

Uma função foi criada para que fossem formadas diversas bases de dados variando o número de jogos, o número de jogos considerado para o cálculo de médias móveis e a quantidade de temporadas que seriam utilizadas nas bases de testes. A partir de cada banco de dados gerado, foi separado o banco de treinamento e o banco de teste. Os dados para teste sempre referiram-se as temporadas mais recentes, ou seja, quando o modelo utilizar duas temporadas testes, essas temporadas foram de 2021-2022 e 2020-2021, e assim por diante. Com isso, informações e medidas dos modelos foram calculadas e apresentadas para que ao final o modelo escolhido fosse o melhor que cumprisse os objetivos deste trabalho.

As técnicas estatísticas utilizadas para as elaborações dos modelos foram de Árvores binárias, *Random Forest*, *Lasso* e Regressão *Ridge*, e os pacotes utilizados no *software R* para a modelagem dessas técnicas foram, respectivamente, *rpart*, *randomForest* e *glmnet* (para as duas últimas técnicas citadas).

A primeira etapa da modelagem destinou-se a encontrar os valores preditos em cada técnica. Com a base de treinamento separada, os valores preditos de cada técnica foram encontrados e armazenados. Após isso, verificou-se o ponto de ótimo da curva ROC pelo método Índice de União (o pacote *pROC* foi utilizado para a obtenção desse valor). Logo após foram realizadas as classificações com base nos dados de teste prevendo se o time iria avançar aos *playoffs* ou não. A partir dessas classificações e dos valores reais observados nos dados de teste, foi elaborada a matriz de confusão.

As matrizes de confusão construídas (utilizou-se do pacote *caret* para a elaboração das matrizes) auxiliam a obtenção de medidas que permitem a comparação dos modelos propostos, por exemplo, os valores de acurácia, sensitividade, precisão, F1. Uma grande tabela foi elaborada utilizando essas medidas para cada modelo construído, assim sendo possível analisar os diferentes cenários com as variações do número de jogos, quantidade de jogos para média móvel e número de temporadas utilizadas para a base de teste de cada técnica. Com base nisso, a seleção de um melhor modelo foi realizada ao final dessa análise.

4 Resultados

Considerando o banco de dados com as estatísticas de todos os jogos de cada uma das 18 temporadas, 2004-2005 até 2021-2022, referentes a cada time, tem-se 101 variáveis, já apresentadas no Quadro 3. Dentre essas variáveis, duas são categóricas, *SEASON* e *TEAM_NAME*, entretanto são utilizadas apenas para identificação da temporada e do time que está sendo tratado. Outra variável é binária, a *PLAYOFFS*, que será considerada como variável resposta dentro da modelagem e as demais são numéricas.

A primeira parte desse estudo é selecionar quais técnicas serão utilizadas de acordo com os dados adquiridos. Um dos principais problemas que podem ocorrer durante a modelagem é em relação a multicolinearidade, para evitar esse problema foi construído o correlograma de todas as 98 variáveis numéricas do banco.

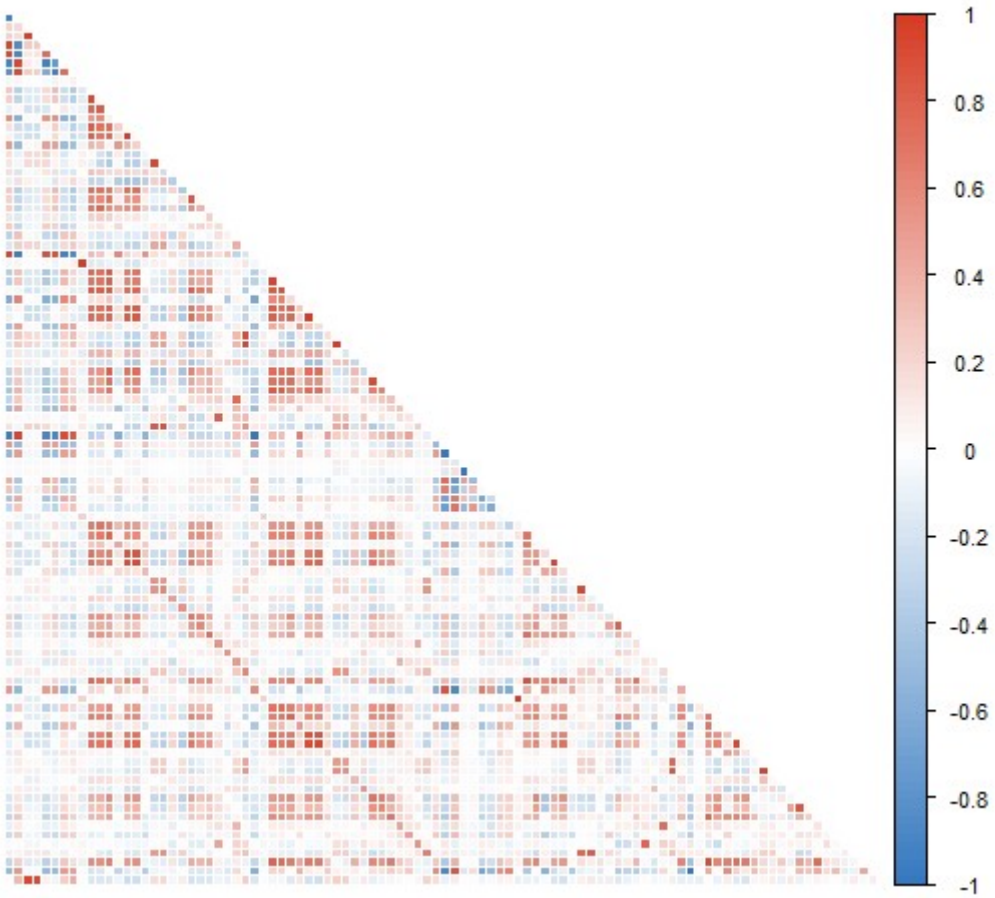

Figura 3: Correlograma das variáveis

A Figura 3 mostra a correlação entre as variáveis, duas a duas, utilizando o método da correlação de *Pearson*. Quanto mais vermelho as variáveis, mais elas são correlacionadas positivamente e quanto mais azuis, mais correlacionadas negativamente, e quando mais brancas representam a não correlação entre as variáveis.

Observa-se após a análise do gráfico ilustrado na Figura 3 que existem correlações fortes entre diversas variáveis que foram utilizadas como preditoras, isso quer dizer que, há grandes indícios de ocorrerem problemas de multicolinearidade. Para contornar essa adversidade foram utilizadas as técnicas de *Lasso* e *Rigde* ao invés da Regressão Logística, pois essas técnicas usam um mecanismo de penalização em coeficientes com alto grau de correlação. Também foram utilizadas as técnicas de Arvore binária e *Random Forest* pois elas são robustas para esse tipo de problema.

4.1 Modelos

Para a escolha de um modelo que satisfizesse os objetivos deste trabalho, foi gerada uma tabela com as medidas que auxiliassem na seleção do mais adequado. Para isso, os modelos foram elaborados para cada número de jogos de 8 a 26, variando entre 3 a 10 a quantidade de jogos calculados para a média móvel. Também foi realizada a alternância da quantidade de temporadas que seriam utilizadas para teste dos modelos, variando de uma a quatro temporadas.

Tabela 1: TOP 20 modelos ordenados por F1

Técnica	**Qnt. Jogos**	**Qnt.Jogos Média Móvel**	**Qnt. Temporadas Teste**	**FN**	**FP**	**VP**	**VN**	**Acurácia**	**Sensibilidade**	**Precisão**	**F1**
Árvore Binária	25	6	1	3	2	13	12	0.835	0.812	0.867	0.839
Random Forest	24	8	3	11	4	37	38	0.838	0.771	0.902	0.831
Random Forest	23	10	3	9	7	39	35	0.823	0.812	0.848	0.830
Random Forest	26	6	3	8	9	40	33	0.810	0.833	0.816	0.825
Árvore Binária	22	4	1	2	4	14	10	0.795	0.875	0.778	0.824
Árvore Binária	22	5	1	2	4	14	10	0.795	0.875	0.778	0.824
Random Forest	16	5	3	11	5	37	37	0.826	0.771	0.881	0.822
Árvore Binária	18	5	3	11	5	37	37	0.826	0.771	0.881	0.822
Random Forest	26	10	3	11	5	37	37	0.826	0.771	0.881	0.822
Random Forest	24	4	3	9	8	39	34	0.811	0.812	0.830	0.821
Random Forest	24	10	3	9	8	39	34	0.811	0.812	0.830	0.821
Random Forest	15	4	3	10	7	38	35	0.812	0.792	0.844	0.817
Random Forest	24	3	3	10	7	38	35	0.812	0.792	0.844	0.817
Random Forest	24	9	3	10	7	38	35	0.812	0.792	0.844	0.817
Random Forest	26	3	3	8	10	40	32	0.798	0.833	0.800	0.816
Árvore Binária	18	10	2	8	3	24	25	0.821	0.750	0.889	0.814
Random Forest	17	5	3	11	6	37	36	0.814	0.771	0.860	0.813
Random Forest	17	8	3	11	6	37	36	0.814	0.771	0.860	0.813
Random Forest	17	9	3	11	6	37	36	0.814	0.771	0.860	0.813
Random Forest	16	3	2	6	6	26	22	0.799	0.812	0.812	0.812

Após a obtenção dos resultados das medidas para cada modelo proposto, verificou-se que, a partir do ordenamento dos valores de F1, a maioria dos modelos com melhores desempenhos utilizaram dados de 3 temporadas para teste, como pode-se perceber na Tabela 1. O número de temporadas para teste não afeta a performance dos modelos, mas

apresentam o desempenho em relação ao *overfitting*. Isso acontece porque quanto mais temporadas tiverem sido utilizadas na base de teste, melhor será a qualidade para verificar como o modelo está se comportando em relação ao *overfitting*, já que esse problema trata de modelos que se adequam muito bem ao banco de treinamento, mas não apresentam o mesmo desempenho para novos dados.

Depois de selecionar apenas os modelos que foram aplicados dados com 3 temporadas de teste, estudou-se com quantos jogos de média móvel foram obtidos os melhores resultados de F1. Para isso, foi realizada, para cada número de jogos aplicados no processamento dos modelos, uma seleção dos 8 modelos que conseguiram os maiores valores de F-score. Em seguida, foi analisado qual número de jogos de média móvel teve maior frequência nesses melhores modelos.

Tabela 2: Frequência de modelos por qnt. de jogos de média móvel

Qnt. Jogos de Média Móvel	3	4	5	6	7	8	9	10
Frequência	20	21	17	18	15	17	12	16

Pelos resultados da Tabela 2, é notório que os modelos que foram utilizados 3 e 4 jogos para o cálculo de médias móveis apresentaram a maior quantidade de modelos com melhores resultados. A análise para a Tabela 2 teve como base a quantidade de jogos de 8 a 26, mas também realizou-se testes com outras variações de número de jogos que comprovaram que utilizando 3 e 4 jogos para a média móvel possuía uma maior quantidade de modelos com os melhores resultados. Com base nesse mesmo ranking, também foram analisadas as técnicas mais frequentes. De acordo com a Tabela 3, nota-se a técnica de *Random Forest* teve maior número de destaques dentre elas.

Tabela 3: Frequência de modelos por técnica

Técnica	Árvore Binária	Lasso	Random Forest	Ridge
Frequência	23	11	92	10

Diante do exposto, o modelo selecionado deve conter 3 ou 4 jogos considerados para o cálculo das médias móveis e dados de 3 temporadas para teste do modelo final. A seguir estão expostas as Tabelas 4 e 5 com os resultados das medidas dos 2 melhores modelos de cada técnica que seguem as condições citadas e as Figuras 4 e 5 mostram o comportamento das medidas de cada técnica de acordo com a quantidade de jogos utilizadas para realizar as previsões.

Tabela 4: Melhores modelos de cada técnica utilizando 3 jogos de média móvel

Técnica	Qnt. Jogos	Qnt.Jogos Média Móvel	Qnt. Temporadas Teste	FN	FP	VP	VN	Acurácia	Sensibilidade	Precisão	F1
Random Forest	24	3	3	10	7	38	35	0.812	0.792	0.844	0.817
Random Forest	26	3	3	8	10	40	32	0.798	0.833	0.800	0.816
Árvore Binária	24	3	3	11	8	37	34	0.790	0.771	0.822	0.796
Árvore Binária	15	3	3	12	7	36	35	0.792	0.750	0.837	0.791
Ridge	24	3	3	15	3	33	39	0.808	0.688	0.917	0.786
Ridge	10	3	3	12	9	36	33	0.768	0.750	0.800	0.774
Lasso	25	3	3	16	3	32	39	0.798	0.667	0.914	0.771
Lasso	10	3	3	14	7	34	35	0.771	0.708	0.829	0.764

Tabela 5: Melhores modelos de cada técnica utilizando 4 jogos de média móvel

Técnica	Qnt. Jogos	Qnt.Jogos Média Móvel	Qnt. Temporadas Teste	FN	FP	VP	VN	Acurácia	Sensibilidade	Precisão	F1
Random Forest	24	4	3	9	8	39	34	0.811	0.812	0.830	0.821
Random Forest	15	4	3	10	7	38	35	0.812	0.792	0.844	0.817
Árvore Binária	22	4	3	7	15	41	27	0.749	0.854	0.732	0.788
Árvore Binária	15	4	3	10	12	38	30	0.753	0.792	0.760	0.776
Ridge	24	4	3	16	3	32	39	0.798	0.667	0.914	0.771
Ridge	25	4	3	16	3	32	39	0.798	0.667	0.914	0.771
Lasso	24	4	3	16	3	32	39	0.798	0.667	0.914	0.771
Lasso	25	4	3	16	3	32	39	0.798	0.667	0.914	0.771

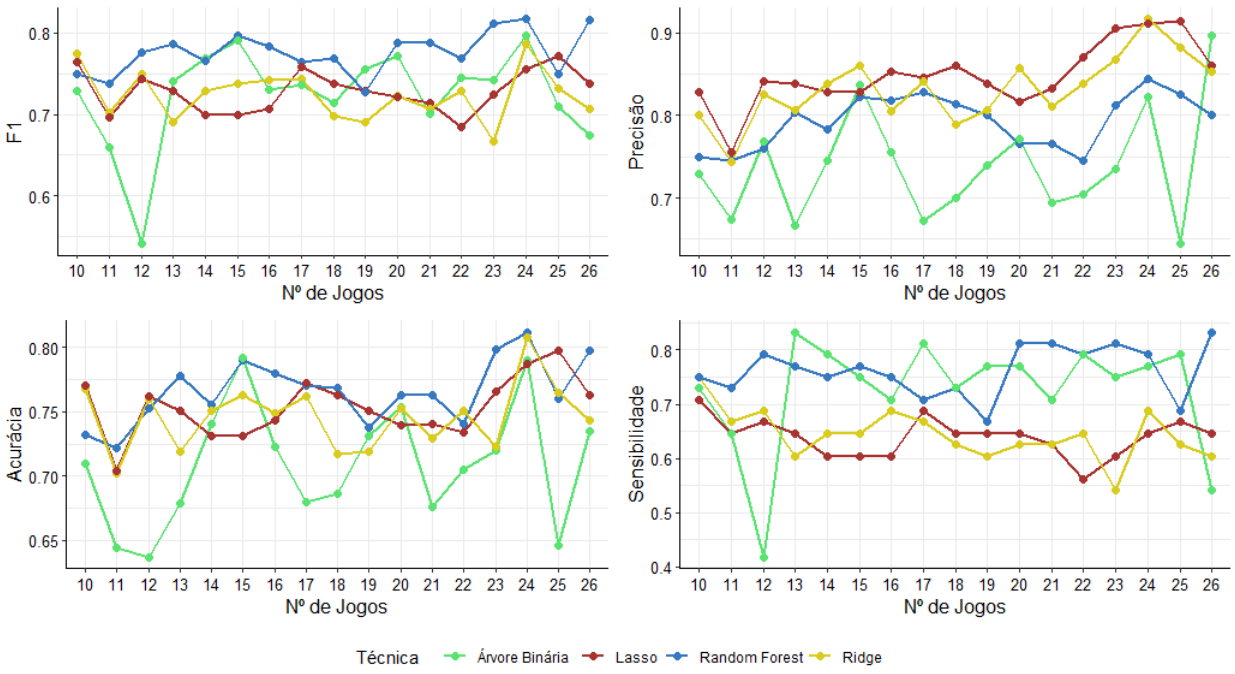

Figura 4: Medidas dos modelos utilizando 3 jogos de média móvel e 3 temporadas para teste

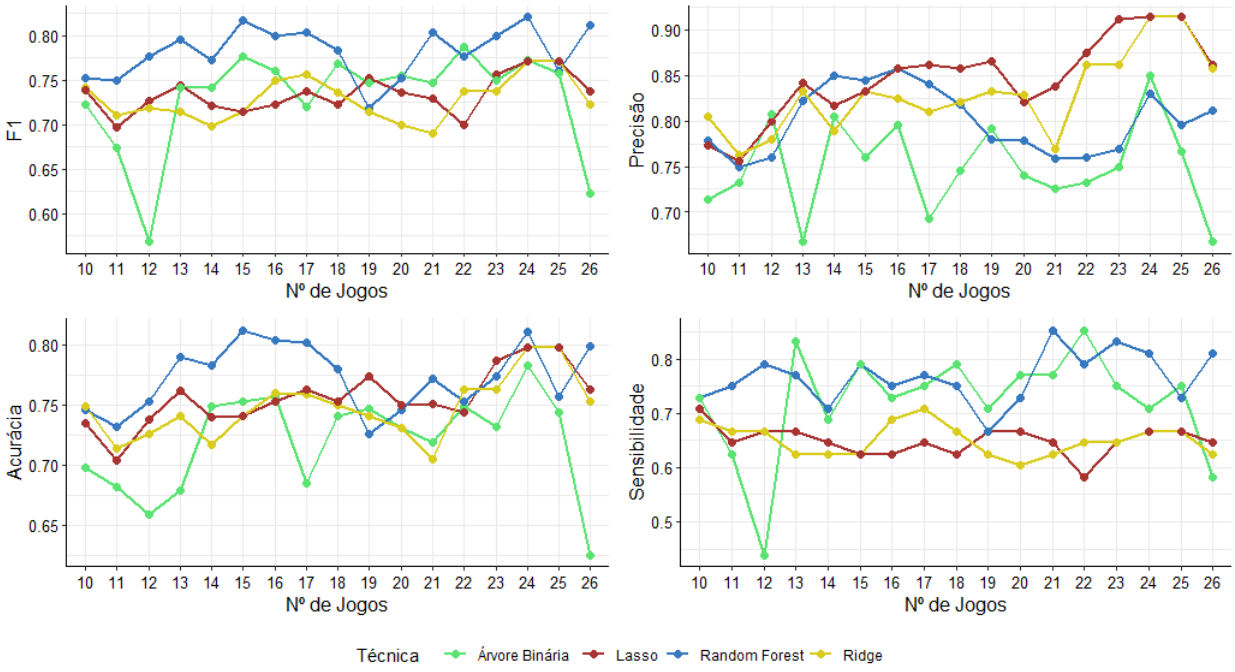

Figura 5: Medidas dos modelos utilizando 4 jogos de média móvel e 3 temporadas para teste

Com base nas Tabelas 4 e 5 e nas Figuras 4 e 5, percebe-se que as medidas das técnicas de *Lasso*, *Ridge* e *Random Forest* variaram pouco ao longo dos jogos, enquanto a técnica de Árvore Binária tem uma oscilação significativa. As técnicas *Lasso* e *Ridge* obtiveram melhores desempenhos na precisão, mas os piores na sensibilidade em razão da forte presença de multicolinearidade nos bancos. A *Random Forest* desempenhou bem na medida de F1, sendo superior às demais técnicas em quase todas as quantidades de jogos analisadas.

Comparando os resultados de acordo com os números de jogos para o cálculo de média móvel, é possível identificar que os resultados das medidas utilizando 4 jogos de média móvel em menor quantidade jogos têm melhores performances em relação aos modelos que utilizaram 3 jogos para a média móvel.

Portanto, o modelo que deve ser selecionado foi o que operou com a técnica de *Random Forest*, utilizando 4 jogos de média móvel. Como o objetivo principal desse trabalho é prever a classificação dos times aos *playoffs*, então o modelo deve atingir um baixo número de falsos positivos e uma grande quantidade de verdadeiros positivos. Para essa análise final, tem-se na Figura 6 para auxiliar nesse estudo.

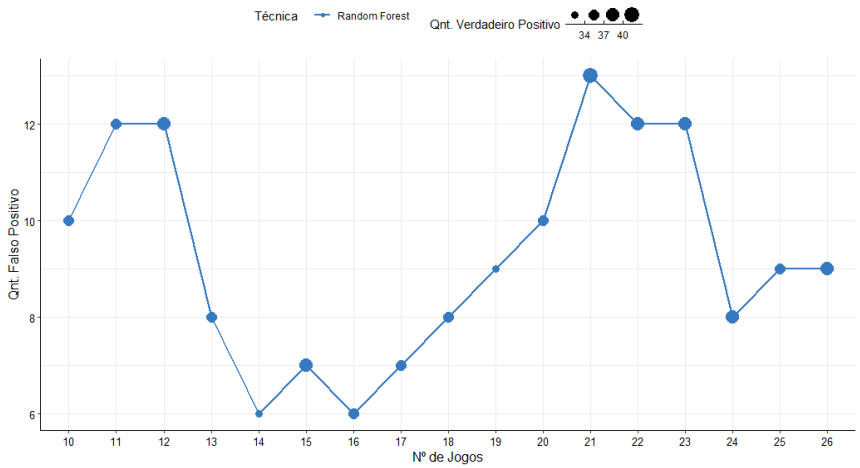

Figura 6: Qnt. de FP, VP e nº de jogos do *Random Forest*

A Figura 6 apresenta que utilizando 14 a 17 jogos, a técnica de *Random Forest* obtém uma pequena quantidade de falsos positivos, e dentre esses jogos citados, utilizando 15 e 16 jogos conseguiu-se uma grande parcela de verdadeiros positivos. Voltando na Figura 5 é perceptível que essa técnica para essas quantidades de jogos obtiveram bons valores para as medidas em análise. O modelo que utiliza 15 jogos se destaca, pois mesmo não sendo o modelo com menor número de falsos positivos, ele apresentou ótimos resultados de previsão de acordo com as medidas de precisão, sensibilidade e acurácia. Considerando que efetuando essa previsão o quanto antes sendo melhor, o modelo escolhido para esse trabalho foi utilizando a técnica de *Random Forest*, as estatísticas dos primeiros 15 jogos de cada equipe e utilizando os 4 últimos jogos desses 15 para o cálculo da média móvel.

4.2 Modelo Escolhido

Com o modelo selecionado, é importante verificar e entender quais variáveis foram mais importantes para as previsões do modelo.

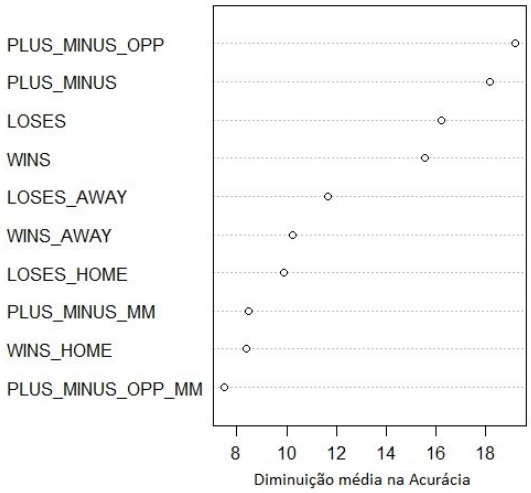

Figura 7: TOP 10 variáveis mais importantes do modelo com base na diminuição média da acurácia

O gráfico da Figura 7 apresenta quais variáveis foram as mais importantes do modelo, sendo apresentadas de cima para baixo. A partir dela, pela diminuição média da acurácia, é possível saber quanto o modelo perde sua capacidade preditiva com as retiradas de variáveis. Percebe-se que quando uma variável que está no topo do gráfico é eliminada o poder de previsão é muito reduzido, mas o impacto no modelo é menor considerando a remoção de um atributo do canto inferior.

Dessa forma, fica evidente a grande importância das variáveis *PLUS_MINUS_OPP*, *PLUS_MINUS*, *LOSES* e *WINS* para o modelo final. Nota-se que essas variáveis são complementares entre elas (PLUS_MINUS_OPP e PLUS_MINUS, LOSES e WINS). Em várias técnicas elas não estariam juntas no modelo, mas na *Random Forest* isso é possível acontecer porque a técnica seleciona as variáveis que obtiveram maiores frequências nas árvores geradas, em busca de ótimos globais.

É possível verificar que as variáveis mais significativas do modelo estão ligadas ao número de vitórias e derrotas dos times, e também na performance a cada jogo, a partir da diferença dos placares contra seus adversários. Esse resultado demonstra que a equipe deve buscar seu melhor desempenho, mesmo nas partidas em que é derrotado, pois caso ela perca por poucos pontos, comprova que não obteve um desempenho abaixo. Então naquele jogo, o time não alcançou vitória, mas ainda pode pensar na classificação aos *playoffs*. O mesmo vale para as vitórias por pouca diferença de pontos, que não afirma se o time teve uma boa atuação, e por consequência, não tendo maiores chances de que ir

aos *playoffs* se continuar com esse comportamento.

Nota-se também que as variáveis sobre vitórias e derrotas nos jogos fora de casa tem maior significância do que os jogos dentro de casa. Juntamente a isso, as únicas variáveis dentre as 10 mais importantes que utilizaram o processo de *feature engineering* estão relacionadas às diferenças de pontuação utilizando a média móvel, tanto da equipe estudada como do adversário. Portanto, deve ser levado em consideração o momento do time naquela fase da competição para determinar uma possível classificação para a próxima fase.

Para melhor entendimento das 10 variáveis mais importantes do modelo, foi realizada análise descritiva para verificar como as estatísticas se comportam entre os times que foram aos *playoffs* e os times que não foram aos *playoffs*. Na Figura 8 foram apresentados resultados de testes estatísticos para verificar se há diferença das médias entre as duas populações. Foram utilizados os testes de *Shapiro-Wilk* (Distribuição normal) e *Levene* (Homogeneidade) para verificar se os pressupostos eram cumpridos para a utilização do teste *T*. Caso esses pressupostos não ocorressem, foi utilizado o teste de *Mann-Whitney*. Todos os testes utilizaram um nível de significância de 0.05 para as tomadas de decisões.

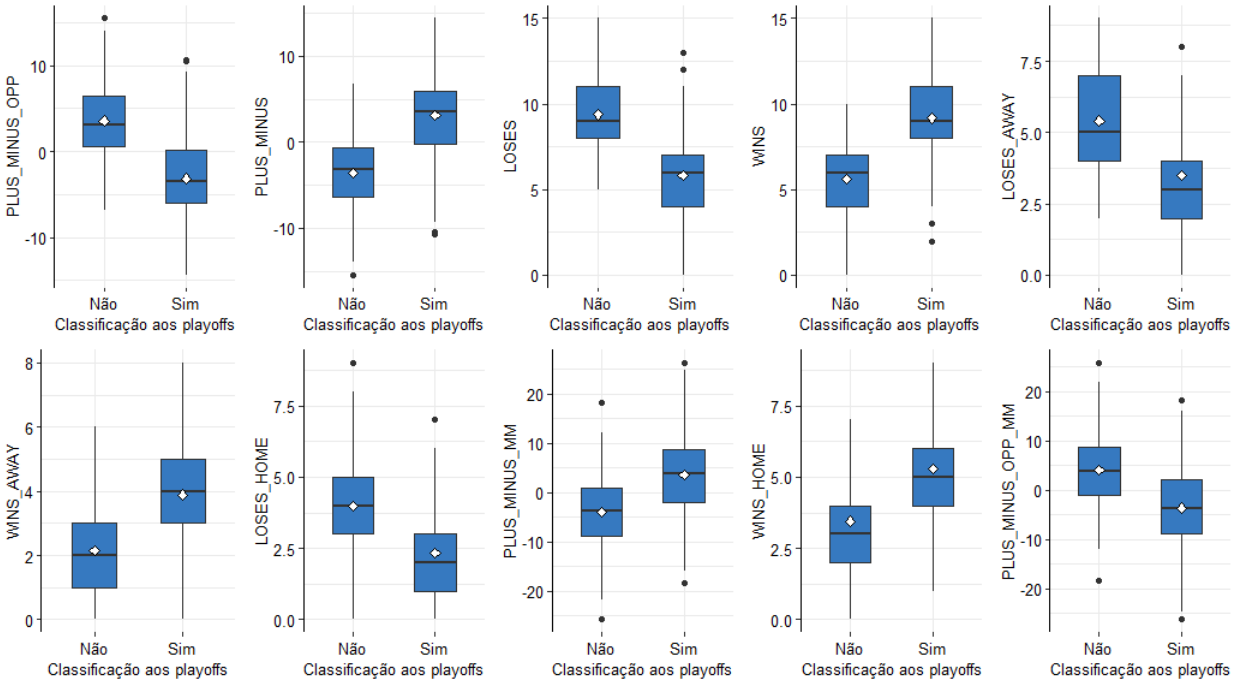

Figura 8: Comportamento das variáveis mais significativas no modelo selecionado de acordo com a classificação ou não aos *playoffs*

Após análises dos gráficos e dos resultados dos testes, percebe-se graficamente que as variáveis que favorecem o time estudado, como vitórias e *PLUS_MINUS* com valores mais altos, estão mais presentes nos times que conseguiram sua classificação aos *playoffs*, enquanto as variáveis que desfavorecem as equipes, como derrotas e *PLUS_MINUS_OPP*

com valores altos, tem maior frequência nos times que não foram aos *playoffs*.

Verificou-se também a partir do teste de *Shapiro-Wilk* que 4 dessas variáveis seguem uma distribuição normal, sendo elas todas ligadas ao *PLUS_MINUS*. Todos os *p-valores* para os testes T e de *Mann-Whitney* foram inferiores a 0.05, isso quer dizer que para todas as 10 variáveis há diferença nas médias entre o grupo que vai aos *playoffs* e o que não vai.

5 Conclusão

Formulou-se diversos modelos com 4 técnicas (Árvores Binárias, *Random Forest*, *Lasso* e *Ridge*) diferentes visando buscar um modelo com ótimo desempenho para previsão de um time ir aos *playoffs* ou não, utilizando as estatísticas dos jogos iniciais e considerando realizar esse estudo com o menor quantidade de jogos possíveis.

Após comparados os modelos, verificou-se que a técnica de *Random Forest* obteve melhor desempenho em quase todas as quantidades de jogos utilizadas, apresentando o comportamento mais estável. A técnica de Árvores Binárias demonstrou bastante oscilação nos valores das medidas de qualidade do ajuste, se mostrando não tão confiável. As técnicas de *Lasso* e *Ridge* não demonstraram bons desempenhos em razão da forte presença de multicolinearidade entre as variáveis consideradas.

Com base nessa etapa exploratória, foi escolhido o modelo utilizando a técnica *Random Forest*, aplicando as estatísticas dos primeiros 15 jogos e aplicando as informações dos 4 últimos jogos desses 15 para o cálculo da média móvel. Os resultados desse modelo foram satisfatórios, resultando em uma acurácia de 0.812 e um F1 de 0.817.

Verificou-se também quais foram os 10 fatores que mais influenciaram para a classificação dos times para a próxima fase. A informação de PLUS_MINUS e suas variações foram as únicas que forneceram a performance real do time a cada partida, pois as outras variáveis mais significativas estão relacionadas a vitórias e derrotas, sendo que deve-se levar em consideração uma maior influência desses indicadores quando se trata de jogos fora de casa. As únicas variáveis obtidas através do processo de *feature engineering* que apresentaram importância no modelo foram relacionadas a média móvel da variável de PLUS_MINUS.

Para trabalhos futuros sugere-se a criação de novos atributos por meio da *feature engineering* que sejam capazes de melhorar o poder preditivo, sendo que se elas forem relacionadas com as variáveis de vitórias e derrotas, tem mais chances de serem significativas. Além disso, fazer um estudo mais aprofundado sobre as variáveis mais significativas utilizando outras técnicas. Verificar se há outros problemas de modelagem além da multicolinearidade. Também é aconselhável utilizar outras técnicas de previsão que sejam menos complexas que a de *Random Forest*.

6 Referências

AGRESTI, A. *An introduction to categorical data analysis*. [S.l.]: John Wiley & Sons, 2018.

BIAU, G. Analysis of a random forests model. *The Journal of Machine Learning Research*, JMLR. org, v. 13, n. 1, p. 1063–1095, 2012.

FEOFILOFF, P. *Árvores binárias*. 2018. ⟨https://www.ime.usp.br/~pf/algoritmos/aulas/bint.html⟩.

IZBICKI, R.; SANTOS, T. M. dos. *Aprendizado de máquina: uma abordagem estatística*. [S.l.]: Rafael Izbicki, 2020.

MACIEL, L. F. V. et al. Regressão linear múltipla na modelagem de resultados na national basketball association (nba). Universidade Federal de Uberlândia, 2019.

NBA. 2022. Disponível em: ⟨https://www.nba.com/stats/teams/traditional/?sort=W_PCT&dir=-1⟩.

NELSON, M. R. *The National Basketball League: A History, 1935-1949*. [S.l.]: McFarland, 2009.

NICK, T. G.; CAMPBELL, K. M. Logistic regression. *Topics in biostatistics*, Springer, p. 273–301, 2007.

SHUNG, K. P. Accuracy, precision, recall or f1? *Towards Data Science*, 2018.

SILVA, G. P. d. Modelos de previsão para os resultados da temporada regular de 2018/19 da nba. 2019.

SILVA, L. M. O. D.

Uma aplicação de árvores de decisão, redes neurais e KNN para a identificação de modelos ARMA não-sazonais e sazonais — PUC-RIO, 2005.

SOBRAL, M. J. R. *Regressão Linear e Árvores de Regressão: Previsão do desempenho na disciplina de Matemática*. Tese (Doutorado), 2014.

SOUZA, E. G. d. Entendendo o que é matriz de confusão com python. *Data Hackers*, 2019.

TARLOW, J. Experience and winning in the national basketball association. 2012.

UNAL, I. Defining an optimal cut-point value in roc analysis: an alternative approach. *Computational and mathematical methods in medicine*, Hindawi, v. 2017, 2017.

WALSH, W. *Basketball dataset*. 2021. Disponível em: ⟨https://www.kaggle.com/wyattowalsh/basketball⟩.

7 Apêndice

Quadro 3: Descrição das variáveis

Variável	Descrição
SEASON	Temporada
TEAM_NAME	Nome do time
CONF_TEAM	Conferência do time (1 - Oeste, 0 - Leste)
PLAYOFFS	Classificação aos playoffs (1 - Sim, 0 - Não)
WINS	Número de vitórias
LOSES	Número de derrotas
HOME	Número de jogos em casa
AWAY	Número de jogos fora de casa
WINS_HOME	Número de vitórias em casa
WINS_AWAY	Número de vitórias fora de casa
LOSES_HOME	Número de derrotas em casa
LOSES_AWAY	Número de derrotas fora de casa
MIN	Média de minutos jogados por jogo
PTS	Média de pontos convertidos por jogo
FGM	Média de tentativas de cestas convertidas por jogo
FGA	Média de tentativas de cestas total por jogo
FG_PCT	Média da porcentagem de tentativas de cestas convertidas por jogo
FG3M	Média de tentativas de cestas de 3 pontos convertidas por jogo
FG3A	Média de tentativas de cestas de 3 pontos total por jogo
FG3_PCT	Média da porcentagem de tentativas de cestas de 3 pontos convertidas por jogo
FTM	Média de tentativas de lances livres convertidos por jogo
FTA	Média de tentativas de lances livres total por jogo
FT_PCT	Média da porcentagem de tentativas de lances livres convertidos por jogo
OREB	Média de rebotes ofensivos por jogo
DREB	Média de rebotes defensivos por jogo
REB	Média do total de rebotes por jogo
AST	Média de assistências por jogo
STL	Média de roubos de bola por jogo
BLK	Média de bloqueios por jogo

Continua

Continuação

Variável	Descrição
TOV	Média de erros por jogo
PF	Média de faltas cometidas por jogo
PLUS_MINUS	Média da diferença de pontos contra os oponentes por jogo
MIN_OPP	Média de minutos jogados por jogo dos oponentes
PTS_OPP	Média de pontos convertidos por jogo dos oponentes
FGM_OPP	Média de tentativas de cestas convertidas por jogo dos oponentes
FGA_OPP	Média de tentativas de cestas total por jogo dos oponentes
FG_PCT_OPP	Média da porcentagem de tentativas de cestas convertidas por jogo dos oponentes
FG3M_OPP	Média de tentativas de cestas de 3 pontos convertida por jogo dos oponentes
FG3A_OPP	Média de tentativas de cestas de 3 pontos total por jogo dos oponentes
FG3_PCT_OPP	Média da porcentagem de tentativas de cestas de 3 pontos convertidas por jogo dos oponentes
FTM_OPP	Média de tentativas de lances livres convertidos por jogo dos oponentes
FTA_OPP	Média de tentativas de lances livres total por jogo dos oponentes
FT_PCT_OPP	Média da porcentagem de tentativas de lances livres convertidos por jogo dos oponentes
OREB_OPP	Média de rebotes ofensivos por jogo dos oponentes
DREB_OPP	Média de rebotes defensivos por jogo dos oponentes
REB_OPP	Média do total de rebotes por jogo dos oponentes
AST_OPP	Média de assistências por jogo dos oponentes
STL_OPP	Média de roubos de bola por jogo dos oponentes
BLK_OPP	Média de bloqueios por jogo dos oponentes
TOV_OPP	Média de erros por jogo dos oponentes
PF_OPP	Média de faltas cometidas por jogo dos oponentes
PLUS_MINUS _OPP	Média da diferença de pontos a favor dos oponentes por jogo dos oponentes
WINS_MM	Número de vitórias considerando a Média Móvel
LOSES_MM	Número de derrotas considerando a Média Móvel
HOME_MM	Número de jogos em casa considerando a Média Móvel

Continua

Continuação

Variável	**Descrição**
AWAY_MM	Número de jogos fora de casa considerando a Média Móvel
WINS_HOME_MM	Número de vitórias em casa considerando a Média Móvel
WINS_AWAY_MM	Número de vitórias fora de casa considerando a Média Móvel
LOSES_HOME_MM	Número de derrotas em casa considerando a Média Móvel
LOSES_AWAY_MM	Número de derrotas fora de casa considerando a Média Móvel
MIN_MM	Média de minutos jogados por jogo considerando a Média Móvel
PTS_MM	Média de pontos convertidos por jogo considerando a Média Móvel
FGM_MM	Média de tentativas de cestas convertidas por jogo considerando a Média Móvel
FGA_MM	Média de tentativas de cestas total por jogo considerando a Média Móvel
FG_PCT_MM	Média da porcentagem de tentativas de cestas convertidas por jogo considerando a Média Móvel
FG3M_MM	Média de tentativas de cestas de 3 pontos convertidas por jogo considerando a Média Móvel
FG3A_MM	Média de tentativas de cestas de 3 pontos total por jogo considerando a Média Móvel
FG3_PCT_MM	Média da porcentagem de tentativas de cestas de 3 pontos convertidas por jogo considerando a Média Móvel
FTM_MM	Média de tentativas de lances livres convertidos por jogo considerando a Média Móvel
FTA_MM	Média de tentativas de lances livres total por jogo considerando a Média Móvel
FT_PCT_MM	Média da porcentagem de tentativas de lances livres convertidos por jogo considerando a Média Móvel
OREB_MM	Média de rebotes ofensivos por jogo considerando a Média Móvel
DREB_MM	Média de rebotes defensivos por jogo considerando a Média Móvel
REB_MM	Média do total de rebotes por jogo considerando a Média Móvel
AST_MM	Média de assistências por jogo considerando a Média Móvel
STL_MM	Média de roubos de bola por jogo considerando a Média Móvel
BLK_MM	Média de bloqueios por jogo considerando a Média Móvel
TOV_MM	Média de erros por jogo considerando a Média Móvel

Continua

Continuação

Variável	Descrição
PF_MM	Média de faltas cometidas por jogo considerando a Média Móvel
PLUS_MINUS_MM	Média da diferença de pontos contra os oponentes por jogo considerando a Média Móvel
MIN_OPP_MM	Média de minutos jogados por jogo dos oponentes considerando a Média Móvel
PTS_OPP_MM	Média de pontos convertidos por jogo dos oponentes considerando a Média Móvel
FGM_OPP_MM	Média de tentativas de cestas convertidas por jogo dos oponentes considerando a Média Móvel
FGA_OPP_MM	Média de tentativas de cestas total por jogo dos oponentes considerando a Média Móvel
FG_PCT_OPP_MM	Média da porcentagem de tentativas de cestas convertidas por jogo dos oponentes considerando a Média Móvel
FG3M_OPP_MM	Média de tentativas de cestas de 3 pontos convertidas por jogo dos oponentes considerando a Média Móvel
FG3A_OPP_MM	Média de tentativas de cestas de 3 pontos total por jogo dos oponentes considerando a Média Móvel
FG3_PCT_OPP_MM	Média da porcentagem de tentativas de cestas de 3 pontos convertidas por jogo dos oponentes considerando a Média Móvel
FTM_OPP_MM	Média de tentativas de lances livres convertidos por jogo dos oponentes considerando a Média Móvel
FTA_OPP_MM	Média de tentativas de lances livres total por jogo dos oponentes considerando a Média Móvel
FT_PCT_OPP_MM	Média da porcentagem de tentativas de lances livres convertidos por jogo dos oponentes considerando a Média Móvel
OREB_OPP_MM	Média de rebotes ofensivos por jogo dos oponentes considerando a Média Móvel
DREB_OPP_MM	Média de rebotes defensivos por jogo dos oponentes considerando a Média Móvel
REB_OPP_MM	Média do total de rebotes por jogo dos oponentes considerando a Média Móvel
AST_OPP_MM	Média de assistências por jogo dos oponentes considerando a Média Móvel

Continua

Continuação

Variável	**Descrição**
STL_OPP_MM	Média de roubos de bola por jogo dos oponentes considerando a Média Móvel
BLK_OPP_MM	Média de bloqueios por jogo dos oponentes considerando a Média Móvel
TOV_OPP_MM	Média de erros por jogo dos oponentes considerando a Média Móvel
PF_OPP_MM	Média de faltas cometidas por jogo dos oponentes considerando a Média Móvel
PLUS_MINUS _OPP_MM	Média da diferença de pontos a favor dos oponentes por jogo dos oponentes considerando a Média Móvel
WINS_OPP	Média das proporções de vitórias dos oponentes que o time enfrentou até aquele momento da competição