



Universidade de Brasília - UnB
Instituto de Ciências Exatas - IE
Departamento de Estatística - EST

Previsão de Resultados da Temporada Regular de 2018/19 da NBA

Gustavo Pompeu da Silva

Orientador: Eduardo Monteiro de Castro Gomes

Brasília

2019

Gustavo Pompeu da Silva

Previsão de Resultados da Temporada Regular de 2018/19 da NBA

Relatório apresentado à disciplina de Trabalho de Conclusão de Curso II de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Orientador: Eduardo Monteiro de Castro Gomes

Brasília

2019

Gustavo Pompeu da Silva

Previsão de Resultados da Temporada Regular de 2018/19 da NBA/ Gustavo Pompeu da Silva. – Brasília, 2019-

57 p. : il. (algumas color.) ; 30 cm.

Orientador: Eduardo Monteiro de Castro Gomes

Relatório Final – Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Trabalho de Conclusão de Curso de Graduação, 2019.

1. NBA. 2. Previsão. 3. Resultados. 4. Regressão. 5. R. 6. Estatística.

Gustavo Pompeu da Silva

Previsão de Resultados da Temporada Regular de 2018/19 da NBA

Relatório apresentado à disciplina de Trabalho de Conclusão de Curso II de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Eduardo Monteiro de Castro Gomes
Orientador

Leandro Tavares Correia
Membro da Banca

Donald Matthew Pianto
Membro da Banca

Brasília
2019

Resumo

Este texto apresenta algumas notas de aula de TCC 1 com o formato da monografia que deve ser apresentada para conclusão do curso de Bacharelado em Estatística na Universidade de Brasília. O objetivo é apenas padronizar a apresentação do Trabalho de Conclusão de Curso, utilizando normas da ABNT e o pacote \LaTeX .

Palavras-chave: \LaTeX , abntex, pesquisa, monografia, slides, poster.

Abstract

This is the english abstract.

Keywords: L^AT_EX, abntex, research, pesquisa, monograph, slides, poster.monografia, slides, poster.

Lista de tabelas

Tabela 1 – Funções e Pacotes utilizados no R para cada método	25
Tabela 2 – Média de pontos sofridos por equipe	34
Tabela 3 – Média de pontos marcados por equipe	35
Tabela 4 – Porcentagem de Acerto das previsões dos jogos da temporada 2018/19 para cada método utilizando dados de 2000/01 à 2017/18 na modelagem	36
Tabela 5 – Porcentagem de Acerto das previsões dos jogos da temporada 2018/19 para cada método utilizando temporadas diferentes na modelagem - Parte 1	37
Tabela 6 – Porcentagem de Acerto das previsões dos jogos da temporada 2018/19 para cada método utilizando temporadas diferentes na modelagem - Parte 2	37
Tabela 7 – Porcentagem de Acerto das previsões dos jogos da temporada 2018/19 para cada método utilizando dados de 2006/07 à 2017/18 na modelagem	38
Tabela 8 – Tempo de execução do código computacional para cada método	39
Tabela 9 – Totais dos acertos das previsões por time separados por categorias - Parte 1	40
Tabela 10 – Totais dos acertos das previsões por time separados por categorias - Parte 2	41
Tabela 11 – Porcentagens dos acertos das previsões por time separados por categorias	42
Tabela 12 – Comparação das vitórias reais com as vitórias previstas - Conferência Leste	43
Tabela 13 – Comparação das vitórias reais com as vitórias previstas - Conferência Oeste	44
Tabela 14 – Variáveis mais significativas no modelo	46

Lista de abreviaturas e siglas

NBA	<i>National Basketball Association</i>
SVM	<i>Support Vector Machine</i> (Máquina de Vetores de Suporte)
NA	Não aplicável

Sumário

1	INTRODUÇÃO	15
2	REVISÃO DE LITERATURA	17
2.1	Regressão Linear	17
2.2	Regressão Logística	18
2.3	Regressão de Probit	19
2.4	Seleção de Variáveis	19
2.5	Máquina de Vetores de Suporte (SVM)	20
2.6	Análise de Discriminante Linear	21
2.7	Árvores de Regressão e Classificação	22
2.8	<i>Random Forest</i>	23
3	MATERIAL E MÉTODOS	25
3.1	Implementação Computacional dos Métodos	25
3.2	Criação das Bases de Dados	25
3.2.1	Lidando com valores faltantes	30
3.3	Casas de Aposta	31
4	RESULTADOS	33
4.1	Resultados Reais da Temporada 2018/19	33
4.1.1	Temporada Regular	33
4.1.2	<i>Playoffs</i>	35
4.2	Previsões	35
4.2.1	“Previsões” das casas de aposta	38
4.3	Tempo de execução computacional de cada método	39
4.4	Modelo “campeão”	39
4.4.1	Acertos por Equipe	40
4.4.1.1	Comparação da tabela de previsões com a tabela real	43
4.5	Acertos das Previsões Durante a Temporada	44
4.6	Variáveis mais significativas	46
4.7	Adicionando jogos de 2018/19 na modelagem	47
5	CONCLUSÃO	49
	REFERÊNCIAS	51

APÊNDICES	55
APÊNDICE A – CÓDIGOS EM R	57

1 Introdução

Mineração de dados em esportes é um tópico que tem crescido rapidamente nos últimos anos. Jogadores de ligas de *fantasy*, apostadores e entusiastas de esportes estão cada vez mais interessados em procurar uma vantagem nas apostas e previsões através de dados e números. Ferramentas e técnicas começaram a ser desenvolvidas para medir desempenho tanto de times quanto de atletas, e esses métodos vem chamando a atenção de grandes franquias esportivas.

Existe uma imensa quantidade de dados disponíveis sobre qualquer esporte. Esses dados podem ser de desempenho individual de jogadores ou da equipe, decisões da comissão técnica, eventos que acontecem nos jogos, entre outros. É preciso saber não só como coletar esses dados, mas também quais podem ser úteis e como fazer o melhor uso possível deles. Achando os meios para transformar esses dados em conhecimento, organizações esportivas tem o potencial de obter uma vantagem competitiva sobre seus oponentes. Não se deve analisar performance no sentido de marcar mais gols ou pontos do que o oponente, pois esse é o objetivo geral de qualquer esporte, o que é interessante é encontrar padrões em outras estatísticas que mostram tendências justamente para chegar às vitórias.

Data Mining envolve procedimentos para descobrir padrões escondidos e descobrir novas informações a partir de fontes de dados. A fundação científica de data mining pode ser dividida em três disciplinas: estatísticas, inteligência artificial e *machine learning*. *Data mining* então pode ser definido como a busca de conhecimento dentro dos dados. (SCHUMAKER; SOLIEMAN; CHEN, 2010)

A NBA (*National Basketball Association*) é a principal liga de basquete profissional do mundo. Atualmente, é composta por 30 times baseados em cidades da América do Norte (29 nos Estados Unidos e 1 no Canadá), divididos em 2 conferências: Leste e Oeste. É a liga onde jogam os melhores atletas de basquete do mundo, e com os maiores salários do esporte. Uma das vantagens de trabalharmos com o basquete e a NBA especificamente é a grande quantidade de dados, pois, atualmente, em uma temporada regular, cada time joga 82 vezes, ou seja, são 1230 jogos por temporada, isso nos permite ter muitas observações para trabalhar. Os 8 melhores times de cada conferência se classificam para os *playoffs* para disputar o título de campeão da NBA.

O objetivo geral desse trabalho é ajustar modelos utilizando diversas técnicas estatísticas para obter previsões para os resultados dos jogos da temporada regular de 2018/19 da NBA e compará-las para chegar em uma conclusão sobre qual técnica funcionou melhor para esse problema em específico, julgando principalmente pela acurácia das previsões.

Este trabalho está organizado de forma que no capítulo 2 está a revisão de literatura com resumos teóricos dos métodos estatísticos que serão aplicados, no capítulo 3 a metodologia utilizada, principalmente a parte computacional, no capítulo 4 os resultados obtidos, e por fim a conclusão.

2 Revisão de Literatura

Os métodos a serem descritos abaixo nesse capítulo serão utilizados para as modelagens com o objetivo de fazer previsões para os jogos da temporada 2018/19 da NBA.

Serão consideradas duas abordagens para a variável dependente, uma quantitativa, que é o saldo de pontos entre os times (subtrair a pontuação de um time pela do outro), de onde se infere qual time vence o jogo pelo sinal do saldo, e a outra qualitativa, que é uma variável dicotômica indicando somente se o time venceu ou perdeu o jogo.

Considerando essas duas abordagens, serão utilizados diferentes métodos, para que os dois tipos de variáveis possam ser aplicadas em vários modelos.

Alguns dos modelos descritos possuem suposições para serem aplicados, porém, nesse trabalho, elas não serão verificadas, pois para o objetivo específico de obter previsões para os resultados dos jogos, não serão levadas em consideração.

As técnicas estatísticas a serem utilizadas para a obtenção das previsões dos jogos serão:

- Regressão Linear
- Regressão Logística
- Regressão de Probit
- Máquina de Vetores de Suporte (SVM)
- Análise de Discriminante Linear
- Árvores de Regressão
- Árvores de Classificação
- *Random Forest*

2.1 Regressão Linear

Regressão linear é uma equação para se descrever o valor esperado de uma variável Y (resposta), dados os valores de outras variáveis X (explicativas). É chamada linear porque se considera que a relação da resposta às variáveis explicativas é uma função linear de alguns parâmetros. A equação que determina a relação entre as variáveis é:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$$

Em que Y é a variável resposta (dependente) e p é o número de variáveis explicativas. β_j com $j = 0, 1, \dots, p$ são constantes, denominados coeficientes de regressão, X_j com $j = 1, \dots, p$ são as variáveis explicativas (independentes) e ϵ representa o erro experimental. O parâmetro β_0 corresponde ao intercepto, e fornece a resposta média de Y quando $X_1 = X_2 = \dots = X_p = 0$. Para $j \geq 1$, os parâmetros β_j indicam uma mudança na resposta média de Y a cada unidade de mudança na variável X_j , quando as demais variáveis são mantidas fixas.

As suposições clássicas necessárias para o Modelo de Regressão Linear Múltipla são (YAN; SU, 2009):

- Os erros não devem ser correlacionados, devem seguir distribuição normal e ter média zero e variância σ^2 , desconhecida. Ou seja, $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$;
- Deve existir uma relação linear entre a variável dependente e as variáveis independentes;

Outra coisa a ser observada, mas que não chega a ser uma suposição para o modelo, é que não deve haver multicolinearidade entre as variáveis independentes, ou seja, elas não devem ter correlação alta entre si.

Nesse trabalho, não serão cumpridas todas as suposições da regressão linear, principalmente porque as observações não são independentes umas das outras, pois os jogos são uma sequência histórica no tempo.

2.2 Regressão Logística

A regressão logística se difere da linear essencialmente pelo fato da variável resposta ser binária, ou seja, Y tem distribuição Bernoulli $(1, \pi)$, com probabilidade de sucesso $P(Y_i = 1) = \pi_i$ e de fracasso $P(Y_i = 0) = 1 - \pi_i$.

No centro da regressão logística está a tarefa de estimar o *log odds* de um evento. Matematicamente, a regressão logística estima uma função de regressão linear múltipla definida por:

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \quad (2.1)$$

Onde $\pi = P(Y = 1)$. Baseado em 2.1, chegamos em:

$$\pi = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)} = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p}}$$

As suposições necessárias para o Modelo de Regressão Logística são (KASSAMBARA, 2018):

- A variável dependente precisa ser binária (dicotômica);
- Não deve haver multicolinearidade entre as variáveis independentes, ou seja, elas não podem ter correlação alta entre si;
- Deve existir linearidade entre as variáveis independentes e a função *logit*;
- Não deve haver valores extremos nas variáveis independentes contínuas.

Novamente, não haverá uma preocupação em cumprir as suposições para o modelo.

2.3 Regressão de Probit

A Análise de Probit ou Regressão de Probit (CARVALHO et al., 2017) é outro tipo de regressão binária, parecida com a regressão logística, a diferença é a função de ligação utilizada, o *link* probit é dado por:

$$probit(\pi) = \Phi^{-1}(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \quad (2.2)$$

Onde $\pi = P(Y = 1)$. Baseado em 2.2, chegamos em:

$$\pi = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)$$

Em que Φ é a Função de Distribuição Acumulada (f.d.a.) da distribuição Normal Padrão.

As suposições necessárias são as mesmas da Regressão Logística, pois são modelos lineares generalizados da mesma família (binomial). Nesse trabalho, as suposições não serão verificadas.

2.4 Seleção de Variáveis

Para os métodos de regressão citados acima, é possível aplicar um método de seleção de variáveis (SUÁREZ et al., 2017), em que modelos são analisados com base em alguma medida, e variáveis são retiradas/adicionadas para que seja obtido o modelo com o melhor valor da medida analisada. A função *step*, que faz parte do R, utiliza a medida AIC para escolher as variáveis.

Existem três formas de aplicar a função *step*:

- *forward*, em que o modelo começa sem nenhuma variável explicativa, e a cada passo vai adicionando a variável que deixaria o modelo com o menor valor de AIC, até que o modelo com o menor AIC seja o atual, sem adicionar mais nenhuma variável.
- *backward*, em que o modelo começa com todas as variáveis explicativas, e a cada passo vai retirando a variável que resultaria em um modelo com o menor valor de AIC, até que o modelo atual tenha o menor AIC.
- *stepwise*, que é uma mistura dos dois métodos acima, em cada passo é aplicada uma iteração do *forward* e uma do *backward*, por exemplo, em um passo, é adicionada a variável que diminuiria mais o valor do AIC, e logo após, é verificado se retirar alguma variável que já estava não diminuiria o AIC. O processo só para quando nem adicionar, nem retirar nenhuma variável diminuiria o valor do AIC.

Nesse trabalho será aplicado o método *forward*, pois as bases de dados utilizadas serão muito grandes e com muitas variáveis, e esse método é o mais rápido de ser aplicado computacionalmente, além disso, os três normalmente obtêm resultados similares.

2.5 Máquina de Vetores de Suporte (SVM)

As Máquinas de Vetores de Suporte (*Support Vector Machines* - SVMs) constituem uma técnica embasada na Teoria de Aprendizado Estatístico (VAPNIK, 1995), em que seu objetivo é reconhecer padrões nos dados.

A teoria da SVM é bastante complexa, mas sua abordagem pode ser esboçada da seguinte forma:

- **Separação das Classes:** Para classificar dados em duas classes diferentes, é procurado um plano que separe as classes no espaço p -dimensional. Esse plano é chamado de hiperplano. O objetivo é determinar o hiperplano ótimo, e isso é feito basicamente através da maximização das “margens” entre os pontos mais próximos das classes (ver Figura 1), os pontos em cima das fronteiras são chamados de vetores de suporte, e o plano no meio das margens é o hiperplano ótimo de separação;
- **Classes sobrepostas:** Observações do lado “errado” da margem discriminante são ponderadas para reduzir suas influências;
- **Não-linearidade:** Quando um separador linear não é encontrado, as observações são projetadas em um espaço de maior dimensão, onde elas se tornam efetivamente linearmente separáveis (essa projeção é feita via técnicas Kernel);
- **Solução de Problemas:** A tarefa toda pode ser formulada como um problema de otimização quadrática que pode ser resolvida por técnicas conhecidas.

Um programa capaz de realizar todas essas tarefas é chamado de uma Máquina de Vetores de Suporte. (MEYER, 2019)

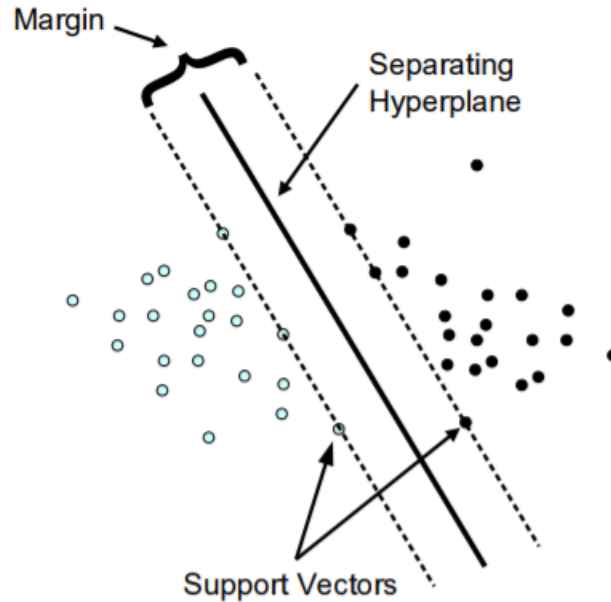


Figura 1 – Classificação (caso de separação linear)

Para a implementação computacional, será utilizada a função *svm* do pacote *e1071* da linguagem R (MEYER et al., 2018), que possui 4 opções de função Kernel: linear, base radial (gaussiana), polinomial e sigmoidal.

A função Kernel a ser utilizada nesse trabalho será a base radial, pois foi a que obteve os melhores resultados para as previsões em testes preliminares. Ela contém um parâmetro específico γ que tem como valor padrão $\frac{1}{\text{dimensão dos dados}}$.

Outro parâmetro a ser utilizado será o custo, que é um parâmetro geral de penalização para esse tipo de classificação. Seu valor padrão é 1.

Serão feitas modelagens com os valores padrão desses dois parâmetros, mas também serão encontrados os melhores valores para eles para os dados utilizados, através da função *tune*, do mesmo pacote da função *svm*.

2.6 Análise de Discriminante Linear

A Análise de Discriminante é uma técnica multivariada que se preocupa em separar observações em grupos, e alocar novas observações em algum dos grupos pré-definidos. A Análise de Discriminante é bastante exploratória em sua natureza. Em geral, o objetivo dessa técnica é descrever algebricamente as características diferenciais das observações,

nós tentamos achar “discriminantes” cujo valores numéricos são tais que as populações são separadas o melhor possível. (JOHNSON; WICHERN, 2007)

A Análise de Discriminante Linear é uma generalização da Discriminante Linear de Fisher. Para duas classes, a alocação de novas observações funciona de uma maneira muito simples. Primeiramente, é feita uma matriz de variância-covariância estimada para os dados (\mathbf{S}_p^2):

$$\mathbf{S}_p^2 = \frac{\sum_{j=1}^{n_1} (\mathbf{x}_{1j} - \bar{\mathbf{x}}_1)^2 + \sum_{j=1}^{n_2} (\mathbf{x}_{2j} - \bar{\mathbf{x}}_2)^2}{n_1 + n_2 - 2}$$

Onde n_1 e n_2 correspondem ao número de observações da população 1 (π_1) e ao número de observações da população 2 (π_2), respectivamente, $\bar{\mathbf{x}}_1$ e $\bar{\mathbf{x}}_2$ correspondem às médias das variáveis independentes para cada população, e \mathbf{x}_{1j} e \mathbf{x}_{2j} são referentes à cada observação j de cada população.

Então, para uma nova observação \mathbf{x}_0 , temos: $\hat{y}_0 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_p^{-1} \mathbf{x}_0$ e $\hat{m} = \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_p^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$, e a regra de alocação será que x_0 pertencerá à π_1 se $\hat{y}_0 - \hat{m} \geq 0$, e x_0 pertencerá à π_2 caso contrário.

As suposições necessárias para a Análise de Discriminante Linear são:

- Normalidade multivariada dos dados;
- Matriz de variância-covariância das populações devem ser iguais.

2.7 Árvores de Regressão e Classificação

As árvores de regressão e classificação são métodos estatísticos não-paramétricos utilizados baseado na teoria de árvores de decisão (ver Figura 2), mas os nós terminais da árvore são resultados numéricos (caso a variável resposta seja quantitativa), ou classes (caso a variável resposta seja qualitativa). No primeiro caso, chamamos o método de Árvore de Regressão, e no segundo de Árvore de Classificação.

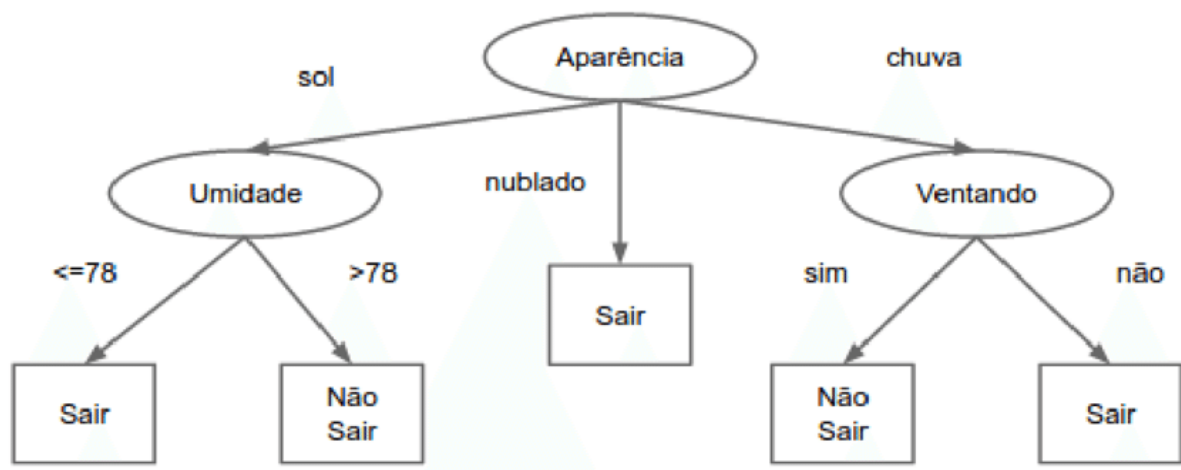


Figura 2 – Exemplo de Árvore de Decisão para sair ou não de um determinado local

As árvores são construídas a partir de um particionamento recursivo binário usando a variável resposta e escolhendo divisões das variáveis explicativas. (RIPLEY, 1996)

O processo é denominado recursivo porque cada subpopulação criada pode ser dividida por um número indefinido de vezes até que o processo de divisão termine após um determinado critério de parada ser atingido.

Algoritmos para construção dessas árvores geralmente trabalham de cima para baixo, escolhendo em cada etapa uma variável que melhor divide o conjunto de observações. Algoritmos diferentes usam métricas diferentes para definir essa “melhor” divisão. Geralmente, é medida a homogeneidade da variável alvo dentro dos subconjuntos.

São alternativas não-paramétricas à regressão linear e à regressão logística, e não necessitam de pressupostos para serem aplicadas.

2.8 Random Forest

Random forests são uma combinação de preditores de árvores (vistos na seção 2.7) tal que cada árvore depende dos valores de um vetor aleatório amostrado de forma independente e com a mesma distribuição para todas as árvores na “floresta”. O erro de generalização para as florestas converge quase certamente para um limite conforme o número de árvores na floresta se torna grande. O erro de generalização de uma floresta de árvores de classificação depende da força das árvores individuais na floresta e da correlação entre elas. É utilizada uma seleção aleatória de observações e de variáveis para a criação de cada árvore de classificação. Estimativas internas monitoram o erro, a força e a correlação e são usadas para mostrar a resposta ao aumento do número de variáveis usados na separação. Estimativas internas também são usadas para medir a importância das variáveis. Todas

essas ideias também são aplicáveis para regressão. (BREIMAN, 2001)

Uma vantagem desse método é a prevenção de *overfitting*, mas em compensação é um método muito mais lento de ser computado do que uma árvore de classificação ou regressão, pois são construídas muitas árvores em vez de uma só.

3 Material e Métodos

A linguagem R (R Core Team, 2018) será a única utilizada nesse trabalho.

3.1 Implementação Computacional dos Métodos

Cada método estatístico mencionado no capítulo 2 tem uma função em um pacote da linguagem R para sua implementação computacional. As que serão utilizadas nesse trabalho se encontram na Tabela 1.

Tabela 1 – Funções e Pacotes utilizados no R para cada método

Método	Função	Pacote
Regressão Linear	<i>lm</i>	<i>stats</i> (R Core Team, 2018)
Regressão Logística	<i>glm</i>	<i>stats</i>
Regressão de Probit	<i>glm</i>	<i>stats</i>
Seleção de Variáveis para as Regressões	<i>step</i>	<i>stats</i>
SVM	<i>svm</i>	<i>e1071</i> (MEYER et al., 2018)
Análise de Discriminante Linear	<i>lda</i>	<i>MASS</i> (VENABLES; RIPLEY, 2002)
Árvores de Regressão e Classificação	<i>tree</i>	<i>tree</i> (RIPLEY, 2019)
<i>Random Forest</i>	<i>randomForest</i>	<i>randomForest</i> (LIAW; WIENER, 2002)

3.2 Criação das Bases de Dados

Para a obtenção dos dados necessários, será utilizada uma técnica de *web scraping*, em que informações são extraídas de alguma página da internet. Com o auxílio do pacote *rvest* (WICKHAM, 2016), serão extraídas informações do site Basketball Reference (BASKETBALL...), um dos maiores sites com dados numéricos sobre a NBA e basquete em geral.

Toda página da internet possui um código HTML por trás, e existe uma extensão do *Google Chrome*, chamada *SelectorGadget* (SELECTOR...), que permite ao usuário clicar nas áreas do site para selecionar as partes que se deseja extrair a partir de seu código HTML, mesmo que o usuário não tenha conhecimento de programação em HTML, e combinando isso com funções do pacote *rvest*, pode-se transformar esse código em texto no R.

Foi escolhido fazer a extração dos dados de resultados de jogos a partir da temporada de 2000/01, pois dela até a de 2017/18 são 18 temporadas para realizar as modelagens e

serem feitas as previsões dos jogos da temporada 2018/19.

O número de jogos por temporada regular varia por alguns motivos. Nas temporadas de 2000/01 a 2004/05, a NBA era composta por apenas 29 times, mas cada time já jogava 82 jogos, isso resultava em 1189 jogos por temporada regular. Nas demais temporadas utilizadas, 30 times faziam parte da liga, resultando em 1230 jogos por temporada. A única exceção foi a temporada de 2011/12, quando aconteceu um *lockout*, quando os donos das equipes se recusaram a deixar os jogos acontecerem, pois o contrato da NBA com os times acabou antes do início da temporada, e a NBA demorou para chegar em um acordo com os donos dos times para assinarem um novo contrato. Um novo acordo foi estabelecido depois de vários meses de negociação, e a temporada começou em 25 de dezembro de 2011, com quase 2 meses de atraso. Isso diminuiu o número de jogos realizados por cada equipe de 82 para 66 jogos, que resultou em um total de apenas 990 jogos na temporada regular.

As informações obtidas de cada um dos jogos realizados das temporadas citadas são: data do jogo, nome do time visitante, pontos marcados pelo time visitante, nome do time mandante, pontos marcados pelo time mandante, se houve prorrogação no jogo, e o público presente no ginásio.

A partir do que foi obtido, podemos criar uma base de dados com muitas variáveis derivadas dessas informações, e então criar modelos para realizar as previsões, utilizando as diversas técnicas estatísticas citadas anteriormente.

É importante ressaltar que a temporada regular de 2018/19 estava em andamento durante a realização desse trabalho, tendo durado de Outubro de 2018 até Abril de 2019, e as informações relacionadas aos jogos dessa temporada foram sendo extraídas gradualmente conforme os jogos foram sendo realizados, mas, ao fim do trabalho a temporada regular já havia sido concluída.

A base de dados inicial, criada a partir das informações extraídas da internet, contém 2 linhas para cada jogo realizado, cada linha tendo informações referentes à um dos times envolvidos na partida, e contém as seguintes variáveis:

Variáveis de identificação das informações do jogo:

- *Team*: Nome do time
- *Opp*: Nome do time adversário
- *Pts_S*: Pontos marcados pelo time nesse jogo.
- *Pts_A*: Pontos marcados pelo time adversário nesse jogo.
- *Home*: Se o time jogou em casa ou não.

- *Attend*: Público presente no ginásio nesse jogo. (Tem a informação apenas se o time jogou em casa)
- *OT*: Indica se ocorreu prorrogação no jogo.

Essas variáveis acima não serão utilizadas nas modelagens, elas são apenas as variáveis extraídas da internet, e terão suas informações repassadas para outras variáveis que serão citadas a seguir.

Variáveis indicadoras do resultado do jogo:

- *Win*: Se o time venceu esse jogo ou não (qualitativa, dicotômica).
- *result*: Saldo de pontos, ou seja, os pontos marcados pelo time menos os pontos marcados pelo seu adversário (quantitativa).

Variáveis de informação sobre o jogo que podem ser identificadas antes da realização da partida:

- *weekday*: Dia da semana em que o jogo foi/será realizado. Essa variável pode ser utilizada nas modelagens, pois sabemos o dia da semana que o jogo ocorrerá mesmo antes do jogo acontecer.
- *Travel*: Variável que indica se o time teve/terá que viajar da partida anterior para essa ou não. Por exemplo, se o jogo anterior foi fora de casa, o time sempre tem que viajar, ou pra voltar pra casa, ou pra ir para outra cidade fora de casa. O time só não viaja quando joga 2 jogos seguidos em casa.

Variáveis referentes à toda informação do time desde o início da temporada até antes do jogo:

- *Games_T*: Total de jogos do time até agora na temporada.
- *Games_H*: Jogos em casa do time até agora na temporada.
- *Games_A*: Jogos fora de casa do time até agora na temporada.
- *Wins_T*: Total de vitórias do time até agora na temporada.
- *Wins_H*: Vitórias em casa do time até agora na temporada.
- *Wins_A*: Vitórias fora de casa do time até agora na temporada.
- *Loss_T*: Total de derrotas do time até agora na temporada.
- *Loss_H*: Derrotas em casa do time até agora na temporada.

- *Loss_A*: Derrotas fora de casa do time até agora na temporada.
- *Streak_T*: Número indicando a sequência de vitórias (positivo) ou derrotas (negativo) do time, considerando jogos em casa e fora de casa. Exemplos: os 5 últimos jogos do time foram vitórias, e o antes desses 5 foi derrota, logo a variável vale +5. O último jogo do time foi derrota, e o penúltimo vitória, então a variável vale -1.
- *Streak_H*: Número indicando a sequência de vitórias (positivo) ou derrotas (negativo) do time, considerando apenas jogos em casa.
- *Streak_A*: Número indicando a sequência de vitórias (positivo) ou derrotas (negativo) do time, considerando apenas jogos fora de casa.
- *Mean_Pts_S_H*, *Max_Pts_S_H*, *Min_Pts_S_H*: Média, máximo e mínimo de pontos marcados do time em jogos em casa até o momento na temporada.
- *Mean_Pts_S_A*, *Max_Pts_S_A*, *Min_Pts_S_A*: Média, máximo e mínimo de pontos marcados do time em jogos fora de casa até o momento na temporada.
- *Mean_Pts_S_T*: Média de pontos marcados do time em todos os jogos até o momento na temporada.
- *Mean_Pts_A_H*, *Max_Pts_A_H*, *Min_Pts_A_H*: Média, máximo e mínimo de pontos sofridos do time em jogos em casa até o momento na temporada.
- *Mean_Pts_A_A*, *Max_Pts_A_A*, *Min_Pts_A_A*: Média, máximo e mínimo de pontos sofridos do time em jogos fora de casa até o momento na temporada.
- *Mean_Pts_A_T*: Média de pontos sofridos do time em todos os jogos até o momento na temporada.
- *Str_Sch*: A “força de calendário” do time até o momento na temporada, ou seja, a proporção de vitórias dos adversários que o time enfrentou até o momento na temporada. Divide-se o total de vitórias de todos os adversários do time pelo total de jogos de todos os adversários do time.
- *mean_attend*: Média de público do time nos jogos em casa, até o momento na temporada.

Variáveis referentes aos últimos 3, 5, 7 ou 10 jogos do time na temporada:

- *Mean_Last_X_A*, *Max_Last_X_A*, *Min_Last_X_A*: Média, máximo e mínimo de pontos marcados do time nos últimos X jogos fora de casa, onde $X = 3, 5, 7, 10$.
- *Mean_Last_X_H*, *Max_Last_X_H*, *Min_Last_X_H*: Média, máximo e mínimo de pontos marcados do time nos últimos X jogos em casa, onde $X = 3, 5, 7, 10$.

- *Mean_Last_X_T*, *Max_Last_X_T*, *Min_Last_X_T*: Média, máximo e mínimo de pontos marcados do time nos últimos X jogos, onde $X = 3, 5, 7, 10$.
- *Mean_Last_X_A_Opp*, *Max_Last_X_A_Opp*, *Min_Last_X_A_Opp*: Média, máximo e mínimo de pontos sofridos do time nos últimos X jogos fora de casa, onde $X = 3, 5, 7, 10$.
- *Mean_Last_X_H_Opp*, *Max_Last_X_H_Opp*, *Min_Last_X_H_Opp*: Média, máximo e mínimo de pontos sofridos do time nos últimos X jogos em casa, onde $X = 3, 5, 7, 10$.
- *Mean_Last_X_T_Opp*, *Max_Last_X_T_Opp*, *Min_Last_X_T_Opp*: Média, máximo e mínimo de pontos sofridos do time nos últimos X jogos, onde $X = 3, 5, 7, 10$.
- *Win_Last_X_A*, *Win_Last_X_H*, *Win_Last_X_T*: Número de vitórias do time nos últimos X jogos fora de casa, em casa, e total, respectivamente, onde $X = 3, 5, 7, 10$.

Variáveis referentes apenas ao último jogo realizado pelo time:

- *OT_Last*: Indica se houve prorrogação no jogo anterior do time.
- *Days_LG*: Quantos dias atrás foi o último jogo do time.

Isso resulta em 1 banco de dados para cada temporada, com 2 linhas para cada jogo realizado na temporada, e 125 variáveis.

Vamos utilizar *Win* (dicotômica, qualitativa) ou *result* (quantitativa) como variáveis dependentes para os modelos, a escolha dessa variável irá depender da técnica utilizada.

Como o objetivo é realizar uma previsão para cada jogo, as duas linhas de cada jogo serão combinadas em uma só, ou seja, uma primeira parte do banco de dados final terá apenas variáveis referentes ao time visitante de cada jogo, e a segunda parte apenas variáveis referentes ao time mandante. Como existe essa separação clara entre as variáveis, é fácil remover as que ficariam duplicadas (como o dia da semana do jogo e as variáveis dependentes), e as que não teriam propósito, (como a média de público do time visitante, que não é aplicável, e a variável que indica se o time viajou do último jogo para o atual para o time visitante, pois ela sempre será *TRUE*). Além disso, também serão retiradas as variáveis referentes à jogos fora de casa para os times mandantes e as referentes à jogos em casa para os times visitantes, pois foi julgado que elas não contribuiriam.

Após a remoção dessas variáveis, e das variáveis de identificação, é obtida uma base de dados final por temporada, com 1 linha por jogo, e 151 variáveis, sendo 2 delas as variáveis dependentes. As variáveis dependentes mantidas foram as referentes ao time

visitante, ou seja, a variável *Win* virou *Win_Vis* e tem valor *TRUE* quando o time visitante vence, e *FALSE* caso contrário, e a variável *result* virou *result_Vis*, ou seja, como é o saldo de pontos, a variável é positiva quando o time visitante vence, e negativa caso contrário. De forma similar, para todas as variáveis restantes na base de dados, foi adicionado a extensão *_Vis* no nome das que são referentes ao time visitante, e a extensão *_Home* no nome das que são referentes ao time mandante. A única variável que não é referente a nenhum dos dois, é a do dia da semana do jogo (*weekday*), e o nome dela foi mantido.

3.2.1 Lidando com valores faltantes

Como existem muitas variáveis que dependem de resultados anteriores, existirão vários valores faltantes no começo de cada temporada, que nesse trabalho serão referenciados como NA (não aplicável). Por exemplo, se o time só realizou 6 jogos na temporada, não é possível obter um valor para a média de pontos marcados nos últimos 7 jogos, ou se é o primeiro jogo do time na temporada, não há como obter nenhuma informação além do dia da semana do jogo.

Na base de dados da temporada 2018/19, que é a que será feita as previsões, das 1230 observações da temporada regular, apenas 875 possuem informação completa, ou seja, nenhum NA em nenhuma variável. Nas outras 355 observações, existe pelo menos um NA em alguma variável.

No R, quando é feita uma modelagem, geralmente, as observações que possuem algum valor NA são ignoradas, e as previsões de observações com algum valor NA são retornadas como NA também. Para tornar possível fazer as previsões de todos os jogos da temporada 2018/19, e não apenas daqueles que tem informação completa, é possível identificar os padrões diferentes de NA's nas linhas para a base dessa temporada transformando a base de dados toda em uma matriz de 0's e 1's, sendo 0 quando a observação tem informação, e 1 quando ela é NA. Assim, cada linha da base se torna um vetor de 0's e 1's, e o padrão de cada linha pode ser identificado concatenando esses 0's e 1's. Feito isso, foram identificados 61 padrões diferentes para a base da temporada 2018/19. Então, para cada padrão, são identificadas as linhas que possuem aquele padrão, e as colunas das variáveis que são NA nesse padrão são retiradas. Por fim, em uma base que contém as observações das temporadas anteriores, retiramos essas mesmas colunas, e depois são retiradas as linhas em que ainda existe algum NA nas colunas que sobraram. Com isso, é possível fazer modelos com essa base das temporadas anteriores, e utilizá-los para fazer previsões para todos os 1230 jogos da temporada regular de 2018/19.

3.3 Casas de Aposta

Em muitos lugares no mundo, existem plataformas onde é possível apostar dinheiro em acontecimentos futuros, e a categoria mais popular é a de esportes. Esses lugares são chamados de casas de aposta, e possuem equipes profissionais que determinam métricas para definir números para as apostas.

A “linha” é um desses artifícios que as casas de aposta dos Estados Unidos usam para equilibrar a aposta em um time em cada partida. Se a casa considera o time A favorito pra ganhar por X pontos, então uma linha é determinada, que faz com que os apostadores apostarem no time A e ganharem, o time A precisa vencer o jogo por uma diferença de pelo menos X pontos. Um exemplo: uma casa de aposta determina que o Golden State Warriors é o favorito em um jogo contra o Portland Trail Blazers por 6.5 pontos, logo, a “linha” é -6.5 para os Warriors e +6.5 para o Trail Blazers, ou seja, quem apostar nos Warriors precisa que o time vença o jogo por uma diferença de 7 pontos ou mais para ganhar a aposta, e quem apostar no Trail Blazers precisa que o time perca por no máximo 6 pontos de diferença, ou vença o jogo, para ganhar a aposta. É comum o uso de meio décimo na linha para evitar empates, mas não é obrigatório, caso uma “linha” seja -7 e o time vença por exatos 7 pontos de diferença, normalmente a aposta é ressarcida ao apostador e nem ele, nem a casa ganha a aposta.

Como é feita a determinação dessas “linhas” é segredo de cada casa de aposta, mas há muitos recursos e pessoal disponíveis, além da possibilidade de levar em conta informações como a situação dos jogadores de cada time, por exemplo, se o jogador principal de um time não vai jogar em um jogo específico, isso é definitivamente levado em conta na determinação da “linha” do jogo.

Utilizando-se da mesma técnica de *web scraping* citada na Seção 3.2, foi possível extrair do site da ESPN Americana (ESPN...,) a “linha” de aposta da maioria dos jogos da temporada. O site da ESPN faz uma média da “linha” de várias casas de aposta diferentes logo antes do início de cada partida, e deixa na página de cada jogo.

4 Resultados

4.1 Resultados Reais da Temporada 2018/19

Para efeito de comparação com as previsões a serem feitas, será listado aqui algumas estatísticas das equipes ao final da temporada regular.

4.1.1 Temporada Regular































Team	W	L	Team	W	L
1  Bucks	60	22	1  Warriors	57	25
2  Raptors	58	24	2  Nuggets	54	28
3  76ers	51	31	3  Trail Blazers	53	29
4  Celtics	49	33	4  Rockets	53	29
5  Pacers	48	34	5  Jazz	50	32
6  Nets	42	40	6  Thunder	49	33
7  Magic	42	40	7  Spurs	48	34
8  Pistons	41	41	8  Clippers	48	34
9  Hornets	39	43	9  Kings	39	43
10  Heat	39	43	10  Lakers	37	45
11  Wizards	32	50	11  Timberwolves	36	46
12  Hawks	29	53	12  Grizzlies	33	49
13  Bulls	22	60	13  Pelicans	33	49
14  Cavaliers	19	63	14  Mavericks	33	49
15  Knicks	17	65	15  Suns	19	63

Figura 3 – Classificação Final da NBA separado por conferência Leste (esquerda) e Oeste (direita)

Na figura 3, é observado o número de vitórias (W) e derrotas (L) de cada time na temporada 2018/19. Os 8 primeiros times de cada conferência se classificaram para os *playoffs*.

Tabela 2 – Média de pontos sofridos por equipe

Fonte: (NBA..., b)

Time	Média de Pontos Sofridos
Indiana Pacers	104.7
Miami Heat	105.9
Memphis Grizzlies	106.1
Utah Jazz	106.5
Orlando Magic	106.6
Denver Nuggets	106.7
Detroit Pistons	107.3
Boston Celtics	108
Toronto Raptors	108.4
Houston Rockets	109.1
Milwaukee Bucks	109.3
San Antonio Spurs	110
Dallas Mavericks	110.1
Portland Trail Blazers	110.5
Oklahoma City Thunder	111.1
Golden State Warriors	111.2
Charlotte Hornets	111.8
Brooklyn Nets	112.3
Philadelphia 76ers	112.5
Chicago Bulls	113.4
Los Angeles Lakers	113.5
New York Knicks	113.8
Minnesota Timberwolves	114
Cleveland Cavaliers	114.1
LA Clippers	114.3
Sacramento Kings	115.3
New Orleans Pelicans	116.8
Phoenix Suns	116.8
Washington Wizards	116.9
Atlanta Hawks	119.4

Na tabela 2, é possível ver que o Indiana Pacers teve a melhor defesa da temporada no quesito pontos, e o Atlanta Hawks a pior.

Tabela 3 – Média de pontos marcados por equipe

Fonte: (NBA..., b)

Time	Média de Pontos Marcados
Milwaukee Bucks	118.1
Golden State Warriors	117.7
New Orleans Pelicans	115.4
Philadelphia 76ers	115.2
LA Clippers	115.1
Portland Trail Blazers	114.7
Oklahoma City Thunder	114.5
Toronto Raptors	114.4
Sacramento Kings	114.2
Washington Wizards	114
Houston Rockets	113.9
Atlanta Hawks	113.3
Minnesota Timberwolves	112.5
Boston Celtics	112.4
Brooklyn Nets	112.2
Los Angeles Lakers	111.8
Utah Jazz	111.7
San Antonio Spurs	111.7
Charlotte Hornets	110.7
Denver Nuggets	110.7
Dallas Mavericks	108.9
Indiana Pacers	108
Phoenix Suns	107.5
Orlando Magic	107.3
Detroit Pistons	107
Miami Heat	105.7
Chicago Bulls	104.9
New York Knicks	104.6
Cleveland Cavaliers	104.5
Memphis Grizzlies	103.5

Da tabela 3, percebe-se que o Milwaukee Bucks foi o time que mais marcou pontos na temporada, e o Memphis Grizzlies o que menos fez.

4.1.2 Playoffs

A final foi disputada por Toronto Raptors e Golden State Warriors.

4.2 Previsões

Utilizando as bases de dados definidas na Metodologia, será possível aplicar os métodos estatísticos propostos na Revisão de Literatura para fazer modelos e obter

previsões para os 1230 jogos da temporada regular de 2018/19.

De início, foram utilizadas todas as bases disponíveis (da temporada 2000/01 até 2017/18) para o ajuste dos modelos. Diferentemente do usual na estatística, a base não foi dividida em treinamento e validação, pois os jogos são uma sequência histórica no tempo, e não faria sentido utilizar jogos do “futuro” para prever jogos que aconteceram antes.

Tabela 4 – Porcentagem de Acerto das previsões dos jogos da temporada 2018/19 para cada método utilizando dados de 2000/01 à 2017/18 na modelagem

Método	Porcentagem de Acerto
Regressão de Probit	0.6723577
Regressão Logística	0.6707317
Análise de Discriminante Linear	0.6682927
Regressão de Probit c/ Forward	0.6682927
Regressão Logística c/ Forward	0.6674797
SVM com $cost = 8$, $gamma = 10^{-4}$	0.6666667
Regressão Linear c/ Forward	0.6658537
Regressão Linear	0.6634146
SVM padrão	0.6577236
Random Forest	0.6373984
Regressão em Árvore	0.6373984
Classificação em Árvore	0.6089431

Na tabela 4, temos a porcentagem de acerto das previsões para cada método utilizado. O melhor resultado foi obtido com a Regressão de Probit, com 0.6723577 de acurácia.

Alguns métodos são mais eficientes computacionalmente, em termos de tempo decorrido para a execução das modelagens, e para esses métodos, é possível realizar as previsões modificando as temporadas a serem utilizadas na modelagem, para ser observada a evolução dos resultados. Esses métodos são: regressão linear, regressão logística, regressão de probit, análise de discriminante linear, regressão em árvore e classificação em árvore.

Nas Tabelas 5 e 6 abaixo, são observadas a porcentagem de acerto das previsões dos jogos da temporada 2018/19 para os métodos citados acima, utilizando diferentes temporadas para as modelagens. A primeira coluna indica que a modelagem foi feita usando os jogos da temporada indicada até a temporada 2017/18. Na primeira linha, que se diz que a temporada de início é 2000/2001, as modelagens foram iguais as usadas na Tabela 4. A última linha indica uma previsão utilizando apenas a temporada 2017/18 na modelagem. O melhor resultado geral foi encontrado utilizando a Regressão Logística quando foram utilizadas as temporadas 2006/2007 à 2017/18 para a modelagem. Em média, a modelagem ser feita utilizando essas temporadas obtém o melhor resultado.

Tabela 5 – Porcentagem de Acerto das previsões dos jogos da temporada 2018/19 para cada método utilizando temporadas diferentes na modelagem - Parte 1

Temporada de Início	Regressão Linear	Regressão Logística	Regressão de Probit
2000/2001	0.663415	0.670732	0.672358
2001/2002	0.664228	0.664228	0.666667
2002/2003	0.669106	0.671545	0.673171
2003/2004	0.667480	0.673171	0.674797
2004/2005	0.667480	0.671545	0.672358
2005/2006	0.668293	0.672358	0.673984
2006/2007	0.674797	0.681301	0.677236
2007/2008	0.669919	0.677236	0.679675
2008/2009	0.669919	0.674797	0.673171
2009/2010	0.661789	0.678049	0.678049
2010/2011	0.669106	0.667480	0.666667
2011/2012	0.667480	0.665041	0.665041
2012/2013	0.660976	0.661789	0.663415
2013/2014	0.660976	0.665041	0.660976
2014/2015	0.668293	0.672358	0.672358
2015/2016	0.653659	0.653659	0.647968
2016/2017	0.648781	0.644715	0.641463
2017/2018	0.605691	0.594309	0.595122

Tabela 6 – Porcentagem de Acerto das previsões dos jogos da temporada 2018/19 para cada método utilizando temporadas diferentes na modelagem - Parte 2

Temporada de Início	Análise de Discriminante Linear	Regressão em Árvore	Classificação em Árvore
2000/2001	0.668293	0.637398	0.608943
2001/2002	0.665041	0.639837	0.610569
2002/2003	0.672358	0.639837	0.635772
2003/2004	0.673984	0.637398	0.634146
2004/2005	0.671545	0.642276	0.642276
2005/2006	0.675610	0.627642	0.642276
2006/2007	0.678862	0.624390	0.644715
2007/2008	0.678862	0.623577	0.647968
2008/2009	0.674797	0.641463	0.646342
2009/2010	0.674797	0.638211	0.612195
2010/2011	0.664228	0.639837	0.614634
2011/2012	0.665041	0.640650	0.639837
2012/2013	0.661789	0.639024	0.613008
2013/2014	0.661789	0.629268	0.613008
2014/2015	0.670732	0.608130	0.607317
2015/2016	0.646342	0.627642	0.591057
2016/2017	0.639837	0.617886	0.590244
2017/2018	0.599187	0.589431	0.573171

Dado os resultados das tabelas acima, serão comparadas novamente as porcentagens de acertos das previsões para todos os métodos, mas dessa vez utilizando apenas as temporadas de 2006/2007 à 2017/2018 na modelagem.

Tabela 7 – Porcentagem de Acerto das previsões dos jogos da temporada 2018/19 para cada método utilizando dados de 2006/07 à 2017/18 na modelagem

Método	Porcentagem de Acerto
Regressão Logística	0.6813008
Análise de Discriminante Linear	0.6788618
Regressão de Probit	0.6772358
Regressão Linear	0.6747967
SVM com $cost = 8$, $gamma = 10^{-4}$	0.6731707
Regressão Linear c/ Forward	0.6707317
Regressão Logística c/ Forward	0.6682927
Regressão de Probit c/ Forward	0.6642276
SVM padrão	0.6569106
Classificação em Árvore	0.6447154
Random Forest	0.6373984
Regressão em Árvore	0.6243902

Comparando a Tabela 4 com a Tabela 7, é observado que na Tabela 7 a maioria dos métodos obtém melhor porcentagem de acerto das previsões, e o melhor resultado geral é o de Regressão Logística quando utilizadas as temporadas de 2006/07 à 2017/2018 na modelagem.

4.2.1 “Previsões” das casas de aposta

Considerando o que foi descrito na seção 3.3, a “linha” de aposta de cada jogo diz qual time era considerado o favorito para vencer o jogo pelas casas de aposta, e podemos obter a porcentagem de acerto média das casas de aposta para os jogos da temporada.

Na página da *web* de 4 dos 1230 jogos da temporada, o valor da “linha” não estava disponível, e não há como recuperar essa informação. Em outros 16 jogos, a “linha” era *even*, ou seja, esses jogos foram julgados tão equilibrados, que em média não foram apontados times favoritos, ou seja, se não foi determinado um time favorito, não seria justo contar esses jogos para as “previsões” das casas de aposta, pois não há possibilidade de acerto.

Excluindo esses 20 jogos citados acima, a porcentagem de acerto calculada para as casas de aposta foi de 0.6727273, em 1210 jogos.

Considerando o melhor modelo obtido nesse trabalho, a taxa de acerto das previsões foi melhor do que as das casas de aposta.

4.3 Tempo de execução computacional de cada método

Para efeito de comparação, foi medido o tempo necessário para a execução computacional dos códigos de modelagem e previsão para cada método. Foi feita a modelagem das temporadas 2006/07 à 2017/18 para todos os métodos.

É lembrado que para cada método é feita não só uma, mas sim 61 modelagens, como visto na seção 3.2.1, por conta dos valores faltantes, para ser possível ter a previsão de todos os jogos da temporada 2018/19.

Os resultados estão apresentados na Tabela 8.

Tabela 8 – Tempo de execução do código computacional para cada método

Método	Tempo (em segundos)
Regressão Linear	9.733
Classificação em Árvore	19.692
Regressão em Árvore	21.069
Regressão Logística	30.542
Regressão de Probit	33.780
Análise de Discriminante Linear	35.057
Regressão Linear c/ Forward	985.550
Regressão de Probit c/ Forward	5359.306
Regressão Logística c/ Forward	6095.090
SVM	9420.622
Random Forest	31367.020

4.4 Modelo “campeão”

Como visto na seção 4.2, o modelo com o melhor resultado geral das previsões foi o de Regressão Logística quando utilizadas as temporadas de 2006/07 à 2017/2018 na modelagem. Além disso, é um dos métodos mais rápidos computacionalmente, demorando em torno de 30 segundos de execução para serem obtidos esses resultados.

Nessa seção serão analisados alguns aspectos das previsões obtidas por esse modelo.

4.4.1 Acertos por Equipe

Tabela 9 – Totais dos acertos das previsões por time separados por categorias - Parte 1

Time	Total de Acertos das Previsões	Acertos nos jogos em casa	Acertos nos jogos fora	Acertos nas vitórias	Vitórias	Acertos nas derrotas	Derrotas
Cleveland Cavaliers	64	29	35	4	19	60	63
Phoenix Suns	62	29	33	2	19	60	63
Denver Nuggets	61	35	26	48	54	13	28
Dallas Mavericks	61	28	33	24	33	37	49
Indiana Pacers	61	29	32	46	48	15	34
Toronto Raptors	61	31	30	53	58	8	24
Portland Trail Blazers	61	32	29	43	53	18	29
New York Knicks	59	27	32	4	17	55	65
Detroit Pistons	59	27	32	25	41	34	41
Minnesota Timberwolves	58	27	31	20	36	38	46
Sacramento Kings	58	27	31	22	39	36	43
Los Angeles Clippers	58	26	32	39	48	19	34
Atlanta Hawks	57	26	31	7	29	50	53
Chicago Bulls	57	28	29	3	22	54	60
Milwaukee Bucks	56	33	23	54	60	2	22
Golden State Warriors	56	30	26	48	57	8	25
Utah Jazz	55	29	26	41	50	14	32
Boston Celtics	55	27	28	41	49	14	33
San Antonio Spurs	55	28	27	35	48	20	34
Brooklyn Nets	54	27	27	23	42	31	40
Charlotte Hornets	54	26	28	28	39	26	43
Orlando Magic	54	29	25	23	42	31	40
Memphis Grizzlies	53	23	30	17	33	36	49
Washington Wizards	53	23	30	16	32	37	50
Philadelphia 76ers	52	32	20	38	51	14	31
New Orleans Pelicans	50	27	23	20	33	30	49
Houston Rockets	50	31	19	36	53	14	29
Oklahoma City Thunder	48	27	21	37	49	11	33
Miami Heat	48	21	27	21	39	27	43
Los Angeles Lakers	46	24	22	20	37	26	45

Tabela 10 – Totais dos acertos das previsões por time separados por categorias - Parte 2

Time	Acertos nas vitórias fora	Vitórias fora	Acertos nas derrotas fora	Derrotas fora	Acertos nas vitórias em casa	Vitórias em casa	Acertos nas derrotas em casa	Derrotas em casa
Cleveland Cavaliers	0	6	35	35	4	13	25	28
Phoenix Suns	0	7	33	34	2	12	27	29
Denver Nuggets	14	20	12	21	34	34	1	7
Dallas Mavericks	4	9	29	32	20	24	8	17
Indiana Pacers	17	19	15	22	29	29	0	12
Toronto Raptors	22	26	8	15	31	32	0	9
Portland Trail Blazers	15	21	14	20	28	32	4	9
New York Knicks	1	8	31	33	3	9	24	32
Detroit Pistons	8	15	24	26	17	26	10	15
Minnesota Timberwolves	4	11	27	30	16	25	11	16
Sacramento Kings	7	15	24	26	15	24	12	17
Los Angeles Clippers	17	22	15	19	22	26	4	15
Atlanta Hawks	2	12	29	29	5	17	21	24
Chicago Bulls	1	13	28	28	2	9	26	32
Milwaukee Bucks	21	27	2	14	33	33	0	8
Golden State Warriors	18	27	8	14	30	30	0	11
Utah Jazz	14	21	12	20	27	29	2	12
Boston Celtics	14	21	14	20	27	28	0	13
San Antonio Spurs	10	16	17	25	25	32	3	9
Brooklyn Nets	6	19	21	22	17	23	10	18
Charlotte Hornets	8	14	20	27	20	25	6	16
Orlando Magic	5	17	20	24	18	25	11	16
Memphis Grizzlies	5	12	25	29	12	21	11	20
Washington Wizards	3	10	27	31	13	22	10	19
Philadelphia 76ers	9	20	11	21	29	31	3	10
New Orleans Pelicans	5	14	18	27	15	19	12	22
Houston Rockets	10	22	9	19	26	31	5	10
Oklahoma City Thunder	12	22	9	19	25	27	2	14
Miami Heat	9	20	18	21	12	19	9	22
Los Angeles Lakers	5	15	17	26	15	22	9	19

Tabela 11 – Porcentagens dos acertos das previsões por time separados por categorias

Time	% Acerto Total	% Acerto Casa	% Acerto Fora	% Acerto Vitórias	% Acerto Derrotas	% Acerto Vitórias Fora	% Acerto Derrotas Fora	% Acerto Vitórias Casa	% Acerto Derrotas Casa
Cleveland Cavaliers	0.780	0.707	0.854	0.211	0.952	0.000	1.000	0.308	0.893
Phoenix Suns	0.756	0.707	0.805	0.105	0.952	0.000	0.971	0.167	0.931
Denver Nuggets	0.744	0.854	0.634	0.889	0.464	0.700	0.571	1.000	0.143
Dallas Mavericks	0.744	0.683	0.805	0.727	0.755	0.444	0.906	0.833	0.471
Indiana Pacers	0.744	0.707	0.780	0.958	0.441	0.895	0.682	1.000	0.000
Toronto Raptors	0.744	0.756	0.732	0.914	0.333	0.846	0.533	0.969	0.000
Portland Trail Blazers	0.744	0.780	0.707	0.811	0.621	0.714	0.700	0.875	0.444
New York Knicks	0.720	0.659	0.780	0.235	0.846	0.125	0.939	0.333	0.750
Detroit Pistons	0.720	0.659	0.780	0.610	0.829	0.533	0.923	0.654	0.667
Minnesota Timberwolves	0.707	0.659	0.756	0.556	0.826	0.364	0.900	0.640	0.688
Sacramento Kings	0.707	0.659	0.756	0.564	0.837	0.467	0.923	0.625	0.706
Los Angeles Clippers	0.707	0.634	0.780	0.812	0.559	0.773	0.789	0.846	0.267
Atlanta Hawks	0.695	0.634	0.756	0.241	0.943	0.167	1.000	0.294	0.875
Chicago Bulls	0.695	0.683	0.707	0.136	0.900	0.077	1.000	0.222	0.812
Milwaukee Bucks	0.683	0.805	0.561	0.900	0.091	0.778	0.143	1.000	0.000
Golden State Warriors	0.683	0.732	0.634	0.842	0.320	0.667	0.571	1.000	0.000
Utah Jazz	0.671	0.707	0.634	0.820	0.438	0.667	0.600	0.931	0.167
Boston Celtics	0.671	0.659	0.683	0.837	0.424	0.667	0.700	0.964	0.000
San Antonio Spurs	0.671	0.683	0.659	0.729	0.588	0.625	0.680	0.781	0.333
Brooklyn Nets	0.659	0.659	0.659	0.548	0.775	0.316	0.955	0.739	0.556
Charlotte Hornets	0.659	0.634	0.683	0.718	0.605	0.571	0.741	0.800	0.375
Orlando Magic	0.659	0.707	0.610	0.548	0.775	0.294	0.833	0.720	0.688
Memphis Grizzlies	0.646	0.561	0.732	0.515	0.735	0.417	0.862	0.571	0.550
Washington Wizards	0.646	0.561	0.732	0.500	0.740	0.300	0.871	0.591	0.526
Philadelphia 76ers	0.634	0.780	0.488	0.745	0.452	0.450	0.524	0.935	0.300
New Orleans Pelicans	0.610	0.659	0.561	0.606	0.612	0.357	0.667	0.789	0.545
Houston Rockets	0.610	0.756	0.463	0.679	0.483	0.455	0.474	0.839	0.500
Oklahoma City Thunder	0.585	0.659	0.512	0.755	0.333	0.545	0.474	0.926	0.143
Miami Heat	0.585	0.512	0.659	0.538	0.628	0.450	0.857	0.632	0.409
Los Angeles Lakers	0.561	0.585	0.537	0.541	0.578	0.333	0.654	0.682	0.474

Nas tabelas 9 e 10, são apresentados o total de acertos das previsões por time, e separados por algumas categorias: apenas nos jogos em casa, nos jogos fora de casa, nas vitórias, nas derrotas, e nas vitórias e derrotas separadas por fora de casa ou em casa. Também nas tabelas se encontram o total de jogos que o time teve em cada categoria, ou seja, o total de vitórias, de derrotas e vitórias e derrotas em casa ou fora de casa. É lembrado que cada time joga 82 jogos no total, sendo 41 em casa e 41 fora de casa.

Na tabela 11, são apresentadas as mesmas informações das tabelas anteriores, mas em porcentagem.

4.4.1.1 Comparação da tabela de previsões com a tabela real

A tabela de classificação final da temporada regular foi apresentada na figura 3, agora, ela será comparada com a previsão do número de vitórias de cada equipe segundo o modelo.

Tabela 12 – Comparação das vitórias reais com as vitórias previstas - Conferência Leste

Time	Vitórias Reais	Vitórias Previstas
Milwaukee Bucks	60	74
Toronto Raptors	58	69
Philadelphia 76ers	51	59
Boston Celtics	49	55
Indiana Pacers	48	54
Brooklyn Nets	42	45
Orlando Magic	42	39
Detroit Pistons	41	39
Charlotte Hornets	39	36
Miami Heat	39	32
Washington Wizards	32	28
Atlanta Hawks	29	14
Chicago Bulls	22	10
Cleveland Cavaliers	19	9
New York Knicks	17	5

Tabela 13 – Comparação das vitórias reais com as vitórias previstas - Conferência Oeste

Time	Vitórias Reais	Vitórias Previstas
Golden State Warriors	57	65
Denver Nuggets	54	65
Portland Trail Blazers	53	63
Houston Rockets	53	60
Utah Jazz	50	59
Oklahoma City Thunder	49	54
Los Angeles Clippers	48	51
San Antonio Spurs	48	49
Sacramento Kings	39	37
Los Angeles Lakers	37	32
Minnesota Timberwolves	36	32
Dallas Mavericks	33	30
Memphis Grizzlies	33	29
New Orleans Pelicans	33	29
Phoenix Suns	19	7

É possível perceber pelas tabelas 12 e 13, que geralmente o modelo prevê mais vitórias do que o verdadeiro para os times da parte de cima da tabela e mais derrotas do que o real para os times da parte de baixo da tabela. Isso é um padrão esperado, pois não é fácil prever “zebras”, isto é, quando um time com números piores acaba vencendo um time com números melhores, o que na prática acontece de vez em quando.

Para contexto, nessa temporada, do jogo de número 300 pra frente, em que a maioria dos times já jogou pelo menos 20 jogos na temporada, ocorreram 434 jogos em que o time mandante tinha mais vitórias no campeonato do que o time visitante, desses 434 jogos, em 107 o time visitante conseguiu a vitória (24.65%), e desses 107, o modelo campeão conseguiu prever essa “zebra” em apenas 14 jogos (13.08%). Com esse exemplo é fácil ver a dificuldade do acerto da previsão em jogos que acontecem resultados improváveis.

Mas também percebe-se que a classificação de ambas as conferências terminaria na mesma ordem se consideradas as vitórias previstas pelo modelo.

4.5 Acertos das Previsões Durante a Temporada

Na figura 4, está representada a porcentagem de acerto das previsões durante a temporada, do modelo campeão e das “previsões” das casas de aposta, para efeitos de comparação entre os dois.

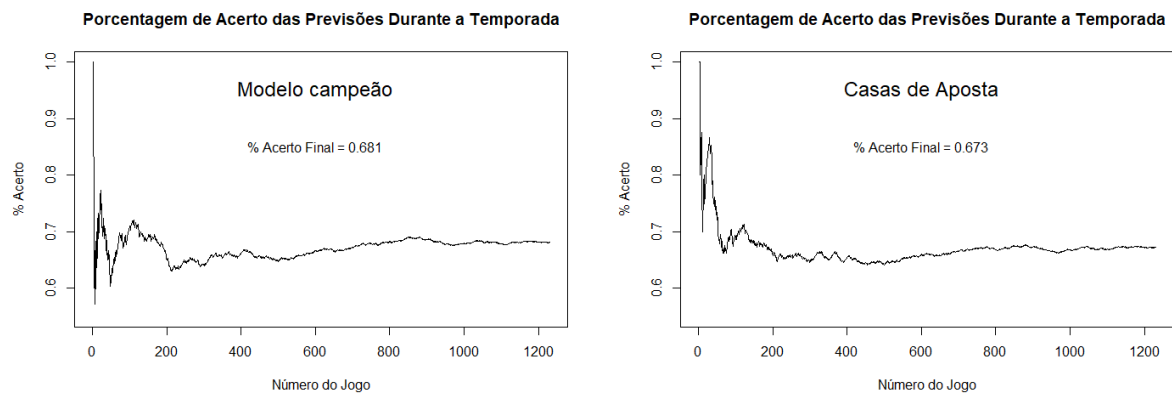


Figura 4 – Evolução da porcentagem de acerto das previsões ao longo da temporada

No começo da temporada, a taxa é bem errática, pois é uma amostra pequena. Percebe-se que para as casas de aposta, a taxa de acerto se estabiliza um pouco mais rápido, o que é um indicativo de que para eles a taxa de acerto é mais constante.

Na figura 5, é colocada a porcentagem de acerto das previsões nos últimos 61 jogos (aproximadamente 5% de 1230), para cada jogo a partir do 61°. Por exemplo, no ponto 61, está representada a porcentagem de acerto das previsões dos jogos 1 ao 61, no ponto 62 está representada a porcentagem de acerto das previsões dos jogos 2 ao 62, e assim por diante.

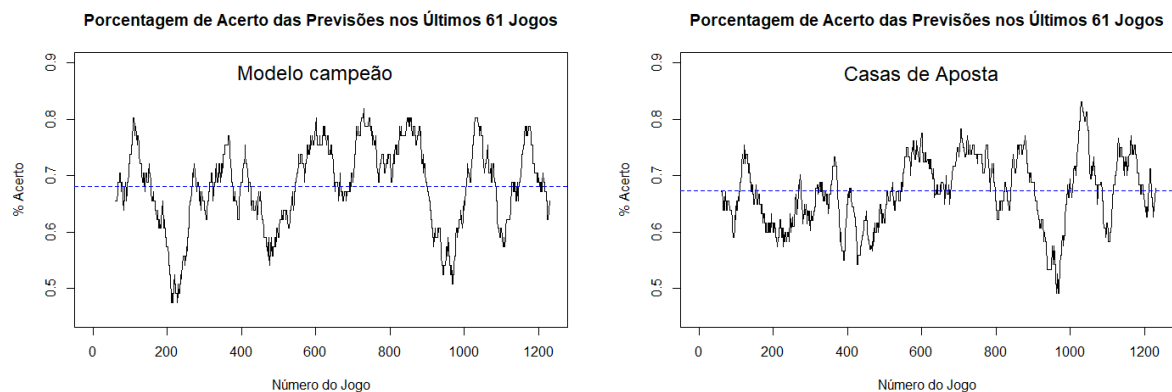


Figura 5 – Porcentagem de acerto das previsões dos últimos 61 jogos ao longo da temporada

A linha azul nesses gráficos representa a porcentagem de acerto das previsões ao final da temporada, os resultados vistos anteriormente no trabalho.

Percebe-se nesses gráficos que alguns padrões são repetidos tanto para o modelo campeão quanto para as previsões das casas de aposta, mas a porcentagem das casas de aposta é bem mais constante, com bem menos valores extremos no gráfico.

A queda brusca no gráfico do modelo campeão, em torno do jogo 200, pode ser

explicada pela falta de informações no começo da temporada. Como foi visto anteriormente, há muitos valores NA no banco de dados no começo da temporada.

A queda percebida nos dois gráficos em torno do jogo 900 pode ter relação com o fim do período de trocas de jogadores entre os times da NBA, pois muitos times trocam jogadores perto do fim desse período, que nessa temporada foi após o jogo de número 816. Nos jogos subsequentes à esse, os times estão se adaptando aos jogadores que chegaram recentemente e como jogar sem os que saíram, consequentemente, alguns times que eram bons ficam piores, e vice-versa, dificultando o acerto das previsões para os modelos.

4.6 Variáveis mais significativas

Para o modelo campeão, foram usadas todas as variáveis do banco de dados na modelagem, e usando a função *standardize* do pacote de mesmo nome (EAGER, 2017), podemos padronizar os dados para mais fácil interpretação dos valores dos parâmetros das variáveis.

Com isso, na tabela 14 seguem as variáveis com p-valor de significância menor que 0.1 na modelagem.

Tabela 14 – Variáveis mais significativas no modelo

Variável	Estimativa do Parâmetro β	Erro Padrão	Z (Estatística de Teste)	p-valor
Mean_Pts_A_T_Vis	-0.34173	0.120916	-2.826	0.00471
Min_Last5home_Home	-0.18985	0.07072	-2.685	0.00726
Loss_T_Vis	-0.59691	0.227052	-2.629	0.00856
Days_LG_Vis	0.062418	0.024913	2.505	0.01223
Mean_Last3_home_opp_Home	-0.34262	0.142105	-2.411	0.01591
Mean_Pts_S_T_Vis	0.295485	0.123971	2.383	0.01715
Min_Last3home_opp_Home	0.176757	0.075284	2.348	0.01888
Mean_Pts_A_T_Home	0.312476	0.133186	2.346	0.01897
Max_Last3_home_opp_Home	0.163982	0.077887	2.105	0.03526
OT_last_HomeTRUE	0.096106	0.045954	2.091	0.0365
Min_Last5total_opp_Home	0.145477	0.069784	2.085	0.0371
Win_Last3_total_Home	0.142964	0.068954	2.073	0.03814
Str_Sch_Vis	0.052763	0.025495	2.069	0.0385
Max_Pts_S_H_Home	-0.08071	0.041444	-1.947	0.05147
Loss_A_Vis	0.231707	0.122455	1.892	0.05847
Wins_H_Home	0.23373	0.126269	1.851	0.06416
Max_Last10_total_Vis	-0.10887	0.059288	-1.836	0.06631
Max_Last5_away_Vis	0.124937	0.072045	1.734	0.08289
Max_Pts_A_H_Home	0.069059	0.040352	1.711	0.087
Days_LG_Home	-0.04212	0.024785	-1.699	0.08924
Min_Last10total_Vis	-0.09652	0.05732	-1.684	0.09221

Isso foi feito para o modelo campeão, que é uma regressão logística, e a variável resposta usada foi a *Win_Vis* (a explicação do que é cada variável se encontra na seção 3.2),

isso quer dizer que parâmetros positivos indicam que a variável contribui para o aumento da probabilidade de vitória do time visitante, e intuitivamente, parâmetros negativos indicam que a variável contribui para o aumento da probabilidade de vitória do time mandante.

A variável mais significativa foi *Mean_Pts_A_T_Vis*. A estimativa do parâmetro dessa variável é negativa, o que significa que quando maior a média de pontos sofridos do time visitante, maior a probabilidade de vitória do time mandante, o que faz todo sentido.

4.7 Adicionando jogos de 2018/19 na modelagem

As previsões até aqui foram feitas sem colocar nenhum jogo da temporada 2018/19 na modelagem, pois para fazer as previsões inserindo os jogos da temporada na modelagem conforme os jogos vão acontecendo aumentaria muito o número de modelos necessários para realizar as previsões, e devido ao tempo de execução de alguns métodos seria impossível implementar dessa maneira para todos eles.

Portanto, foi decidido realizar as previsões dessa maneira apenas para o modelo campeão. Como acontecem vários jogos no mesmo dia, e muitas vezes no mesmo horário, as previsões foram sendo feitas para os jogos de cada dia em que houve partidas, e as partidas já realizadas até o dia anterior foram sendo adicionadas na modelagem.

Isso foi feito de duas maneiras: a primeira foi deixando todos os jogos de 2006/07 até 2017/18 e apenas adicionando os jogos de 2018/19, e a outra maneira foi retirando o mesmo número de jogos que foram sendo adicionados, ou seja, conforme foram entrando as partidas de 2018/19, as partidas mais antigas de 2006/07 foram saindo, desse jeito, o números de jogos para a modelagem se manteve constante.

Do primeiro jeito, a porcentagem de acerto das previsões foi 0.6756098, e do segundo foi 0.6731707. Surpreendentemente, ambos os resultados foram piores do que o obtido sem inserir nenhum jogo da temporada 2018/19 na modelagem.

5 Conclusão

O objetivo desse trabalho foi obter previsões para os jogos da temporada regular da NBA de 2018/19, e os melhores resultados obtidos foram superiores aos inferidos de casas de aposta dos Estados Unidos, o que indica que foram bons resultados.

Os métodos de Regressão Linear, Regressão Logística, Regressão de Probit, e a Análise de Discriminante Linear se mostraram os melhores tanto em relação ao acerto das previsões, quanto em relação ao tempo necessário para a execução dos códigos computacionais.

A grande desvantagem da base de dados aplicada nas modelagens realizadas nesse trabalho é a falta de informação sobre os jogadores, pois dessa maneira não é possível ter a influência de quando um jogador importante não vai jogar um jogo específico, quando acontecem lesões, quando um jogador que começou a temporada machucado volta a jogar, ou quando um jogador troca um time por outro durante a temporada. Todas essas informações levam tempo para a base de dados se “atualizar” sozinha, pois jogadores importantes impactam bastante os números dos times, por isso as variáveis que consideram as estatísticas dos últimos jogos são muito importantes.

A implementação pode ser replicada para realizar previsões de jogos de temporadas futuras.

Referências

- AGRESTI, A. *An Introduction to Categorical Data Analysis*. Wiley, 2007. (Wiley Series in Probability and Statistics). ISBN 9780470114742. Disponível em: <<https://books.google.com.br/books?id=OG9Eqwd0Fh4C>>. Nenhuma citação no texto.
- BASKETBALL Reference. <<https://www.basketball-reference.com/>>. Acessado em: 28/09/2018. Citado na página 25.
- BREIMAN, L. Random forests. 2001. Acessado em: 01/06/2019. Disponível em: <<https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf>>. Citado na página 24.
- CARVALHO, J. et al. *ANÁLISE DE PROBIT APLICADA A BIOENSAIOS COM INSETOS*. [S.l.: s.n.], 2017. ISBN 978-85-64937-08-6. Citado na página 19.
- EAGER, C. D. *standardize: Tools for Standardizing Variables for Regression in R*. [S.l.], 2017. R package version 0.2.1. Disponível em: <<https://CRAN.R-project.org/package=standardize>>. Citado na página 46.
- ESPN NBA Scores. <<http://www.espn.com/nba/scoreboard>>. Acessado em: 16/05/2019. Citado na página 31.
- GROTHENDIECK, G. *sqlf: Manipulate R Data Frames Using SQL*. [S.l.], 2017. R package version 0.4-11. Disponível em: <<https://CRAN.R-project.org/package=sqlf>>. Nenhuma citação no texto.
- JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, 2007. (Applied Multivariate Statistical Analysis). ISBN 9780131877153. Disponível em: <<https://books.google.com.br/books?id=gFWcQgAACAAJ>>. Citado na página 22.
- KASSAMBARA, A. *Machine Learning Essentials: Practical Guide in R*. CreateSpace Independent Publishing Platform, 2018. ISBN 9781986406857. Disponível em: <<https://books.google.com.br/books?id=745QDwAAQBAJ>>. Citado na página 19.
- KUTNER, M.; NACHTSHEIM, C.; NETER, J. *Applied Linear Regression Models*. McGraw-Hill Higher Education, 2003. (The McGraw-Hill/Irwin Series Operations and Decision Sciences). ISBN 9780072955675. Disponível em: <<https://books.google.com.br/books?id=0nAMAAAACAAJ>>. Nenhuma citação no texto.
- LIAW, A.; WIENER, M. Classification and regression by randomforest. *R News*, v. 2, n. 3, p. 18–22, 2002. Disponível em: <<https://CRAN.R-project.org/doc/Rnews/>>. Citado na página 25.
- MEYER, D. Support vector machines, the interface to libsvm in package e1071. 2019. Acessado em: 31/05/2019. Disponível em: <<https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf>>. Citado na página 21.
- MEYER, D. et al. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. [S.l.], 2018. R package version 1.7-0.

Disponível em: <<https://CRAN.R-project.org/package=e1071>>. Citado 2 vezes nas páginas 21 e 25.

NBA Logos. <http://www.sportslogos.net/teams/list_by_league/6/National_Basketball_Association/NBA/logos/>. Acessado em: 13/10/2018. Nenhuma citação no texto.

NBA Stats. <<https://stats.nba.com/>>. Acessado em: 28/09/2018. Citado 2 vezes nas páginas 34 e 35.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2018. Disponível em: <<https://www.R-project.org/>>. Citado na página 25.

RIPLEY, B. *tree: Classification and Regression Trees*. [S.l.], 2019. R package version 1.0-40. Disponível em: <<https://CRAN.R-project.org/package=tree>>. Citado na página 25.

RIPLEY, B. D. *Pattern Recognition and Neural Networks*. [S.l.]: Cambridge University Press, 1996. Citado na página 23.

SCHUMAKER, R. P.; SOLIEMAN, O. K.; CHEN, H. *Sports Data Mining*. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2010. ISBN 144196729X, 9781441967299. Citado na página 15.

SELECTOR Gadget. <<https://selectorgadget.com/>>. Acessado em: 25/05/2019. Citado na página 25.

SUÁREZ, E. et al. Selection of variables in a multiple linear regression model. In: _____. *Applications of Regression Models in Epidemiology*. John Wiley & Sons, Ltd, 2017. cap. 5, p. 77–86. ISBN 9781119212515. Disponível em: <<https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119212515.ch5>>. Citado na página 19.

UUDMAE, J. Predicting nba game outcomes. Acessado em: 28/09/2018. Disponível em: <<http://cs229.stanford.edu/proj2017/final-reports/5231214.pdf>>. Nenhuma citação no texto.

VAPNIK, V. N. *The Nature of Statistical Learning Theory*. Berlin, Heidelberg: Springer-Verlag, 1995. ISBN 0-387-94559-8. Citado na página 20.

VENABLES, W. N.; RIPLEY, B. D. *Modern Applied Statistics with S*. Fourth. New York: Springer, 2002. ISBN 0-387-95457-0. Disponível em: <<http://www.stats.ox.ac.uk/pub/MASS4>>. Citado na página 25.

WICKHAM, H. *rvest: Easily Harvest (Scrape) Web Pages*. [S.l.], 2016. R package version 0.3.2. Disponível em: <<https://CRAN.R-project.org/package=rvest>>. Citado na página 25.

WICKHAM, H. *stringr: Simple, Consistent Wrappers for Common String Operations*. [S.l.], 2019. R package version 1.4.0. Disponível em: <<https://CRAN.R-project.org/package=stringr>>. Nenhuma citação no texto.

WICKHAM, H. et al. *dplyr: A Grammar of Data Manipulation*. [S.l.], 2018. R package version 0.7.8. Disponível em: <<https://CRAN.R-project.org/package=dplyr>>. Nenhuma citação no texto.

YAN, X.; SU, X. G. *Linear Regression Analysis: Theory and Computing*. River Edge, NJ, USA: World Scientific Publishing Co., Inc., 2009. ISBN 9789812834102, 9812834109. Citado na página 18.

Apêndices

APÊNDICE A – Códigos em R

Quisque facilisis auctor sapien. Pellentesque gravida hendrerit lectus. Mauris rutrum sodales sapien. Fusce hendrerit sem vel lorem. Integer pellentesque massa vel augue. Integer elit tortor, feugiat quis, sagittis et, ornare non, lacus. Vestibulum posuere pellentesque eros. Quisque venenatis ipsum dictum nulla. Aliquam quis quam non metus eleifend interdum. Nam eget sapien ac mauris malesuada adipiscing. Etiam eleifend neque sed quam. Nulla facilisi. Proin a ligula. Sed id dui eu nibh egestas tincidunt. Suspendisse arcu.