Desempenho Individual no *College Basketball* e Sucesso Profissional na *National Basketball Association*: uma Abordagem de Aprendizagem de Máquina

Enzo Bertoldi Oestreich
IME-UFRGS
Porto Alegre-RS, Brasil
enzobertoldi.o@gmail.com

João Henrique Ferreira Flores
IME-UFRGS
Porto Alegre-RS, Brasil
joao.flores@ufrgs.br

Resumo

This study aims to evaluate the predictive power of college basketball statistics for success in the NBA, using players selected in the drafts between the 2010 and 2017 seasons. The goal is to identify the variables that contribute positively or negatively to an athlete's Value Over Replacement Player (VORP), aiming to establish a relationship that assists in identifying selection preferences for the NBA draft.

To analyze the dataset, dimensionality reduction techniques such as Principal Component Analysis (PCA) were applied to evaluate the predictive capacity of the selected variables. Multiple Linear Regression and Artificial Neural Networks were also used for VORP prediction. The results show that, although it is not possible to distinguish a new star using college statistics alone, machine learning techniques serve as an auxiliary tool to improve draft ranking.

Keywords

Basquete, Aprendizagem da Máquina, Modelagem, Predição

1 Introdução

A maior inserção de novos talentos na *National Basketball Association* (NBA) ocorre através do *draft*. Um evento anual, realizado desde 1947, onde as equipes selecionam candidatos advindos do basquete universitário ou internacional. Além do acréscimo de novos jogadores aos plantéis, o *draft* possui o intuito de estabelecer a paridade na liga. De acordo com o formato atual, equipes com as piores campanhas na temporada anterior possuem maiores chances de serem as primeiras a escolher, consequentemente as melhores equipes serão as últimas a selecionarem seus jogadores [1].

Devido a esta definição do ordenamento nas escolhas, a cada temporada novas equipes se submetem ao processo de *tanking* [1] na esperança de obterem as primeiras posições do *draft* e, quem sabe, *draftar* a nova estrela da franquia.

Anualmente, os maiores meios de comunicação esportiva divulgam rankings dos melhores candidatos elegíveis para o *draft* [2]. Com a ampliação da cultura orientada a dados e o alto investimento direcionado a *scouts* [3], pressupõe-se que a primeira escolha seja destinada ao melhor jogador disponível, sendo acompanhado pelo segundo e subsequentemente até a última seleção. Tal premissa não se solidifica como verdade absoluta ao olharmos para o passado, uma vez que inúmeras equipes falharam ao prever a próxima estrela. O fracasso mais recente decorreu no draft de 2013, a equipe do *Cleveland Cavaliers* selecionou Anthony Bennett, versátil e atlético ala de força ([4]), como número 1 da noite. Após 4 temporadas com pontuação média abaixo dos 5 pontos, Anthony não se encontra em nenhum elenco da liga americana e até hoje é considerado a pior primeira escolha da história. Na mesma noite, o décimo quinto jogador a ser selecionado foi Giannis Antetokounmpo que recebeu dois prêmios de *Most Valuable Player* (MVP) e é nome confirmado no hall da fama do basquetebol [5].

Com base nesta recorrência de jogadores ficarem para trás no *draft*, objetiva-se entender quais fatores discriminam um atleta de alto nível em relação aos demais. Para avaliar o possível sucesso de um jogador recém-chegado à liga, fatores externos que possam influenciar o seu desenvolvimento precisam ser levados em conta [6]. Como muitas dessas variáveis são não-mensuráveis, este trabalho visa a analisar dados pré-NBA e sua capacidade de predizer a performance de um novato no basquete profissional. O foco do estudo está especificamente nos dados do basquetebol colegial, visto que cerca de 75% dos atletas selecionados no *draft* provêm do *college*. Desta forma, o trabalho possui três objetivos: (i) Definir o melhor modelo a ser utilizado para predição da variável resposta; (ii) Estabelecer balizador para mensurar o sucesso de um jogador e (iii) Identificar as variáveis que contribuem de forma positiva para a variável resposta.

2 Métodos

Quando tratamos da análise de performance no basquetebol profissional nos deparamos com um impasse que se resume à definição de sucesso: quais estatísticas devem ser utilizadas para parametrizar e separar uma estrela de um atleta de baixo desempenho. Ao utilizarmos o conceito de sucesso não estamos atrelando-o a fama, dado que este pode ser monitorado através de votos do *All Star Game*, vendas de camisetas e outros fatores, mas sim àquele jogador cuja presença no plantel de sua equipe seja indispensável para a prosperidade da mesma.

No intuito de explorar tais métricas de performance, encontramos estudos que utilizam diferentes abordagens para responder a mesma pergunta, revelando uma certa subjetividade no que diz respeito ao conceito de sucesso. No trabalho de [7] temos uma visão baseada em longevidade de carreira onde, a partir de um *threshold* pré-definido pelo autor, separamos os atletas entre bem-sucedidos ou não a partir da quantidade de partidas do mesmo na liga profissional, tendo como resposta uma variável binária; já [8] traz uma visão metrificada de desempenho utilizando como medida de performance a estatística *Win Shares Per 48*, a qual baseia-se no *box score*[2]

[1] Processo no qual equipes propositalmente mantém um desempenho abaixo da média ao longo da temporada para obterem mais chances de possuir a primeira escolha no *draft*

[2] Tabela de estatísticas contabilizadas durante uma partida

para dividir entre os jogadores o crédito pelas vitórias da equipe, buscando estimá-la a partir de dados do basquetebol colegial.

Como notamos acima, a definição de um atleta bem sucedido profissionalmente varia entre autores. Partindo desse pressuposto, será utilizado como balizador de sucesso, o conceito de avaliar o quão essencial um jogador é para sua equipe, ou seja, compreender o impacto que o mesmo possui nas partidas. Considerando o interesse na comparação e avaliação de atletas e na estimação da variável resposta, o sucesso será mensurado por meio de uma métrica quantitativa denominada *Value Over Replacement Player* (VORP).

2.1 *Value Over Replacement Player*

Com o avanço das tecnologias de monitoramento e obtenção de dados referentes a jogadores ao longo das partidas, o chamado *play-by-play data*, surgem, cada vez mais, novas métricas avançadas para avaliação de desempenho, porém estes dados não possuem grande disponibilidade. Deste modo, optou-se pela utilização de uma métrica baseada no *box score* tradicional. Para uma melhor compreensão do VORP é necessário compreender o conceito de *Box Plus/Minus* (BPM), visto que o VORP é um derivado do BPM.

O BPM também é baseado em estatísticas presentes no *box score*, além de levar em conta a posição do atleta e a performance geral da equipe. Esta métrica foi criada por [9] para avaliar o impacto de um jogador em quadra quando comparado a média da liga, estimando o número de pontos por 100 posses de bola. Para visualização de uma escala para esta métrica, visto que quanto maior o seu valor, melhor é o desempenho do jogador, apresenta-se na Tabela 1 uma exemplificação criada pelo desenvolvedor do BPM. O jogador LeBron James, na temporada de 2009-2010, terminou o ano com BPM +11.8, ou seja, a sua equipe era 11.8 pontos melhor (por 100 posses de bola) quando ele estava em quadra.

Tabela 1: Escala de Performance BPM

Escala	Classificação	Interpretação
+10	*all-time season*	Desempenho Histórico
+8	*MVP season*	Desempenho de *MVP*
+6	*all-NBA season*	Desempenho de *all-NBA*
+4	*all-star consideration*	Desempenho de *all-star*
+2	*good starter*	Desempenho de um bom titular
+0	*starter or 6th man*	Desempenho de titular ou sexto homem
-2	*bench player*	Desempenho de reserva

Dado os conceitos apresentados na 1, podemos seguir para a compreensão da variável resposta. O VORP é uma métrica que leva em conta o tempo em quadra de um atleta, convertendo o BPM em uma estimativa da contribuição do jogador para a equipe, tendo como referência o *replacement player*, sendo este jogador considerado um reserva com BPM no valor de -2. Para obtermos o VORP, faz-se:

$$VORP = [BPM - (-2)] \times (\%PP) \times (G/82). \quad (1)$$

Onde *BPM* é o escore; −2 por ser comparado a um jogador definido como *replacement player*; *%PP* é o número de minutos jogados dentre os minutos possíveis e *G* é o número de jogos disputados.

Utilizando a Equação 1, obtém-se o número de pontos que um jogador produz a mais que um *replacement player* por 100 posses da equipe, desta forma pode-se comparar os diferentes atletas de acordo com seu VORP, sendo que, ao seguir esta regra, quanto maior o VORP melhor.

3 Banco de dados

Tabela 2: Variáveis de Estatísticas Individuais do Basquete (NBA e NCAA)

Variável	Descrição
Season	Temporada em que os dados foram registrados.
School	Instituição/Universidade do jogador.
Conf	Conferência em que a equipe competiu (ex: Pac-12).
G	Jogos (Games) disputados.
GS	Jogos como Titular (Games Started).
MP	Minutos Jogados (Minutes Played) por jogo.
FG	Cestas de Campo Feitas (Field Goals made).
FGA	Cestas de Campo Tentadas (Field Goals Attempted).
FG%	Porcentagem de Cestas de Campo (Field Goal Percentage).
2P	Cestas de 2 Pontos Feitas (2-Point Field Goals made).
2PA	Cestas de 2 Pontos Tentadas (2-Point Field Goals Attempted).
2P%	Porcentagem de Cestas de 2 Pontos.
3P	Cestas de 3 Pontos Feitas (3-Point Field Goals made).
3PA	Cestas de 3 Pontos Tentadas (3-Point Field Goals Attempted).
3P%	Porcentagem de Cestas de 3 Pontos.
FT	Lances Livres Feitos (Free Throws made).
FTA	Lances Livres Tentados (Free Throws Attempted).
FT%	Porcentagem de Lances Livres (Free Throw Percentage).
ORB	Rebotes Ofensivos (Offensive Rebounds).
DRB	Rebotes Defensivos (Defensive Rebounds).
TRB	Rebotes Totais (Total Rebounds).
AST	Assistências (Assists).
STL	Roubos de Bola (Steals).
BLK	Tocos (Blocks).
TOV	Desperdícios de Posse de Bola (Turnovers).
PF	Faltas Pessoais (Personal Fouls).
PTS	Pontos (Points) marcados por jogo.

O horizonte de dados para este estudo foram os anos de 2010 à 2017, buscando informações de todos os jogadores *draftados* no período em questão, possuindo um $N = 480$. Para possibilitar a análise

dos dados foram criados 3 bancos de dados distintos, sendo estes construídos de forma a se conectarem entre si através do nome do atleta. O primeiro banco de dados se refere as informações do *draft* (*rank*, *round*, equipe, *college*); já as outras bases se dividem entre dados dos atletas no *college* e na NBA. Todos as informações utilizadas neste trabalho foram coletadas na base de dados pública disponível em [10] através de um *scraper* criado na linguagem Python pelo autor. A Tabela 2 apresenta as principais variáveis utilizadas dos atletas. Assim, pode-se acompanhar a carreira individualmente ao longo do tempo de cada atleta, desde seu período colegial, o *draft* para a NBA e os anos seguintes jogando profissionalmente. Em relação a variável resposta VORP, salienta-se que a mesma é medida de forma anual, ou seja, ao longo das temporadas. Portanto para cada ano de um jogador na NBA ele terá um VORP atribuído.

Tabela 3: VORP de Kevin Durant ao longo de sua carreira

Season	VORP
2007-08	1.3
2008-09	3.8
2009-10	7.5
2010-11	5.3
2011-12	5.8
2012-13	8.9
2013-14	9.6
2014-15	2.8
2015-16	7.8
2016-17	5.7
2017-18	5.5
2018-19	5.1
2020-21	2.7
2021-22	4.8

Para fins de exemplo, na Tabela 3 pode-se ver como esta métrica é coletada ao longo das temporadas em que o jogador esteve em quadra. Assim sendo, o número de observações para cada atleta varia de acordo com sua longevidade na liga profissional.

4 Modelos

Ao todo, dois diferentes modelos serão avaliados neste trabalho: modelo de regressão múltipla e modelos de redes neurais (*feed-forward* com diferentes configurações). Tanto o modelo de regressão como o modelo de redes neurais serão feitos em duas diferentes abordagens: com o conjunto de variáveis originais e com o uso de componentes principais (ACP, análise de componentes principais).

4.1 Análise de Componentes Principais

Dado um conjunto de dados formado por uma série de variáveis correlacionadas, a aplicação de componentes principais visa a sumarização dos dados em uma representação de espaço dimensional reduzido, a qual retém a maior porção de variação possível [11]. Tal processo é realizado através de uma transformação ortogonal dos dados, alterando as coordenadas originais.

Definindo uma série de covariáveis $X_1, X_2, ..., X_p$, os componentes principais Z_i são combinações lineares normalizadas dessas variáveis que seguem o modelo:

$$\begin{aligned} Z_1 &= \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{1p}X_p, \\ Z_2 &= \phi_{21}X_1 + \phi_{22}X_2 + \ldots + \phi_{2p}X_p, \\ &\vdots \\ Z_p &= \phi_{p1}X_1 + \phi_{p2}X_2 + \ldots + \phi_{pp}X_p. \end{aligned} \tag{2}$$

Onde ϕ_p representa a carga de cada componente principal. Para se obter estes componentes deve-se seguir as seguintes restrições:

- A variabilidade explicada por cada componente decresce em relação ao anterior, ou seja, $\text{Var}(Z_1) > \text{Var}(Z_2) > ... > \text{Var}(Z_p)$;
- Cada componente deve ser não-correlacionado com o anterior (ortogonal);
- A soma dos quadrados das cargas ϕ devem somar 1;

No contexto algébrico, pode-se interpretar cada vetor de cargas como a direção no espaço de variáveis onde os dados apresentam maior variabilidade. Outro resultado que se pode obter ao utilizar este método se diz respeito a variáveis latentes, visto que variáveis com altas cargas nos mesmos componentes são correlacionadas entre si, o que leva a suposição, mesmo que de forma empírica, da presença de padrões nos dados que antes não eram perceptíveis [12]. Para a seleção do número de componentes a serem retidos serão utilizados dois métodos distintos descritos abaixo.

- Critério de Kayser: esta abordagem apresentada por Kayser e Guttman retém apenas aqueles componentes cujos autovalores possuem valor próximo ou acima de 1 [13];
- *Scree Plot*: esta abordagem busca encontrar um ponto de corte baseado na suavização do gráfico de variância explicada, retendo todos os componentes antes deste ponto [14];

A Tabela 4 apresenta as variáveis do banco de dados que foram selecionadas para a composição de componentes principais.

Tabela 4: Variáveis selecionadas para o PCA

Abreviação	Descrição
FG%	*Field Goal Percentage*
2PT%	*Two-Point Percentage*
3PT%	*Three-Point Percentage*
FT%	*Free-Throw Percentage*
RPG	*Rebounds Per Game*
APG	*Assists Per Game*
SPG	*Steals Per Game*
BPG	*Blocks Per Game*
TPG	*Turnovers Per Game*
PPG	*Points Per Game*
USG%	*Usage Percentage*

Com a quantidade de componentes definida, é possível analisar a carga de cada componente a fim de identificar quais variáveis que são correlacionadas com os componentes selecionados. Apesar de não se tratar de um contexto de análise fatorial, o formato em que as variáveis se organizaram permite, ao menos de forma empírica, caracterizar cada um dos componentes de acordo com a relação presente entre suas principais variáveis.

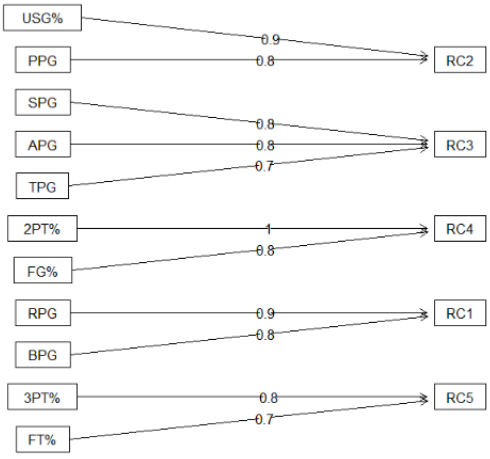


Figura 1: Relação entre Variáveis e Componentes

Baseado na Figura 1, pode-se construir os seguintes conceitos:

- Componente 1: protetores de garrafão;
- Componente 2: alto volume de jogo;
- Componente 3: controle do jogo;
- Componente 4: seleção de arremessos;
- Componente 5: arremessadores;

4.2 Modelo de Regressão Linear Múltipla

Este método é uma generalização da Regressão Linear Simples, visto que o contexto é de duas ou mais variáveis preditoras. A regressão em questão é utilizada para predizer um vetor quantitativo (Y) através do ajuste de uma função para modelar a relação entre o mesmo com as variáveis explicativas (X); tal função assume que a associação entre X e Y é linear ao relação aos parâmetros [11].

Dado um conjunto de dados com p variáveis de entrada, o modelo para análise de regressão múltipla é dado por

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \tag{3}$$

Onde $\beta_1, \ldots, \beta_p$ representam os coeficientes relacionados a cada variável $X_1, \ldots, X_p$ e β_0 o intercepto da equação. Como estes valores são desconhecidos, a estimação é feita utilizando a abordagem de mínimo quadrados ordinários [11].

$$RSS = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip} \right)^2. \tag{4}$$

Onde RSS é a soma do quadrado dos resíduos.

No momento em que se trata de uma abordagem quantitativa, é possível extrair informações referentes a contribuição de cada variável em relação a resposta, facilitando a interpretação dos coeficientes do modelo. Além deste resultado, pode-se obter outras conclusões através da regressão, entre eles:

- Quais variáveis são úteis para explicar a variável dependente;
- Quão bem o modelo se ajusta aos dados;
- Possibilidade de prever novos valores;

4.3 Redes Neurais Artificiais

As Redes Neurais Artificiais, ou simplesmente Redes Neurais, são, como o próprio nome sugere, uma arquitetura computacional formada por uma estrutura de camadas que se assemelha com o cérebro humano, assim como seu funcionamento busca replicar a atividade das sinapses entre neurônios [15]. A estrutura básica de uma rede neural possui 3 camadas, sendo elas 2:

- *Input Layer*: entrada de dados para aprendizado;
- *Hidden Layer*: camada(s) onde os cálculos são realizados e o modelo aprende as relações presentes nos dados;
- *Output Layer*: saída do modelo, variando número de nodos de acordo com a variável resposta;

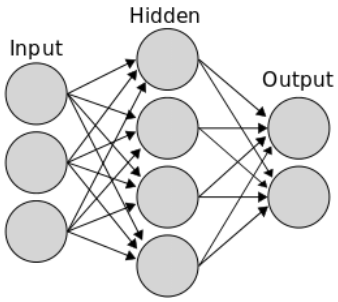


Figura 2: Multi-Layer Network

Dentro das redes neurais tem-se que um neurônio é o componente formado através da composição de peso, *bias* e *input*, onde uma função de ativação é aplicada, sendo este o valor que o neurônio irá repassar para a próxima camada até chegarmos na saída do modelo. Para que a aprendizagem por parte do neurônio aconteça, este processo é repetido inúmeras vezes em conjunto com outras funções, no intuito de obter os pesos que minimizam uma função de custo, normalmente o erro quadrático [15].

A arquitetura selecionada para o estudo em questão é uma das mais clássicas e utilizadas dentre as diversas opções de redes neurais, sendo o modelo na modalidade de *feed forward*, ou alimentadas adiante. Neste modelo as diferentes camadas de neurônios se postulam de maneira sequencial, onde a informação viaja em apenas uma direção, da entrada para a saída dos dados. Dentro deste grupo, encontra-se a classe dos *Multilayer Perceptrons*, objeto de interesse para este trabalho. Embora o significado de MLP se confunda diretamente com qualquer arquitetura *Feed Forward*, o mesmo tem suas particularidades. Em [15], o MLP é definido como uma rede neural com uma ou mais camadas ocultas, sendo a generalização do *Single Layer Perceptron*, onde cada camada é totalmente conectada com a camada subsequente. Utiliza-se o algoritmo *backpropagation* para correção e ajuste de pesos utilizados no processo de aprendizagem da rede neural, visando a redução do erro preditivo. A utilização do termo propagação reversa ocorre devido a ordem que este ajuste ocorre, sendo partir da última camada, procedendo até a camada inicial.

4.4 Avaliação dos modelos

Para avaliar a assertividade da predição dos modelos ajustados neste trabalho, serão utilizadas três métricas distintas que buscam

mensurar o erro nestas predições. Dado as observações do banco definidas como y_i e as predições do modelo $\hat{y}_i$, temos:

- MAE: média do valor absoluto da diferença entre valores observados e preditos;

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}| \tag{5}$$

- MSE: média do quadrado da diferença entre valores observados e preditos;

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{6}$$

- RSME: raiz do MSE;

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

No intuito de obter o melhor modelo de acordo com seu poder preditivo, minimizando um possível *overfitting*, opta-se por um ajuste com base em uma validação cruzada de 5 grupos *5-fold cross-validation*.

5 Resultados

Inicialmente foi realizado uma análise descritiva do banco de dados para apresentar algumas características dos atletas presentes no *draft*, assim como explorar a variável resposta quando avaliada em conjunto de outras variáveis.

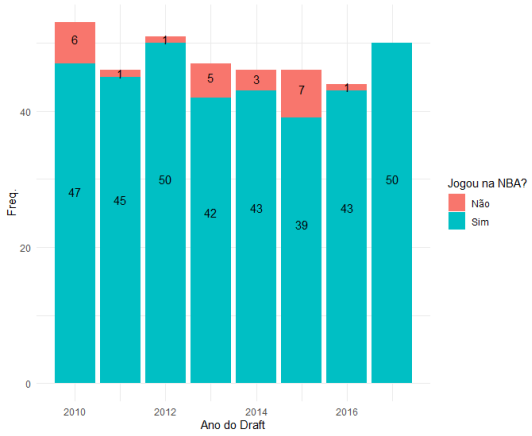


Figura 3: Frequência (Freq.) de atletas *draftados* com passagem pelo *college*

Como o *draft* é composto de 60 escolhas divididas em duas rodadas (30 escolhas cada *round*), pode-se observar na Figura 3 que cerca de 75% dos atletas selecionados no processo de *draft* são advindos do *college*, sendo esta a principal fonte de novos talentos na liga. Tem-se um total de 480 atletas selecionados onde apenas 90 não são advindos do *college*. Além disso, pode-se perceber que apesar do *draft* ser uma coroação do desempenho pré-NBA dos novatos, uma pequena parcela destes jogadores nunca põe o pé em quadra entre os profissionais. Também foi analisada em qual etapa do *draft* se encontram estes jogadores que não tiveram oportunidade na NBA.

Tabela 5: Número de *draftados* por *round* que jogaram (ou não) na NBA

	Jogou na NBA	
Draft Round	Sim	Não
Primeiro	204	0
Segundo	155	24

Através da Tabela 5 nota-se que todos os atletas *draftados* entre 2010 e 2017 que não possuíram uma oportunidade de pisar nas quadras como atleta profissional da NBA foram escolhidos no segundo *round* do *draft*.

Tabela 6: VORP médio por *round*

Round	VORP médio
1	0.572
2	0.043

Corroborando com as afirmações acima, ao se analisar a variável do estudo (VORP) por *rounds* através da Tabela 6, nota-se que esta segue os mesmos padrões, ou seja, atletas selecionados entre as 30 primeiras escolhas possuem maiores oportunidades para se desenvolver e obter sucesso na liga profissional se comparados com atletas de segundo *round*. De forma a se ter uma visualização gráfica da variável de interesse deste trabalho, tem-se o histograma apresentado na Figura 4.

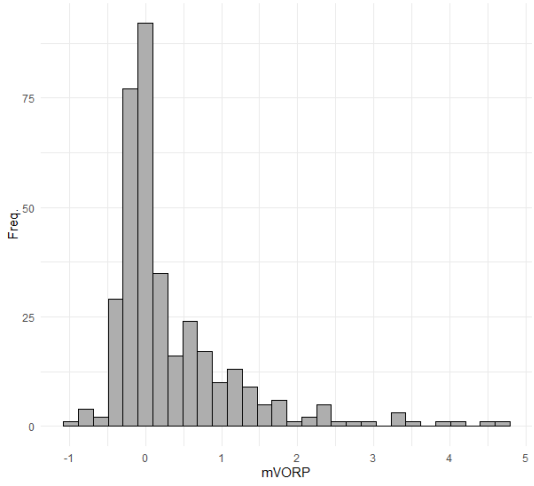


Figura 4: Histograma do VORP Médio

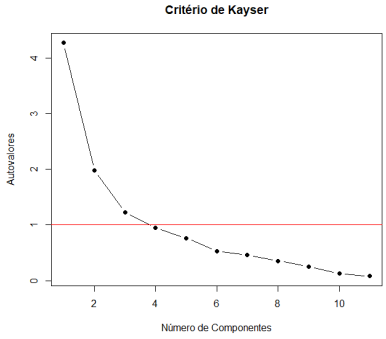
Nota-se através da Figura 4 que a variável possui uma distribuição assimétrica a esquerda, onde grande parte dos jogadores se encontram com um VORP próximo de 0, ou seja, a grande maioria dos jogadores selecionados nos *drafts* de 2010 a 2017 se encontram no nível de *replacement player*. Tal resultado já se mostra esperado, visto que em uma liga profissional de qualquer esporte há mais jogadores medianos do que estrelas. Nossa análise se complementa com a Tabela 7, a qual nos mostra a dificuldade de um atleta na NBA

se manter em alto nível ao longo de toda extensão de sua carreira, dado que o VORP máximo se encontra no valor próximo de +5.

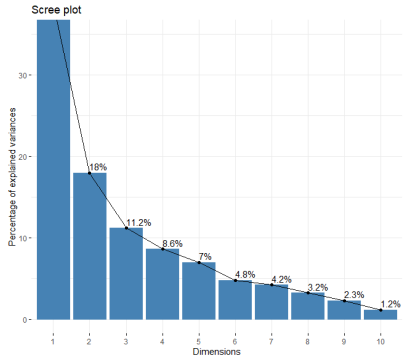
Tabela 7: Estatísticas descritivas do VORP Médio

Estatísticas	VORP Médio
Mediana	0.0125
Desvio Padrão	0.8309
Mínimo	-0.9
Média	0.3435
Máximo	4.7667

As estatísticas descritivas apresentadas até agora e resumidas através da Tabela 7 se referem a todos os 358 jogadores com passagem pelo basquetebol colegial que jogaram pelo menos uma partida profissional na NBA (de um total de 480 previamente mencionados). Em relação ao VORP, ainda não está sendo considerada uma faixa de temporadas, ou seja, a média apresentada pela Tabela 7 se refere a todos os anos da carreira de cada jogador, portanto o cálculo inclui atletas com mais e outros com menos experiência. A seguir, passa-se a definição dos componentes principais com base nas variáveis listadas na Tabela 4.



(a) *Critério de Kayser*



(b) *Scree Plot*

Figura 5: Critérios utilizados na escolha dos componentes

Analisando através da Figura 5, ambos os critérios acabam por se complementar. Dado que se tem aproximadamente 5 autovalores bem próximos do limiar 1 e que o ponto de suavização do *scree plot* ocorre na quinta dimensão, optou-se por fixar em 5 componentes. Ao utilizar 5 componentes, o percentual de variância acumulada é igual a 83.62%.

5.1 Comparação entre os modelos

Ao todo, 4 modelos foram ajustados: Regressão Linear com variáveis originais e com os componentes principais, Redes Neurais com variáveis originais e com os componentes principais. No caso das Redes Neurais, diferentes configurações foram testadas e apenas o modelo com os menores erros (MAE, MSE, RMSE) é apresentado. Foi utilizado 90% do banco para treino e o restante para teste, tendo sido realizada uma amostra estratificada por posição dos jogadores. Após a validação cruzada o melhor resultado para redes neurais utilizando PCA foi a combinação de 3 camadas ocultas, sendo a primeira com 4 neurônios, a segunda também com 4 neurônios e a terceira com 18 neurônios, ou seja, (4, 4, 18), enquanto para as redes com variáveis originais foram 3 camadas ocultas (4,9,11), onde ambas foram treinadas utilizando 100.000 épocas ou até atingir a convergência do algoritmo, aplicando função de ativação logística em todas as camadas ocultas e linear na camada de saída. Na Tabela 8 tem-se a comparação das métricas entre os melhores modelos (modelos com menores métricas de erros de teste):

Tabela 8: Métricas de Desempenho - Primeiro Ano

Modelo	RMSE	MSE	MAE
Regressão	0.8931	0.7976	0.6073
Regressão (PCA)	0.8866	0.7861	0.6003
Redes Neurais	0.8896	0.7913	0.5895
Redes Neurais (PCA)	**0.8830**	**0.7797**	**0.5850**

Através de Tabela 8 nota-se que os modelos com menores erros foram aqueles que utilizaram os *scores* do PCA, tanto no caso das redes neurais como na regressão.

Tendo escolhido o modelo com menores métricas, o próximo passo é comparar o quão melhor (ou pior) seria o uso do modelo proposto (Redes Neurais com PCA) em relação ao *draft* que foi efetivamente utilizado. Para isso ordena-se os jogadores do *draft* de 2017 de acordo com seu VORP e campara-se esta medida com o ordenamento do *draft* atual e do modelo.

Tabela 9: Exemplo de comparação

Ordenação Modelo	Ordenação *draft*	Ordenação VORP	Jogador
5	15	1	Jogador 1
10	3	2	Jogador 2
3	7	3	Jogador 3

De acordo com o exemplo da Tabela 9, o erro calculado para a ordenação do modelo seria:

$$\begin{aligned} MSE &= \frac{((5-1)^2 + (10-2)^2 + (3-3)^2)}{3} = 26.7 \\ MAE &= \frac{(|(5-1)| + |(10-2)| + |(3-3)|)}{3} = 4 \end{aligned} \tag{8}$$

Enquanto para o *draft* atual seria:

$$\begin{aligned} MSE &= \frac{((15-1)^2 + (3-2)^2 + (7-3)^2)}{3} = 71 \\ MAE &= \frac{(|(15-1)| + |(3-2)| + |(7-3)|)}{3} = 6.3 \end{aligned} \tag{9}$$

De acordo com os valores presentes nas equações 8 e 9 a conclusão é que o modelo do exemplo se sai melhor que o *draft* conduzido à época da NBA (apresenta erros menores). Esta metodologia é então expandida para todo o conjunto de teste, com os resultados apresentados na Tabela 10. Cabe destacar que esta análise leva em consideração apenas o primeiro ano após o *draft* para o cálculo do VORP.

Tabela 10: Comparação com o *draft*

Modelo	RMSE	MSE	MAE
Modelo RN (PCA)	**13.51**	**182.56**	**9.96**
draft	14.73	217.24	11.24

Através da Tabela 10 nota-se que em todas as métricas o modelo se sai melhor quando comparado ao *draft* real da NBA. Portanto, mesmo que o modelo enfrente dificuldades em entender a relação entre covariáveis e variável resposta, isto não o impede de melhorar as predições do *draft*.

6 Conclusão

O presente trabalho tinha como finalidade explorar e analisar o poder preditivo das estatísticas do basquetebol colegial para o sucesso na NBA, utilizando de técnicas de *machine learning*. Quanto a determinação de uma variável resposta, utilizou-se como balizador de sucesso o VORP de um atleta, métrica que busca sumarizar todas as contribuições em quadra através de um número, possibilitando, desta forma, a comparação do sucesso entre jogadores. Em relação a avaliação dos diferentes métodos: redes neurais artificiais e regressão linear múltipla, as melhores métricas de avaliação se dá em relação as redes, embora o desempenho de ambos seja semelhante.

Quanto as variáveis que influenciam no aumento ou diminuição do VORP, ao utilizar a técnica de componentes principais que jogadores com características de proteger o garrafão, controlar o jogo ou selecionar arremessos de forma a ter alto aproveitamento tendem a possuir um VORP mais alto, tanto no seu primeiro ano na NBA quanto ao longo do seu contrato de novato, enquanto as características de alto volume de jogo assim como arremessadores não tem influência significativa nesta métrica de sucesso. Os resultados corroboram com [8] pois, embora um alto volume de jogo esteja relacionado com posições mais altas no *draft*, tal variável não se relaciona com a produção e desempenho a nível profissional. Em relação as variáveis que influenciam no aumento do VORP, pode-se avaliar que por mais que elas não aumentem o VORP de um atleta, são características mais fáceis de serem mantidas na transição do *college* para o basquetebol profissional e, como o VORP se baseia em dados do *box score*, consequentemente estes atletas devem ter um desempenho mais alto.

Com relação a variável resposta, pode-se afirmar que as estatísticas do basquetebol colegial se mostram mais úteis para prever o VORP no primeiro ano de um atleta na liga. Não foi avaliado o VORP médio ou mediano de uma longa carreira. O VORP de um maior período de tempo influencia em questões de adaptação de um atleta, podendo sofrer com rotações de equipe, lesões ou até mesmo o efeito contrário, jogadores com pior *ranking* se tornarem peças fundamentais em equipes vencedoras, sendo este tipo de análise indicado para trabalhos futuros. De qualquer forma, prever uma nova estrela na NBA foge do alcance dos modelos apresentados. Dentro do que foi proposto no trabalho, onde o foco se deu apenas em estatísticas do *box score*, concluímos que apesar do modelo não ser capaz de diferenciar um LeBron James ou Michael Jordan dos demais, o mesmo aparece como uma boa alternativa para estimar o potencial de um atleta ao entrar na liga profissional, dado a comparação direta com o *draft* baseado no primeiro ano de VORP de um jogador.

Vale mencionar que este trabalho se mostra como um estudo pontual considerando uma faixa relativamente atual de jogadores selecionados no *draft*. Dado que a NBA possui mudanças em seu estilo de jogo assim como a adaptação e evolução dos atletas de acordo com estas mudanças, a realização deste estudo em outras épocas do *draft* poderia obter diferentes resultados dos obtidos neste trabalho em questão. Embora o foco seja no sucesso de jogadores do basquetebol profissional, este estudo, assim como os métodos do mesmo, poderiam ser estendidos para outros esportes que considerassem a universidade como principal meio de inserção de novos talentos para a liga profissional, como é o caso do futebol americano.

Referências

[1] NBA.com. NBA Draft Lottery: Schedule, odds and how it works. NBA.com, 2021, 2021. URL https://www.nba.com/nba-draft-lottery-explainer. Acesso em: 28 set. 2021.
[2] CBS. 2021 NBA Draft Prospect Rankings. CBS Sports, 2021, 2021. https://www.nba.com/nba-draft-lottery-explainer. Acesso em: 28 set. 2021.
[3] Ben Alamar. Rockets, spurs lead the way in nba draft analytics. ESPN, 2021, 2021. https://www.espn.com/nba/story/id/23762871/rockets-spurs-celtics-most-analytical-draft-teams-nba. Acesso em: 28 set. 2021.
[4] Anthony Bennett. Basketball recruiting – player profiles – espn. ESPN, 2013, 2013. http://insider.espn.com/college-sports/basketball/recruiting/player/id/103629/anthony-bennett. Acesso em: 28 set. 2021.
[5] Ernesto Soliven. Giannis antetokounmpo makes the case for the hall of fame in just one year. Basketball Network, 2021, 2021. https://www.basketballnetwork.net/giannis-antetokounmpo-makes-the-case-for-the-hall-of-fame-in-just-one-year/. Acesso em: 28 set. 2021.
[6] Scott Barry Kaufman. What predicts nba success? Scientific American, 2014, 2014. https://blogs.scientificamerican.com/beautiful-minds/what-predicts-nba-success/. Acesso em: 28 set. 2021.
[7] Adarsh Kannan, Brian Kolovich, Brandon Lawrence, and Sohail Rafiqi. Predicting national basketball association success: A machine learning approach. *SMU Data Science Review*, 1(3):7, 2018.
[8] David J Berri, Stacey L Brook, and Aju J Fenn. From college to the pros: Predicting the nba amateur player draft. *Journal of Productivity Analysis*, 35(1):25–35, 2011.
[9] Daniel Myers. About box plus/minus (bpm). Basketball-Reference, 2020, 2020. https://www.basketball-reference.com/about/bpm2.html. Acesso em: 15 mar. 2022.
[10] Basketball-Reference. Basketball-Reference. https://www.basketball-reference.com/, 2024.
[11] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.
[12] Christopher M Bishop. Pattern recognition. *Machine learning*, 128(9), 2006.
[13] Frank J Floyd and Keith F Widaman. Factor analysis in the development and refinement of clinical assessment instruments. *Psychological assessment*, 7(3):286, 1995.
[14] A. Dmitrienko, C. Chuang-Stein, and R.B. D'Agostino. *Pharmaceutical Statistics Using SAS: A Practical Guide*. SAS Institute, 2007. ISBN 9781629590301. URL https://books.google.com.kw/books?id=Gym2BQAAQBAJ.
[15] Simon Haykin. *Redes neurais: princípios e prática*. Bookman Editora, 2001.