# Utilização de algoritmos de inteligência artificial na predição de partidas de basquetebol

**Marcos Vinicius Fernandes Vital**[1]

**Resumo:** O basquetebol e um dos esportes mais assistidos é praticado no mundo com isso surgiu diversos interesses na predição de usas partidas. Por isso os sistemas computacionais que utiliza a predição passaram a ser usado, para auxiliar na formação de novas estratégias para aqueles que necessitam de dados mais preciso e confiáveis. Este projeto teve por objetivo utilizar algoritmos de inteligência artificial para a predição de partidas de basquete. Atualmente existe diverso algoritmos de predição para com diversas características, com isso nesse projeto ao se utilizar de diversos algoritmos poderá se saber quais os são os mais adequados para a predição de dados nas partidas de basquetebol.
**Palavras-chave:** Basquetebol. Predição. inteligencia artificial.

## Use of Artificial Intelligence Algorithms in Predicting Basketball Games

**Abstract:** Basketball and one of the most watched sports is practiced in the world with this has arisen several interests in the prediction of using matches. Therefore the computational systems that use the prediction have been used to assist in the formation of new strategies for those who need more accurate and reliable data. This project aimed to use artificial intelligence algorithms to predict basketball games. Currently there are several prediction algorithms for several characteristics, so in this project when using several algorithms you can know which are the most suitable for predicting data in basketball games.
**Keywords:** Basketball. Prediction. Artificial intelligence.

[1]Estudante de Ciência da Computação, IFTM, Campus Ituiutaba, vinicius.pm901@hotmail.com

# 1 INTRODUÇÃO

O basquete é um dos esportes mais populares do mundo,e também, uma esporte de equipe, em que o objetivo é atirar uma bola através de uma cesta posicionada horizontalmente, para marcar pontos, com um conjunto de regras. Normalmente, há duas equipes de cinco jogadores jogando em uma quadra retangular marcado, cada lado com uma cesta. Com sua popularidade surgiu o interesses na predição de suas partidas.

*National Basketball Associativo* (NBA) desde a sua origem tem mais de 60 anos. Durante esta organização, há 30 equipes formadas e divididas em Conferência Leste e Conferência Oeste. Para a temporada regular terá 82 jogos para cada equipe e pós temporada usando um esquema de melhor de sete séries. Portanto, uma estimativa que, haverá pelo menos cerca de 2.300 jogos gerados. Com uma massa de dados e gerada depois de cada jogo da NBA, esses dados existentes nos permitem descobrir dados valioso.

Por isso os sistemas computacionais que utiliza a predição passaram a ser usado, para auxiliar na formação de novas estratégias para aqueles que necessitam de dados mais preciso e confiáveis. A matemática tornaram-se uma parte importante do esporte e muito esforço é dedicado a prever os resultados de eventos esportivo.

Já tendo sido usado em vários processos de tomada de decisão, a predição pode fornecer para os treinadores uma visão sobre o desempenho de suas equipes durante um jogo. Assim podendo simular como uma partida é estudar cenários que podem surgir em diferentes circunstâncias na quadra. Assim ajudará os treinadores a entender como uma equipe pode aumentar suas chances de ganhar, como as habilidades em jogos individuais afetam o desempenho da equipe e qual desempenho pode ser esperado usando diferentes abordagens.

# 2 REFERENCIAL TEÓRICO

A análise computacional é uma maneira objetiva de registrar o desempenho, de modo que os eventos críticos nesse desempenho podem ser quantificados de maneira consistente e confiável. Essa análise permite que o treinador e o gerente avaliem objetivamente o desempenho competitivo e, portanto, melhorem-no (FRANKS, 2004).

A precisão e a velocidade das previsões dependerão da seleção manual ou automática adequada dos recursos mais significativos e altamente correlacionados. Kahn avaliou características primárias e empregou o método sugerido por (PURUCKER, 1996) para selecionar cinco carac-

terísticas finais para predição (KAHN, 2003).

As métricas da Associação de Pesquisadores de Basquete Profissional (ABPR) são semelhantes às da sabermetrics, que é uma das primeiras métricas para avaliar o desempenho dos jogadores de beisebol, mas as métricas da ABPR tentam visualizar as estatísticas em termos de desempenho de equipe e não de desempenho individual. Existem muitos fatores incertos para influenciar o resultado, no entanto, a mineração de dados ainda tem seu próprio valor na previsão do resultado na previsão de resultados de jogos de basquete.(SCHUMAKER; SOLIEMAN; CHEN, 2010)

Bernard, Earl e W (2009) fizeram uma pesquisa sobre a previsão de jogos da NBA usando redes neurais. Autores exploraram subconjuntos obtidos a partir de relações sinal-ruído e opiniões de especialistas para identificar um subconjunto de recursos de entrada para as redes neurais. Os resultados obtidos a partir dessas redes foram comparados com as previsões feitas por vários especialistas no campo do basquete. Após o experimento, o projeto teve 70,33% de precisão.

Embora o treinamento de um Máquina de vetores de suporte (MVS) leve mais tempo comparado a outros métodos, acredita-se que o algoritmo tenha alta precisão devido à sua alta capacidade de construir limites de decisão complexos e não-lineares. Também é menos propenso a *overfitting* (HAN; KAMBER; PEI, 2017).

Cao empregou um MVS, um classificador logístico simples uma combinação de algoritmos cujo núcleo é a regressão logística e usa o *LogitBoost* como uma função de regressão simples (LANDWEHR; HALL; FRANK, 2005), e uma rede neural multicamada para prever resultados de basquete.

Witten et al. (2017), no momento no qual deseja-se estimar o valor de uma variável numérica e os atributos do conjunto de dados também são numéricos, a escolha pela técnica de regressão linear é natural, a mesma vem sendo utilizada por décadas na aplicação de problemas estatísticos, de modo que mesmo quando o conjunto de dados não apresenta uma dependência linear a aplicação do algoritmo serve como um ponto de partida para a utilização de outros algoritmos mais complexos.

## 2.1 OBJETIVO GERAL

O objetivo deste trabalho é a análise e predição de partidas de basquete utilizando dados das partidas e dos jogadores.

## 2.2 OBJETIVOS ESPECÍFICOS

- comparar e demonstrar a eficácia para os classificadores utilizados no estado da arte de predição de partidas de basquete.
- comparar e demonstrar a eficácia das bases de dados existentes no estado da arte na predição de partidas de basquete.
- comparar e demonstrar a eficácia dos métodos seletores de características utilizados no estado da arte de predição de partidas de basquete.

# 3 DESENVOLVIMENTO

## 3.1 MATERIAIS

Os materiais usados no trabalho foram duas bases de dados ambas da NBA Advanced Stats[1] sendo uma da *season* de 2014 a 2018 com 9.840 jogos com os dados armazenados em um arquivo csv, a outra indo da *season* de 2007 a 2019 com 30.000 jogos com os dados armazenados em um banco de dados e sendo acessado através da nba-api PyPI.

A ferramenta utilizada para o desenvolvimento foi o *JupyterLab*, linguagem de programação python e o uso das bibliotecas pandas, numpy, *sklearn*, *seaborn*, matplotlib.

A NBA advances stats e um site patrocinado pela SAP com o proposito e de manter um registros de toda liga da NBA e facilitar e acesso a essa informações pelas equipes e organizações. A nba-api PyPI e uma API para acesso a www.nba.com, o principal objetivo e mapear e analisar o maior numero posivel de jogos.

O *jupyterlab* é um ambiente de desenvolvimento interativo baseado na *web* para *notebooks*. O *jupyterlab* é facil de configurar e organizar, a interface do usuário para suporta uma ampla variedade de fluxos de trabalho em ciência de dados, computação científica e aprendizado de máquina.O *jupyterlab* é extensível e modular e fácil de adicionar os *plug-ins*, que adicionam novos componentes e se integram aos já existentes(JUPYTER, 2019).

Python é uma linguagem de programação criada por Guido van Rossum em 1991. Os objetivos do projeto da linguagem eram produtividade e legibilidade, é uma linguagem de alto nível, multi paradigma, suporta o paradigma orientado a objetos, imperativo, funcional e procedural. Possui tipagem dinâmica e uma de suas principais características é permitir a fácil

[1] National Basketball Associativo

leitura do código e exigir poucas linhas de código se comparado ao mesmo programa em outras linguagens(TECHNOLOGY, 2019).

O pandas é uma biblioteca de código aberto, licenciada por BSD [2], que fornece estruturas de dados de alto desempenho e fáceis de usar e ferramentas de análise de dados para a linguagem de programação *python*. O pandas ajuda a preencher essa lacuna, permitindo que você execute todo o fluxo de trabalho de análise de dados no *python* sem precisar mudar para uma linguagem(PANDAS, 2019).

O numPy é uma biblioteca *python* que é usada para realizar cálculos em *arrays* multidimensionais. Fornecendo um grande conjunto de funções e operações que ajudam os programadores a executar facilmente cálculos numéricos.(SANTIAGO, 2019)

O *scikit learn* é uma biblioteca *python* que é usada para aprendizado de máquina. Ela possui uma variedade de algoritmos incluindo vários algoritmos de classificação, regressão e agrupamento incluindo máquinas de vetores de suporte, florestas aleatórias, *gradient boosting*, *k-means*(VAROQUAUX, 2013).

O matplotlib é uma biblioteca de plotagem 2D do python, é uma biblioteca que tenta facilitar e facilitar a gerar gráficos, histogramas, espectros de potência, gráficos de barras, gráficos de erros, gráficos de dispersão etc(MATPLOTLIB, 2019).

O seaborn é uma biblioteca de visualização de dados Python baseada no matplotlib . Ele fornece uma interface de alto nível para desenhar gráficos estatísticos atraentes e informativos.(SEABORN, 2019)

## 3.2 METODOLOGIA

Os algoritmos usados para as predições são os de regressão linear, regressão logística, k-NN[3], arvore de decisão, floresta aleatória, máquinas de vetores de suporte.

O algoritmo de regressão linear responsável por modelar uma associação entre uma ou mais variáveis de saída e entrada. O processo de regressão pode ser dividido em duas categorias, as paramétricas, no qual o relacionamento entre as variáveis é conhecido, e não paramétricas onde não existe conhecimento preexistente entre as variáveis. As técnicas de regressão linear procuram a relação entre duas variáveis por meio de uma equação de uma linha reta(BOGONI, 2019).

[2] Berkeley Software Distribution
[3] k-nearest neighbors

A regressão logística é uma técnica utilizada para a estimação de uma variável de natureza binária, estimando o valor em 0 ou 1, sendo que as variáveis independentes podem ser de natureza categórica ou não. Igualmente como na regressão linear é necessário aplicar pesos onde ajustam-se aos dados de treinamento do algoritmo, porém a regressão logística não procura a melhor reta que se ajuste aos dados, mas sim a melhor curva. A regressão logística calcula uma razão de probabilidade da variável alvo, que posteriormente é convertida em uma variável de base logarítmica, permitindo assim a classificação com base na aproximação de um dos valores(WITTEN, 2011).

O algoritmo k-NN é um método não paramétrico usado para classificação e regressão . Nos dois casos, a entrada consiste nos k exemplos de treinamento a saída depende se k-NN é usado para classificação ou regressão. Na classificação k-NN, a saída é uma associação de classe. Um objeto é classificado pelo voto de pluralidade de seus vizinhos, sendo o objeto atribuído à classe mais comum entre os k vizinhos mais próximos. Os vizinhos são obtidos de um conjunto de objetos para os quais a classe ou o valor da propriedade do objeto é conhecida(KAMGAR-PARSI; KANAL, 1985).

O processo de classificação em uma árvore de decisão, acontece de maneira recursiva, de modo que o nó inicial representa o conjunta de dados, em seguida deve ser avaliado se os objetos são da mesma classe, sendo esse o caso o nó é considerado um nó folha, caso contrário um atributo precisa ser usado para dividir os dados. Este processo deve ser executado recursivamente, ele pode ser descontinuado caso faltarem atributos para realizar testes de divisão ou caso todos os registros forem da mesma classe(CASTRO, 2016).

Florestas aleatórias são um grupo de árvores de decisões, nos quais juntos formam uma floresta. Estas árvores são geradas com base em um atributo aleatório que é o responsável pela divisão em cada nó da árvore. A precisão de uma floresta aleatória é determinada de acordo com a força de cada classificador da árvore, e também o nível de dependência entre eles, o melhor modo de atingir essa precisão é mantendo a força dos classificadores e não aumentar a correlação entre eles(CASTRO, 2016).

A técnica de máquinas de vetores de suporte, têm como fundamento o aprendizado em cima da estatística, o algoritmo apresenta ótima performance na utilização de dados de alta dimensionalidade. O mesmo funciona através de um conceito de hiperplano, sendo definido um limite linear neste plano para realizar a classificação, o algoritmo possuí a função de detectar o hiperplano de margem máxima, aquele com a maior margem separação entre as classes, com

o objetivo de apresentar menos erros de generalização em relação a margens menores(TAN, 2009).

Para desenviolamento e testes foi necessário a escrita dos algoritmos,para o começo foi importadas as bibliotecas, e foi carregada as bases de dados, apos a base de dados ser carregadas em um *dataframe*. foi feita a avaliação de ambas as bases e feita a escolha das características que seria usadas para a predição.

A base1[4] contendo 40 características sendo elas *team*, *game*, *date*, *home*, *opponent*, *winorloss*, *team points*, *opponent points*, *field goals*, *field goals attempted*, *X3 point shots*, *X3 point shots attempted*, *x3 point shots*, *free throws*, *free throws attempted*, *free throws*, *off rebounds*, *Total rebounds*, *assists*, *steals*, *blocks*, *turnovers*, *total fouls*, *opp field goals*, *opp field goals attempted*, *opp field goals* , *opp 3 point shots*, *opp 3 point shots attempted*, *opp free throws* , *opp off rebounds*, *opp total rebounds*, *opp assists*, *opp steals*, *opp blocks*, *opp turnovers*, *opp total fouls* depois de um analise as características que não foram relevante foram excluídas da base se dados as que foram retirada sao as *team*, *game*, *date*, *home*, *opponent*.

A base2[5] contendo 30 características sendo elas *season id*, *team id*, *team abbreviation*, *team name*, *game id*, *team out*, *match up gamedate*, *(W/L)win loss*, *minutes*, *played*, *points*, *field goals made*, *field goals attempted*, *field goal percentage*, *3 point field goals made*, *3 point field goals attempted*, *3 point field goal percentage*, *free throws made*, *free throws attempted*, *free throw percentage*, *offensive rebounds*, *defensive rebounds*, *rebounds*, *assists*, *steals*, *blocks*, *turnovers*, *personal*, *fouls*, *plus minus* depois de um analise as características que não foram relevante foram excluídas da base se dados as que foram retirada sao as *season id*, *team id*, *team abbreviation*, *team name*, *game id*, *team out*, *match up*, *gamedate*.

Apos a retirada das características que não serão usadas foi feito um processamento nos dados, transformando as colunas *(W/L)win loss* e *winorloss* que continha dos dados de vitoria como ”L”e derrota como ”L”, as linha contendo ”W”foi convertida para ”1”as com ”L”para ”0”. É verificando se a dados faltantes na base de dados, caso houve-se dados faltante as lacunas foi preenchida com a media dos dados da receptiva coluna.

Com as base de dados preparada foi feito uma divisão na base, sendo dividida em duas parte uma para teste e outra para treino, com a base de teste contendo 30 % dos dados e a base treino contendo os outra 70 %. com a base separa em quatro vetores sendo os vetores x_treino, x_teste, y_treino, y_teste.

[4]base de dados de 2014 a 2018 com 9.840

[5]base de dados de 2007 a 2019 com 30.000 jogos

Logo após a divisão dos vetores foi feita a instanciação do algoritmos que sera utilizado, foi chamada a função de treino do algoritmo. Em seguida ao treino da base ser realizado foi feita a chamada do função de predição, para mostrar os dados foi realizado um plote contendo os dados reais e o que foi previsto, também foi realizado exibição das métricas de erro do algorítimo. A seguir foi realizados a predição usando o método do *Cros Validation* para ver se haveria melhoria na predição dos dados.

# 4 CONCLUSÃO

Represao Linerar com a base2[5]

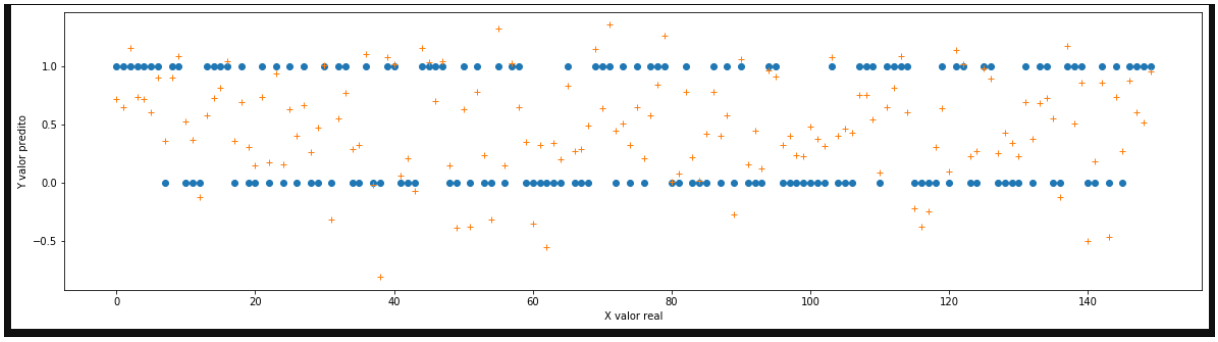


Figura 1: relação entre os dados reais e os previstos

[5]base de dados de 2007 a 2019 com 30.000 jogos

## Referências


BERNARD, L.; EARL, B.; W, B. K. Predicting nba games using neural networks. *Journal of Quantitative Analysis in Sports*, v. 5, n. 1, p. 1–17, 2009.

BOGONI, J. P. *APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS PARAPREVISÃO DE JOGOS DE BASQUETE*. [S.l.]: UNIVERSIDADE DO VALE DO TAQUARI, 2019.

CASTRO, L. D. *Introdução À Mineração De Dados: CONCEITOS BÁSICOS, ALGORITMOS E APLICAÇÕES*. SARAIVA EDITORA, 2016. ISBN 9788547200985. Disponível em: ⟨https://books.google.com.br/books?id=7HxSvgAACAAJ⟩.

DEGENNARO, K. *BWorld Robot Control Software*. 2019. ⟨https://news.sap.com/2017/08/corporate-sponsorships-reimagined-nba/⟩. [Online; accessed 19-Nov-2019].

FRANKS, I. M. *Notational Analysis of Sport*. Taylor & Francis Ltd, 2004. ISBN 0415290058. Disponível em: ⟨https://www.ebook.de/de/product/3473295/notational_analysis_of_sport.html⟩.

GRIFFITHS, M. Online video gaming: what should educational psychologists know? *Educational Psychology in Practice*, Informa UK Limited, v. 26, n. 1, p. 35–40, mar 2010.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. Elsevier LTD, Oxford, 2017. ISBN 0123814790. Disponível em: ⟨https://www.ebook.de/de/product/14641128/jiawei_han_micheline_kamber_jian_pei_data_mining_concepts_and_techniques.html⟩.

JUPYTER, P. *Jupyter*. 2019. ⟨https://jupyter.org/⟩. [Online; accessed 19-Nov-2019].

KAHN, J. Neural network prediction of nfl football games. *World Wide Web Electronic Publication*, 01 2003.

KAMGAR-PARSI, B.; KANAL, L. N. An improved branch and bound algorithm for computing k-nearest neighbors. *Pattern Recognition Letters*, Elsevier BV, v. 3, n. 1, p. 7–12, jan 1985.

KONONENKO, I. On biases in estimating multi-valued attributes. Morgan Kaufmann, p. 1034–1040, 1995.

LANDWEHR, N.; HALL, M.; FRANK, E. Logistic model trees. *Machine Learning*, v. 59, n. 1, p. 161–205, May 2005. Disponível em: ⟨https://doi.org/10.1007/s10994-005-0466-3⟩.

MATPLOTLIB. *Entendendo a biblioteca matplotlib*. 2019. ⟨https://matplotlib.org/⟩. [Online; accessed 20-Nov-2019].

PANDAS. *O projeto dos pandas*. 2019. ⟨https://pandas.pydata.org/⟩. [Online; accessed 20-Nov-2019].

PAPIĆ, V.; ROGULJ, N.; PLEŠTINA, V. Identification of sport talents using a web-oriented expert system with a fuzzy module. *Expert Systems with Applications*, Elsevier BV, v. 36, n. 5, p. 8830–8838, jul 2009.

PURUCKER, M. Neural network quarterbacking. *IEEE Potentials*, Institute of Electrical and Electronics Engineers (IEEE), v. 15, n. 3, p. 9–15, 1996.

SANTIAGO, L. *Entendendo a biblioteca NumPy*. 2019. ⟨https://medium.com/ensina-ai/entendendo-a-biblioteca-numpy-4858fde63355⟩. [Online; accessed 20-Nov-2019].

SCHUMAKER, R. P.; SOLIEMAN, O. K.; CHEN, H. Sports knowledge management and data mining. *Annual Review of Information Science and Technology*, Wiley, v. 44, n. 1, p. 115–157, 2010.

SEABORN. *Entendendo a biblioteca Seaborn*. 2019. ⟨https://seaborn.pydata.org/⟩. [Online; accessed 20-Nov-2019].

STEKLER, H.; SENDOR, D.; VERLANDER, R. Issues in sports forecasting. *International Journal of Forecasting*, Elsevier BV, v. 26, n. 3, p. 606–621, jul 2010.

TAN. *Introducão ao datamining : mineracão de dados*. Rio de Janeiro (RJ: Ciencia Moderna, 2009. ISBN 8573937610.

TECHNOLOGY, J. *Sobre o Python*. 2019. ⟨https://www.python.org/about⟩. [Online; accessed 19-Nov-2019].

VAROQUAUX, L. B. e Gilles Louppe e Mathieu Blondel e Fabian Pedregosa e Andreas Mueller e Olivier Grisel e Vlad Niculae e Peter Prettenhofer e Alexandre Gramfort e Jaques Grobler e Robert Layton e Jake VanderPlas e Arnaud Joly e Brian Holt e Ga ”e l. In: *ECML PKDD Oficina : Línguas de dados Mining e Máquina de Aprendizagem*. [S.l.: s.n.], 2013.

WITTEN, I. et al. *Data Mining*. Elsevier LTD, Oxford, 2017. ISBN 0128042915. Disponível em: ⟨https://www.ebook.de/de/product/26440029/ian_witten_eibe_frank_mark_a_hall_christopher_j_pal_data_mining.html⟩.

WITTEN, I. H. *Data mining practical machine learning tools and techniques*. Burlington, MA: Morgan Kaufmann, 2011. ISBN 0123748569.