

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

Marcos Vinicius Fernandes Vital¹;Rodrigo Grassi²

Resumo: O basquetebol é um dos esportes mais assistidos e praticado no mundo. Com um grande publico passou a ser utilizadas instrumentos para análise técnicas, que pode ser utilizada com fins estratégicos por dirigentes e até apostadores. Assim surgiram diversos interesses na predição do resultados de suas partidas. Os sistemas computacionais que utilizam a predição passaram a ser utilizados, para auxiliar na formação de novas estratégias, técnicas e táticas de jogo. Este projeto tem por objetivo utilizar algoritmos de inteligência artificial para a predição de partidas de basquete. Atualmente existem diversos 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 is one of the most watched and practiced sports in the world. With a large public began to be used tools for technical analysis, which can be used for strategic purposes by managers and even punters. Thus arose several interests in predicting the outcome of their matches. The computer systems that use the prediction started to be used to help in the formation of new strategies, techniques and tactics of the game. This project aims to use artificial intelligence algorithms to predict basketball matches. There are currently several prediction algorithms for different characteristics. With this project using several algorithms you can know which are the most suitable for predicting data in basketball matches.

Keywords: Basketball. Prediction. Artificial intelligence.

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

²Professor do IFTM, Campus Ituiutaba, MG, rodrigograssi@iftm.edu.br

1 INTRODUÇÃO

O basquete é um dos esportes mais populares do mundo, e também, um 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 um grande público passou a ser utilizadas instrumentos para análise técnicas, que pode ser utilizada com fins estratégicos por dirigentes e até apostadores. Assim surgiram diversos interesses na predição do resultados de suas partidas.

National Basketball Association (NBA) desde a sua origem tem mais de 60 anos. A organização tem 30 equipes formadas e divididas em Conferência Leste e Conferência Oeste. Uma temporada regular tem 82 jogos para cada equipe além da 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 gerada depois de cada jogo da NBA, esses dados existentes nos permitem descobrir informações valiosas.

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

Já tendo sido usado em vários modelos de predição. Eles podem fornecer para os treinadores uma visão sobre o desempenho de suas equipes durante um jogo. Também permitem simular uma partida e estudar cenários que podem surgir em diferentes circunstâncias na quadra. Esses modelos ajudam 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, melhorar o mesmo (FRANKS, 2004).

A precisão e a velocidade das previsões dependerão da seleção manual ou automática ade-

quada dos recursos mais significativos e altamente correlacionados. (KAHN, 2003) avaliou características primárias e empregou o método sugerido por (PURUCKER, 1996) para selecionar cinco características finais para predição.

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 basquete, 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. Os 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 acurácia.

Embora o treinamento de um Máquina de vetores de suporte (SVM) 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).

O SVM, 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, 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 da NBA Advanced Stats¹, sendo uma da *season* de 2014 a 2018 com 9.840 jogos com os dados armazenados em um arquivo csv, a outra é 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 NBA advances stats um site patrocinado pela SAP com o proposito de manter um registro de toda liga da NBA e facilitar acesso a essa informação pelas equipes e organizações. A nba-api PyPI é uma API para acesso a www.nba.com, o principal objetivo é mapear e analisar o maior número possível de jogos.

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

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

O *jupyterlab* é um ambiente de desenvolvimento interativo baseado na *web*. O *jupyterlab* é fácil de configurar e organizar, a interface do usuário suporta uma ampla variedade de fluxos de trabalho em ciência de dados, computação científica e aprendizado de máquina.O

¹National Basketball Associativo

jupyterlab é extensível e modular é fácil de adicionar os *plug-ins*, que se integram aos já existentes(JUPYTER, 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 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 previsões são os de regressão linear, regressão logística, k-NN³, árvore 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 linear(BOGONI, 2019).

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 natu-

²Berkeley Software Distribution

³k-nearest neighbors

reza categórica ou não. 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 conjunto 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 decisão, 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 possui 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).

A escrita dos algoritmos para realizar os testes foi iniciada, com a importação das bibliotecas

e foi carregada as bases de dados, apos a base de dados ser carregadas em um *dataframe* da biblioteca padas, foi feita a avaliação de ambas as bases é feita a escolha das características que foi usadas para a predição.

A base¹⁴ contendo 34 características sendo elas:

Características	Descrição
<i>team</i>	sigla do <i>time</i>
<i>game</i>	id do jogo
<i>date</i>	data do jogo
<i>opponent</i>	sigla do oponente
<i>winorloss</i>	vitoria e derrota
<i>team points</i>	pontos do time
<i>opponent points</i>	pontos do oponente
<i>field goals</i>	cesta marcada em qualquer arremesso ou toque que não seja lance livre
<i>field goals attempted</i>	tentativa cesta marcada em qualquer arremesso ou toque que não seja lance livre
<i>X3 point shots</i>	cesta de 3 pontos
<i>X3 point shots attempted</i>	tentativa de cesta de 3 pontos
<i>X3 point shots opp</i>	cesta de 3 pontos oponente
<i>X3 point shots attempted opp</i>	tentativa de cesta de 3 pontos oponente
<i>free throws</i>	arremessos livres
<i>free throws attempted</i>	tentativa de arremessos livres
<i>free throws opp</i>	arremessos livres oponente
<i>free throws attempted opp</i>	tentativa de arremessos livres oponente
<i>rebounds</i>	rebotes
<i>Total rebounds</i>	total de rebotes
<i>assists</i>	assistência
<i>steals</i>	roubada de bola
<i>blocks</i>	bloqueio de bola
<i>turnovers</i>	rotatividade
<i>total fouls</i>	falta total

¹⁴base de dados de 2014 a 2018 com 9.840 jogos

<i>opp field goals</i>	cesta marcada em qualquer arremesso ou toque que não seja lance livre do oponente
<i>opp field goals attempted</i>	tentativa cesta marcada em qualquer arremesso ou toque que não seja lance livre do oponente
<i>opp off rebounds</i>	rebotes dos oponentes
<i>opp total rebounds</i>	total de rebote dos oponentes
<i>opp assists</i>	assistência oponente
<i>opp steals</i>	roubada de bola oponente
<i>opp blocks</i>	bloqueio de bola oponente
<i>opp turnovers</i>	rotatividade do oponente
<i>opp total fouls</i>	total de faltas do oponente

Depois de um análise as características que não foram relevantes foram excluídas da base de dados, sendo retiradas as *team*, *game*, *date*, *home*, *opponent*.

A base²⁵ contendo 30 características sendo elas:

Características	Descrição
<i>season id</i>	id da temporada
<i>team id</i>	id time
<i>team abbreviation</i>	abreviação do time
<i>team name</i>	nome do time
<i>game id</i>	id do jogo
<i>team out</i>	time de fora
<i>match up</i>	confronto individua
<i>gamedate</i>	data do jogo
<i>win loss (W/L)</i>	vitoria e derrota
<i>minutes</i>	tempo do jogo
<i>played</i>	numero de jogadas
<i>points</i>	pontuação
<i>field goals</i>	cesta marcada em qualquer arremesso ou toque que não seja lance livre
<i>field goals attempted</i>	tentativa cesta marcada em qualquer arremesso ou toque que não seja lance livre

²⁵base de dados de 2007 a 2019 com 30.000 jogos

<i>field goal percentage</i>	porcentagem cesta marcada em qualquer arremesso ou toque que não seja lance livre
<i>3 point field goals</i>	cesta de 3 pontos marcada em qualquer arremesso ou toque que não seja lance livre
<i>3 point field goals attempted</i>	tentativa cesta de 3 pontos marcada em qualquer arremesso ou toque que não seja lance livre
<i>3 point field goal percentage</i>	porcentagem de cesta de 3 pontos marcada em qualquer arremesso ou toque que não seja lance livre
<i>free throws</i>	arremessos livres
<i>free throws attempted</i>	tentativa de arremessos livres
<i>free throw percentage</i>	porcentagem
<i>offensive rebounds</i>	rebotes ofensivos
<i>defensive rebounds</i>	rebotes defensivos
<i>rebounds</i>	rebotes
<i>assists</i>	assistência
<i>steals</i>	roubada de bola
<i>blocks</i>	bloqueio de bola
<i>turnovers</i>	rotatividade
<i>fouls</i>	faltas
<i>plus minus</i>	minutos extra

Depois de um análise as características que não foram relevantes foram excluídas da base de dados, sendo retiradas as *season id*, *team id*, *team abbreviation*, *team name*, *game id*, *team out*, *match up*, *gamedate*.

Após a retirada das características que não serão usadas foi feito um processamento nos dados, transformando as colunas (*W/L*)*win loss* e *winorloss* que continham os dados de vitória como "W" e derrota como "L", as linha contendo "W" foi convertida para "1" e as com "L" para "0". Foi verificado a ocorrência de dados faltantes na base de dados, e as lacunas foram preenchidas com a média dos dados da coluna.

Com as bases de dados preparada foi feito uma divisão, sendo em duas partes, uma para teste e outra para treino com a base de teste contendo 30 % dos dados e a base treino contendo os outros 70 %. Com a base separada em quatro vetores sendo os vetores *x_treino*, *x_teste*, *y_treino*, *y_teste*.

Logo após a divisão dos vetores foi feita a instanciação do algoritmos utilizados, e foi cha-

mada a função de treino do algoritmo. Após o treino ser realizado foi feita a chamada do função de predição. Para mostrar os dados foi realizado um plot de um gráfico contendo os dados reais e o que foi previsto, também foi realizada exibição das métricas de erro do algoritmo. A seguir foi realizada a predição usando o método do *Cross Validation* para ver se haveria melhoria nos dados.

3.3 RESULTADOS

O algoritmo de regressão linear com a base⁴ teve uma acurácia de 66.6% veja o gráfico de relação entre os dados reais e os previstos. Após aplicar *Cross Validation* essa taxa teve um aumento de 2.3%.

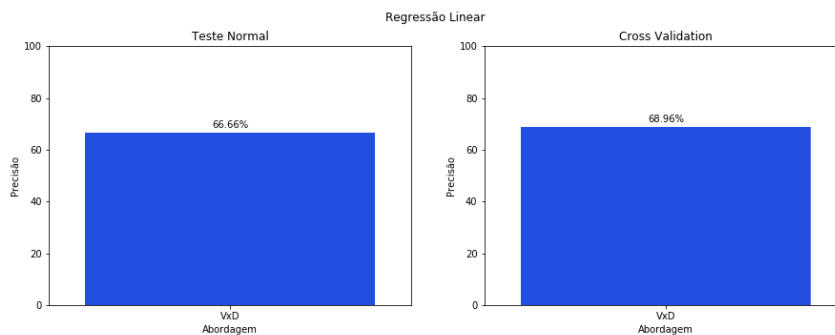


Figura 1: relação entre os dados reais e os previstos

O algoritmo de regressão linear com a base⁵ teve uma acurácia de 68% veja o gráfico de relação entre os dados reais e os previstos. Após aplicar o *Cross Validation* essa taxa teve um aumento de 3.6%.

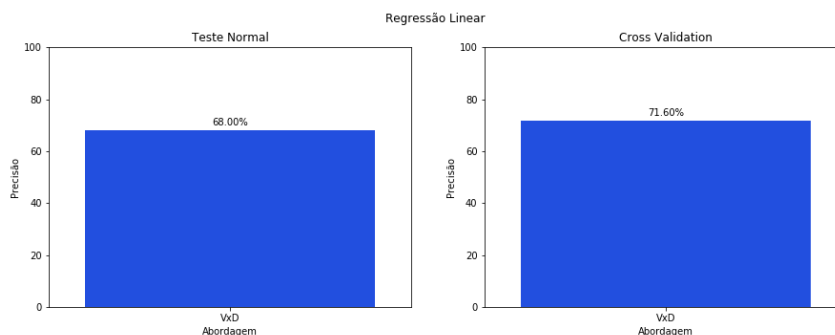


Figura 2: relação entre os dados reais e os previstos

⁴base de dados de 2014 a 2018 com 9.840 jogo

⁵base de dados de 2007 a 2019 com 30.000 jogos

O algoritmo de árvore decisão com a base1⁴ teve uma acurácia de 89.7% veja o grafico de relação entre os dados reais e os previstos. Após aplicar o *Cross Validation* essa taxa teve um aumento de 4%.

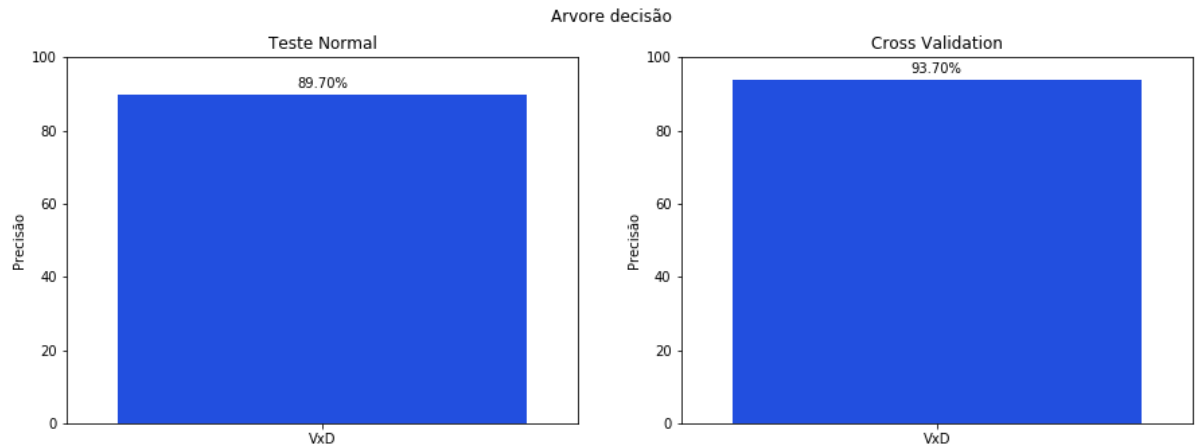


Figura 3: relação entre os dados reais e os previstos

O algoritmo de árvore decisão com a base2⁵ teve um taxa de acerto de 97% veja o grafico de relação entre os dados reais e os previstos. Após aplicar o *Cross Validation* essa taxa teve um aumento de 3%.

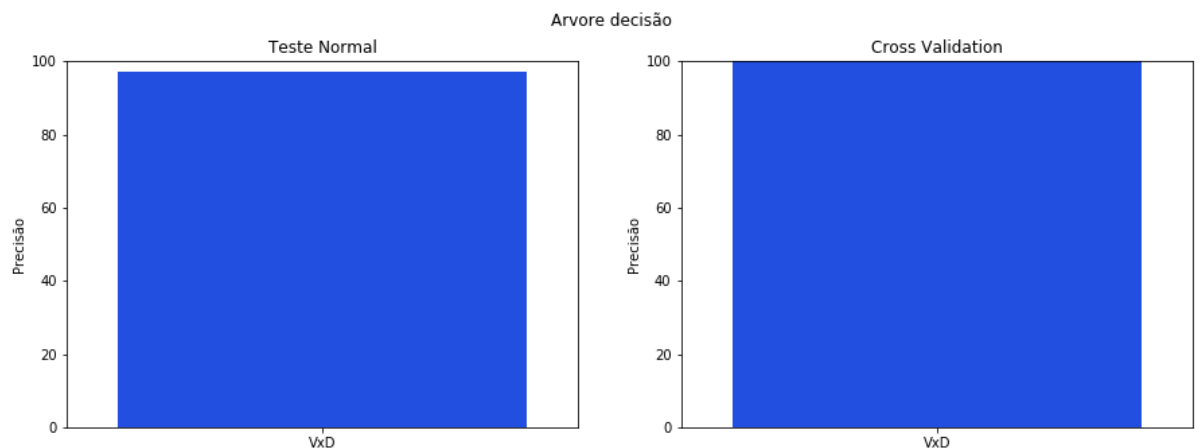


Figura 4: relação entre os dados reais e os previstos

⁴base de dados de 2014 a 2018 com 9.840 jogo

⁵base de dados de 2007 a 2019 com 30.000 jogos

O algoritmo de k-NN³ com a base1 teve uma acurácia de 84.5% veja o grafico de relação entre os dados reais e os previstos. Após aplicar o *Cross Validation* essa taxa teve um aumento de 3.3%.

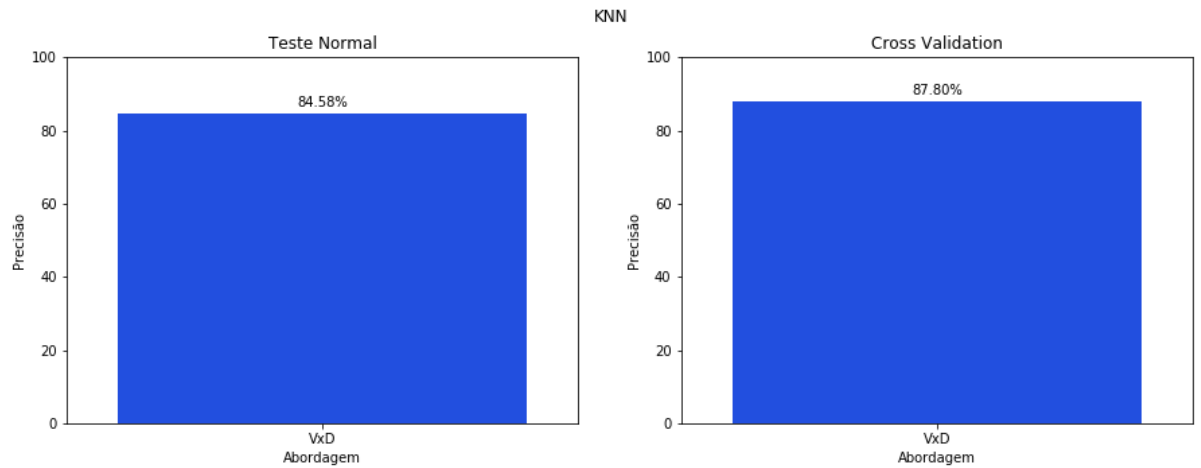


Figura 5: relação entre os dados reais e os previstos

O algoritmo de k-NN com a base2 teve uma acurácia de acerto de 87% veja o grafico de relação entre os dados reais e os previstos. Após aplicar o *Cross Validation* essa taxa teve um aumento de 3.3%.

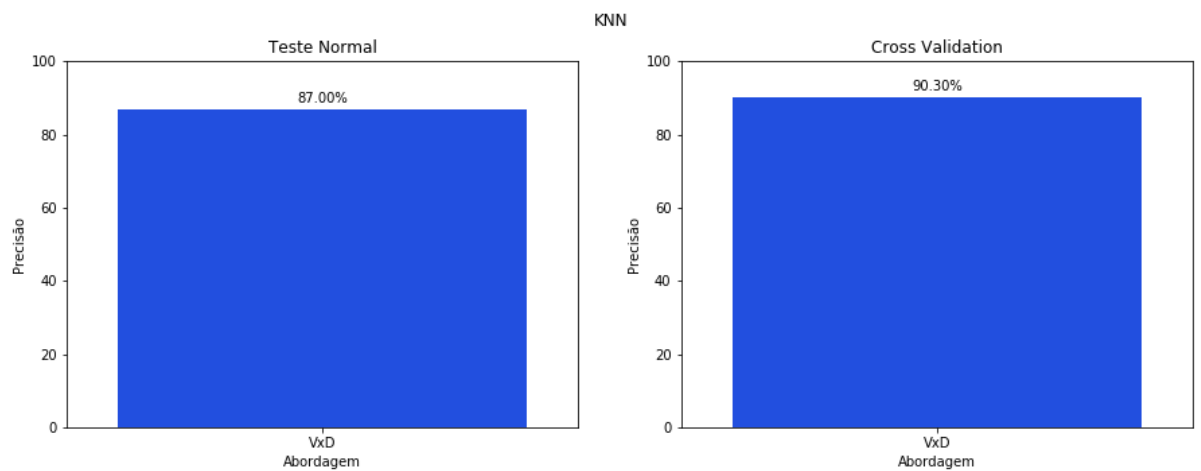


Figura 6: relação entre os dados reais e os previstos

³k-nearest neighbors

O algoritmo de floresta aleatória com a base2 inicialmente foi testado com 10 arvores chegando a um resultado de 91% e 94 % com 25 arvores de acurácia.

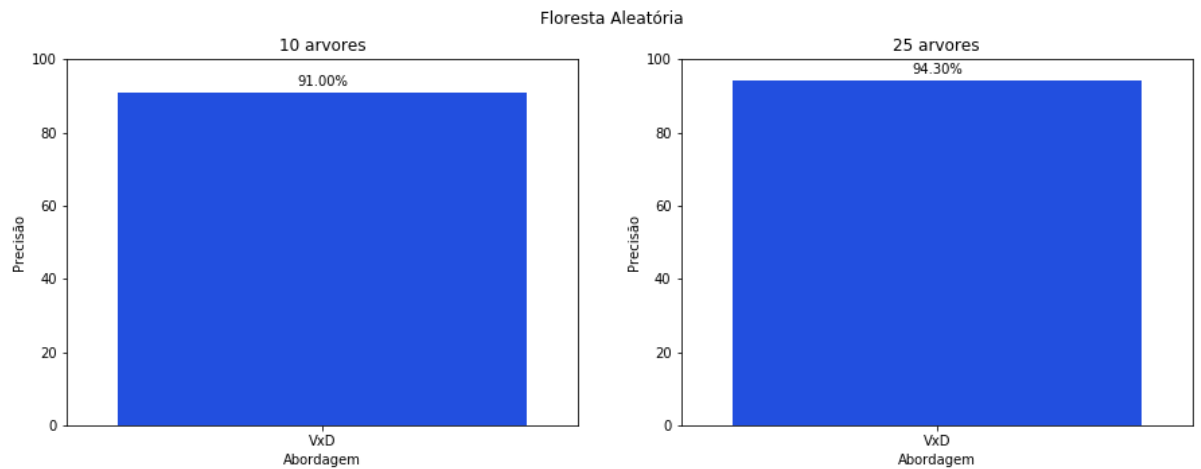


Figura 7: relação entre os dados reais e os previstos com 10 e 25 arvores

O algoritmo de floresta aleatória com a base2 com 53 arvores conseguiu atingir uma faixa de 99% e 100% de acurácia. Para ter um taxa contante de 100% de acurácia são necessária mais de 100 arvores.

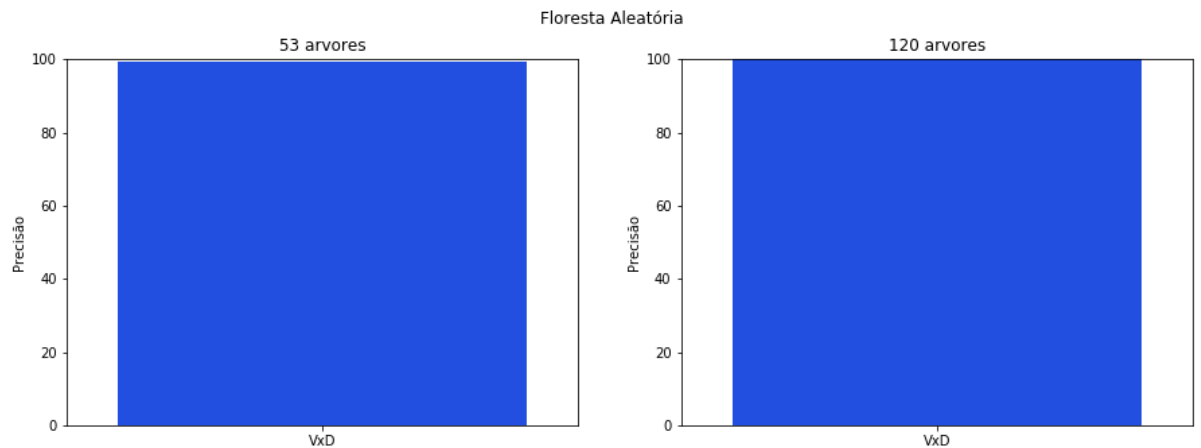


Figura 8: relação entre os dados reais e os previstos com 53 e 120 arvores

O algoritmo de regressão logística tanto com a base1⁴ é com a base2⁵ teve uma acurácia de 99% a 100% veja o gráfico de relação entre os dados reais e os previstos.

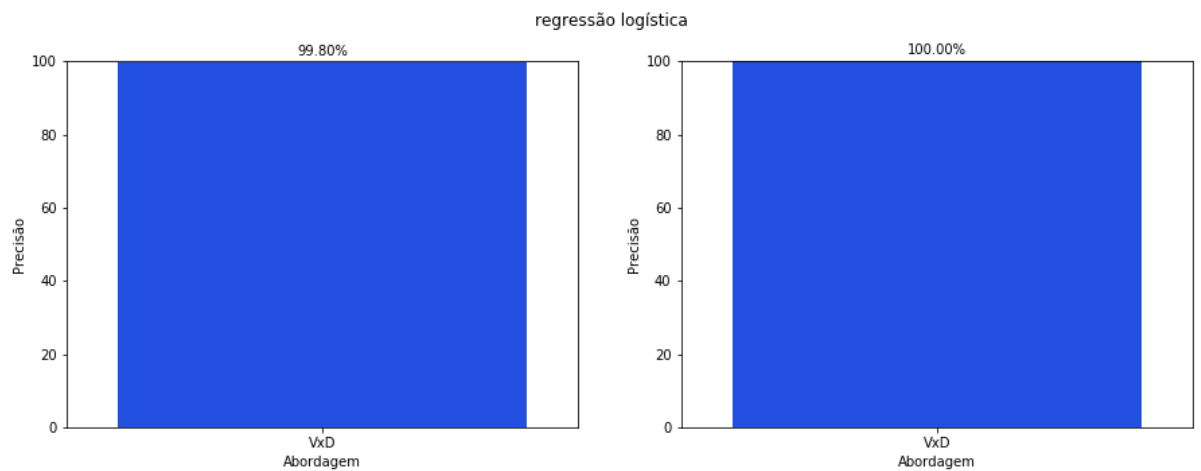


Figura 9: relação entre os dados reais e os previstos

O algoritmo de máquinas de vetores de suporte com a base1 teve uma acurácia de 80% veja o gráfico de relação entre os dados reais e os previstos. Após aplicar o *Cross Validation* essa taxa teve um aumento de 2.7%.

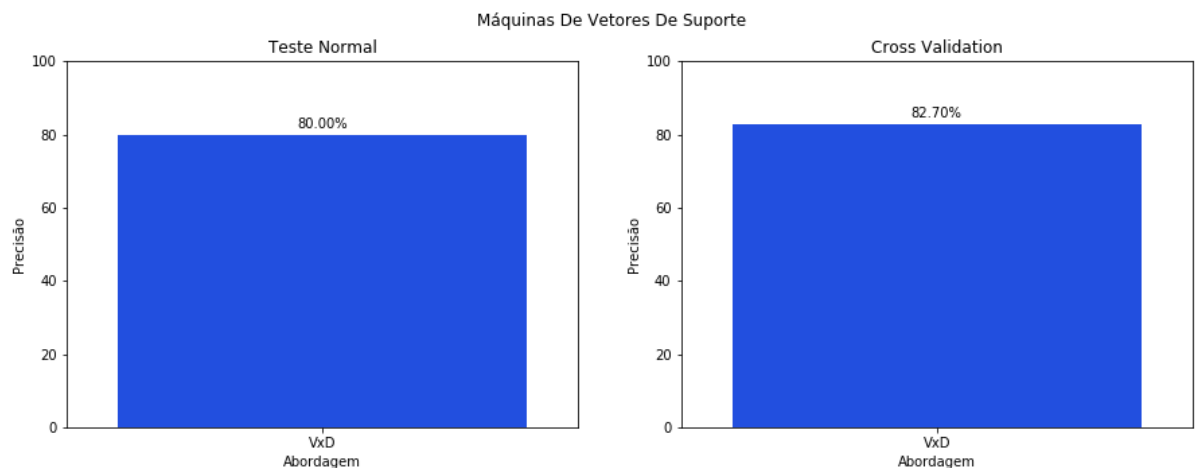


Figura 10: relação entre os dados reais e os previstos

⁴base de dados de 2014 a 2018 com 9.840 jogo

⁵base de dados de 2007 a 2019 com 30.000 jogos

O algoritmo de máquinas de vetores de suporte com a base⁵ teve uma acurácia de 87%, veja o gráfico de relação entre os dados reais e os previstos. Após aplicar o *Cross Validation* essa taxa teve um aumento de 3.7%.

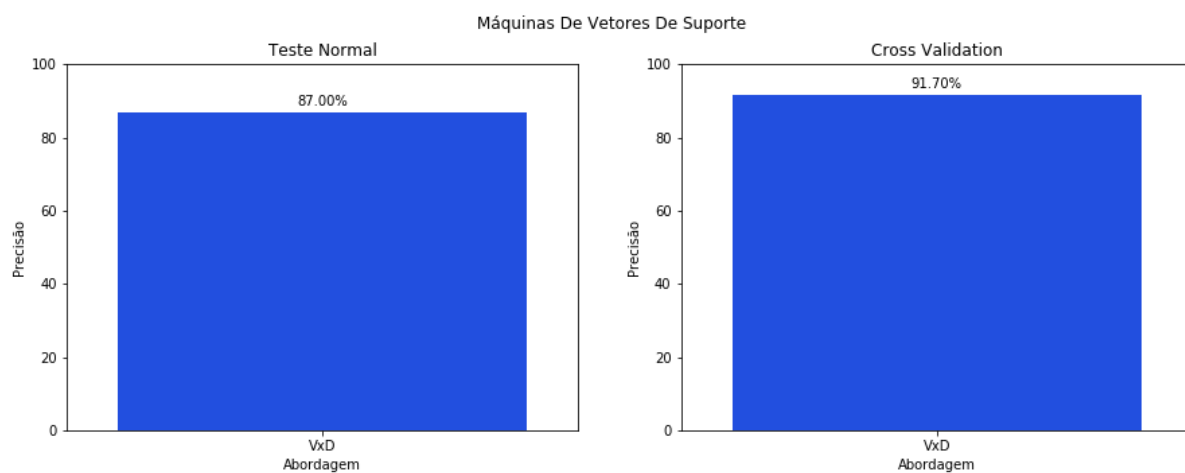


Figura 11: relação entre os dados reais e os previstos

⁵base de dados de 2007 a 2019 com 30.000 jogos

4 DISCUSSÃO E CONCLUSÃO

Conforme observado nos teste conclui-se que os algoritmos testados funciona melhor com dados de classe binaria,sendo assim dados como vitoria x derrota.

O algoritmo de regressão linear que teve o pior despenho no teste, devido a construção dele trabalhar melhor com dados lineares. O que dado que foi utilizado consistia em prever derrota ou vitoria ou seja, o valor da predição era de forma binaria. Assim a técnicas de regressão linear procuram a relação entre duas variáveis por meio de uma equação de uma linha reta.

O algoritmos arvore de decisão, como funciona na forma de um fluxograma em pra suas tomadas de decisão vai depender da quantidade e qualidade dos dados com a qual essa arvore foi treinada. Como pode ser visto nos resultado a taxa de acertos com a utilização da base1⁴, quando testado um a base2⁵ que possui mais que o dobre de dados a taxa teve um grande aumento.

O algoritmo k-NN³ trabalho com os vizinhos mais deve ser considerado como um método no qual baseia-se por instâncias, isto é, ele vai determinar a classe de um objeto desconhecido através da classe de outras instâncias.

Floresta aleatória são um conjunto de árvores de decisão trabalhando em conjunto, com um maior numero de arvore a taxa de predição também aumenta. Nos teste realizados com 10 árvores tinha uma aula porcentagem de acertos quando chegando acima de 53 arvore a taxa varia de 99% a 100%, mas quanto o maior numero de árvores o tempo do teste vai aumentando significativamente. Para ter uma taxa de constante de 100% seria necessario mas 100 árvores isso exige um grande tempo.

A regressão logística foi o algoritmo com a maior taxa de acertos, pois ele trabalha com os fatores binário de predição,sendo assim o opostos da regressão linear. Como os teste foram feitos para prever a derrota e a vitoria.Assim 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.

A máquinas de vetores trabalha definindo um limite linear logo para realizar a classificação ele separa os dados é os analisa para reconhecer padrões, assim que a uma entrada de um conjunto de dados e adicionada ele vai realizar a analise e dividir em duas classes, na qual as duas possíveis classes faz parte do classificador linear binário não probabilístico.

⁴base de dados de 2014 a 2018 com 9.840 jogo

⁵base de dados de 2007 a 2019 com 30.000 jogos

³k-nearest neighbors

Podemos concluir que os algoritmos que obtiveram a melhor performance mais vezes foram nesta ordem: regressão lógica, floresta aleatória, árvore de decisão, knn, máquina de vetores de suporte, regressão linear.

Todos os algoritmos abordados neste trabalho foram implementados em Python e encontram-se disponíveis em <https://github.com/marcos901/Projeto-TCC/tree/master/Teste%20dos%20algoritmos>

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. *Introdução ao datamining : mineração 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 1. 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.