

UTILIZAÇÃO DE ALGORITMOS DE INTELIGÊNCIA ARTIFICIAL NA PREDIÇÃO DE PARTIDAS DE BASQUETE

ALUNO : MARCOS VINICIUS FERNANDES VITAL

ORIENTADOR : DR. RODRIGO GRASSI MARTINS

INTRODUÇÃO

- Basquete
 - NBA ha 30 equipes.
 - Temporada regular tem em media 82 com 7 series.
 - 2.300 jogos gerados
- Predição
 - Estratégias
 - Desempenho
 - Técnicas
 - Tácticas de jogo

INTRODUÇÃO

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 adequada dos recursos mais significativos e altamente correlacionados (PURUCKER, 1996).

INTRODUÇÃO

REFERENCIAL TEÓRICO

- Embora o treinamento de um Máquina de vetores leve mais tempo com parâdo a outros métodos, acredita-se que o algoritmo tenha alta precisão devido a sua alta capacidade de construir limites de decisão complexos. (HAN; KAMBER; PEI, 2017)
- Bernard, Earl e W (2009) em uma pesquisa sobre a previsão de jogos da NBA usando redes neurais. Autores exploraram subconjuntos obtidos a partir de especialistas para identificar um subconjunto de recursos de entrada para as redes neurais.

INTRODUÇÃO

- OBJETIVO GERAL

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

- 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.

DESENVOLVIMENTO

MATÉRIAS

- NBA Advanced Stats
 - season de 2014 a 2018 com 9.840 jogos com os dados armazenados em um arquivo csv.
 - 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 é um site patrocinado pela SAP com o propósito de manter um registro de toda a liga da NBA e facilitar o acesso a essas informações pelas equipes e organizações.

DESENVOLVIMENTO

MATÉRIAS

- A ferramenta utilizada para o desenvolvimento foi o JupyterLab, linguagem de programação python e o uso das bibliotecas pandas, numpy, sklearn, seaborn, matplotlib.
- O jupyterlab é um ambiente de desenvolvimento interativo baseado na web para notebooks. Fácil de configurar e organizar e extensível e modular e fácil de adicionar os plug-ins, que adicionam novos componentes e se integram aos já existentes(JUPYTER, 2019).

DESENVOLVIMENTO

MATÉRIAS

- Python é uma linguagem de programação criada por Guido van Rossum em 1991. Os objetivos do projeto da linguagem eram produtividade e legibilidade, e uma linguagem de alto nível, multi paradigma, suporta o paradigma orientado a objetos, imperativo, funcional e procedural (TECHNOLOGY, 2019).
- O pandas é uma biblioteca de código aberto, licenciada por BSD, que fornece estruturas de dados de alto desempenho e fáceis de usar e ferramentas de análise de dados para a linguagem de programação python (PANDAS, 2019).

DESENVOLVIMENTO

MATÉRIAS

- 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, k-means(VAROQUAUX, 2013).

DESENVOLVIMENTO

MATÉRIAS

- O matplotlib é uma biblioteca de plotagem 2D do python, e uma biblioteca que tenta facilitar e facilitar a gerar gráficos, histogramas, espectros de potencia, 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)

DESENVOLVIMENTO

METODOLOGIA

- 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 (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 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(WITTEN, 2011).

DESENVOLVIMENTO

METODOLOGIA

- 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 (KAMGARPARSI; KANAL, 1985).
- 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(CASTRO, 2016).

DESENVOLVIMENTO

METODOLOGIA

- 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 (CASTRO, 2016).
- 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 (TAN, 2009).

DESENVOLVIMENTO

METODOLOGIA

- Para desenvolvimento e testes foi necessário a escrita dos algoritmos, para o começo foi importadas as bibliotecas, e foi carregada as bases de dados, após 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 contendo 34 características

DESENVOLVIMENTO

METODOLOGIA

Características	Descrição
<i>team</i>	sigla do time
<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

<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

DESENVOLVIMENTO

METODOLOGIA

- A base2 contendo 30 características

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 individual
<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

<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

DESENVOLVIMENTO

METODOLOGIA

- Processamento dos dados, transformando as colunas W/L e winorloss que continha dos dados de vitoria e derrotas.
- 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.
- A base foi dividida em duas parte uma para teste e outra para treino
 - 30 % dos dados base treino.
 - 70 % dos dados base teste.
 - 4 vetores x_{treino} , x_{teste} , y_{treino} , y_{teste} .

DESENVOLVIMENTO

METODOLOGIA

- Logo após a divisão dos vetores foi feita a instanciação do algoritmos que sera utilizado.
- Algoritmo foi treinado.
- Predição.
- Foi realizado um plote do gráfico contendo os dados reais e o que foi previsto.

DESENVOLVIMENTO

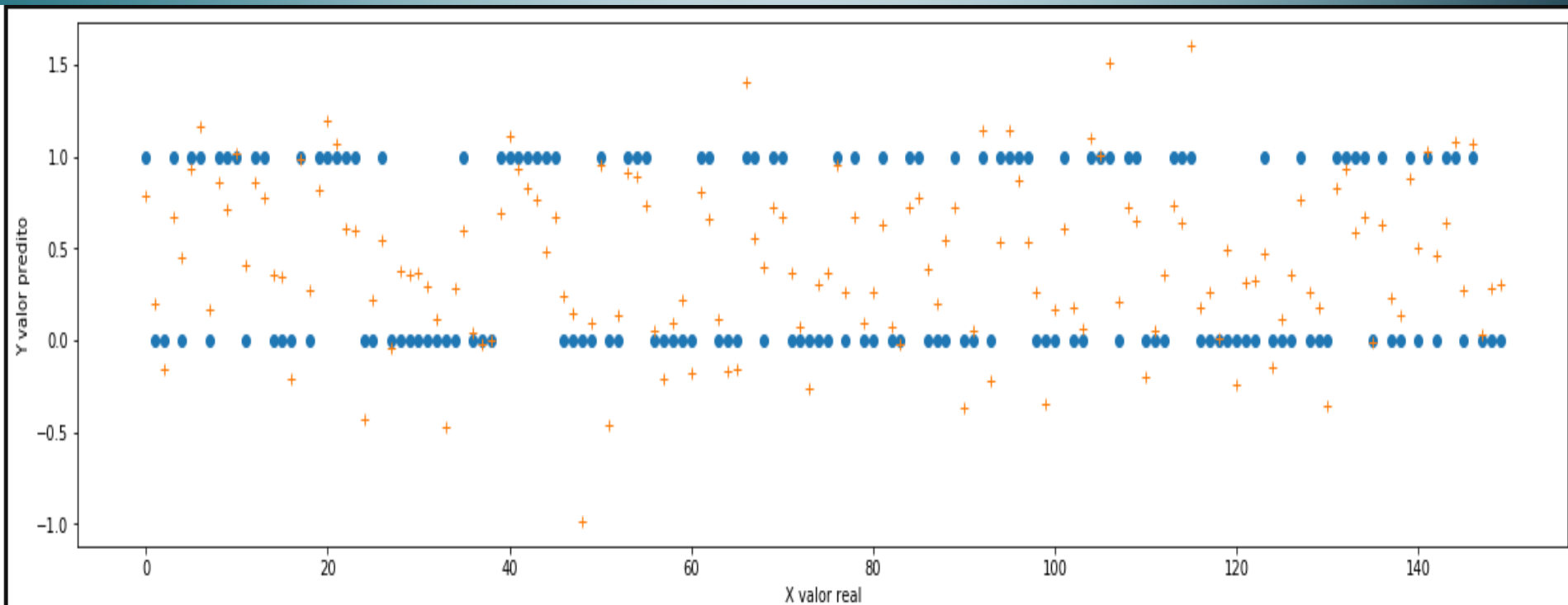
METODOLOGIA

- Foi realizado exibição das métricas de erro do algoritmo.
- A seguir foi realizados a predição usando o método do Cros Validation para ver se haveria melhoria na predição dos dados.

DESENVOLVIMENTO

RESULTADO

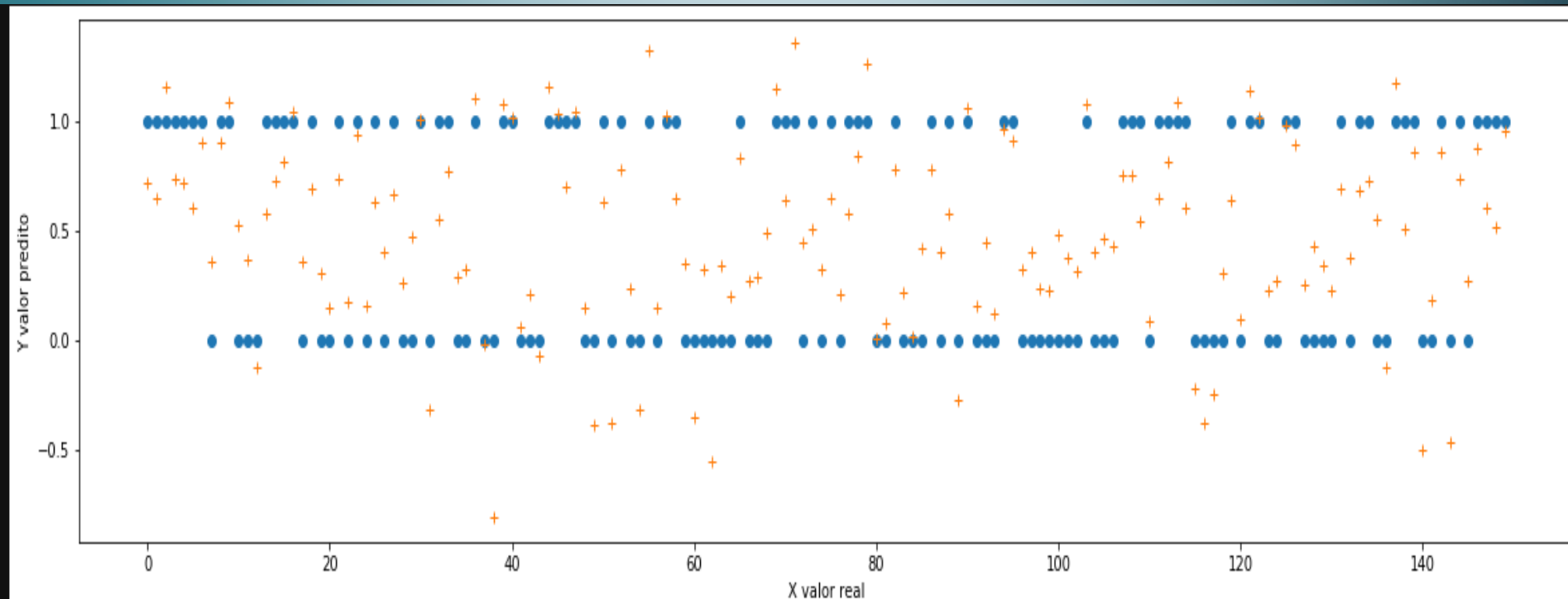
- O algoritmo de regressão linear com a base1 de 65%,
- Cros Validation o essa taxa teve um aumento de 2 %.



DESENVOLVIMENTO

RESULTADO

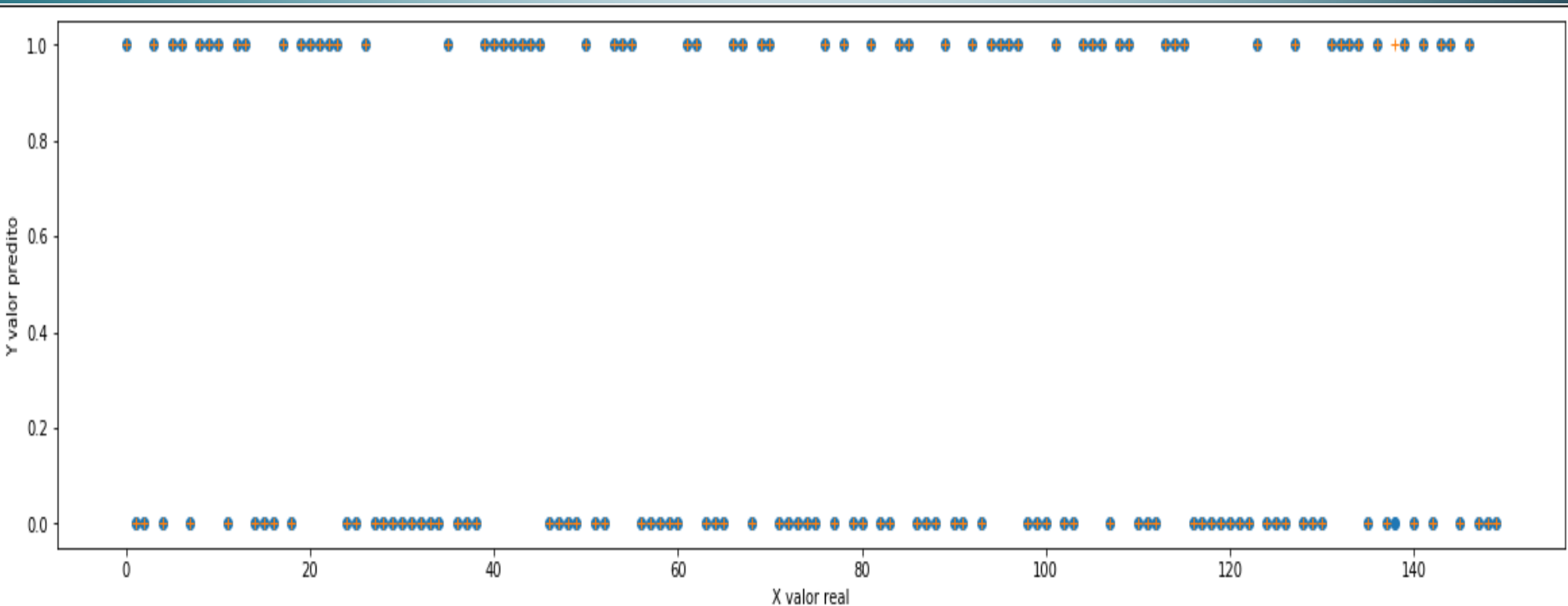
- O algoritmo de regressão linear com a base2 de 68%,
- Cros Validation o essa taxa teve um aumento de 3.6%.



DESENVOLVIMENTO

RESULTADO

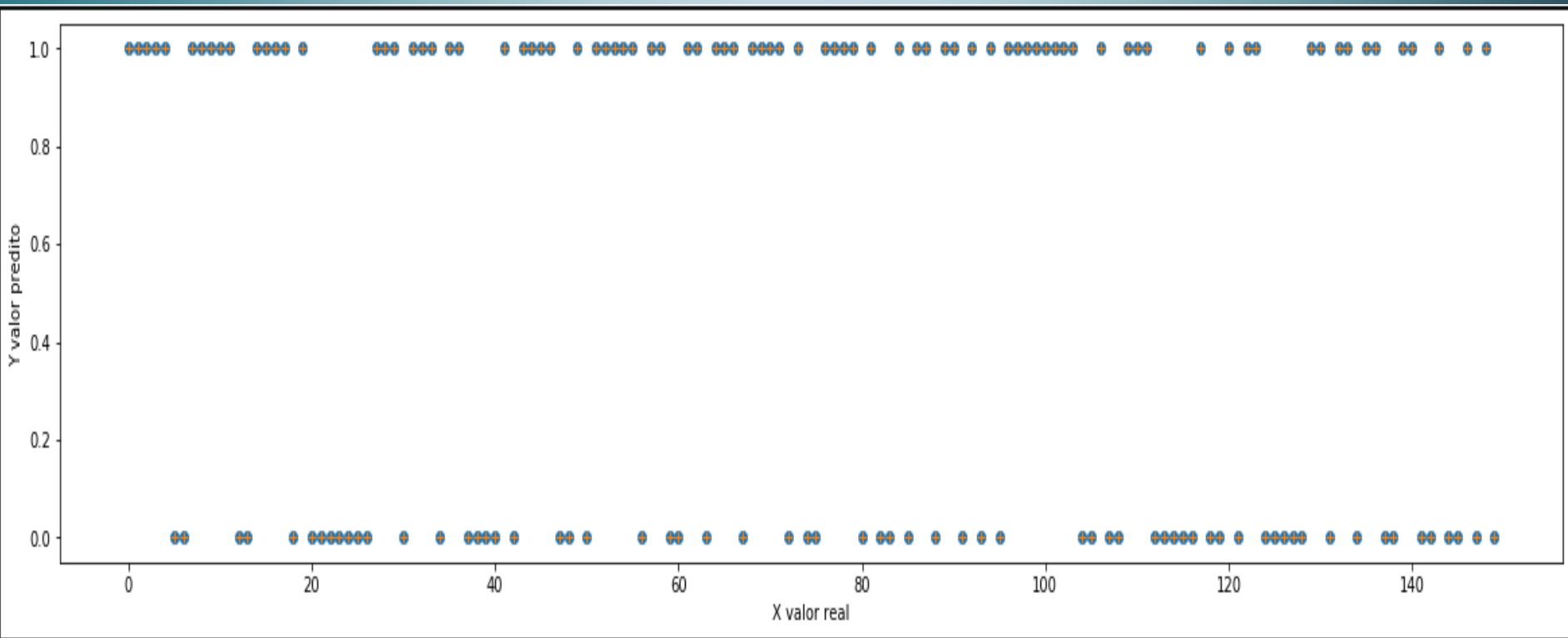
- O algoritmo de árvore decisão com a base1 89.7%,
- Cros Validation essa taxa teve um aumento de 4%.



DESENVOLVIMENTO

RESULTADO

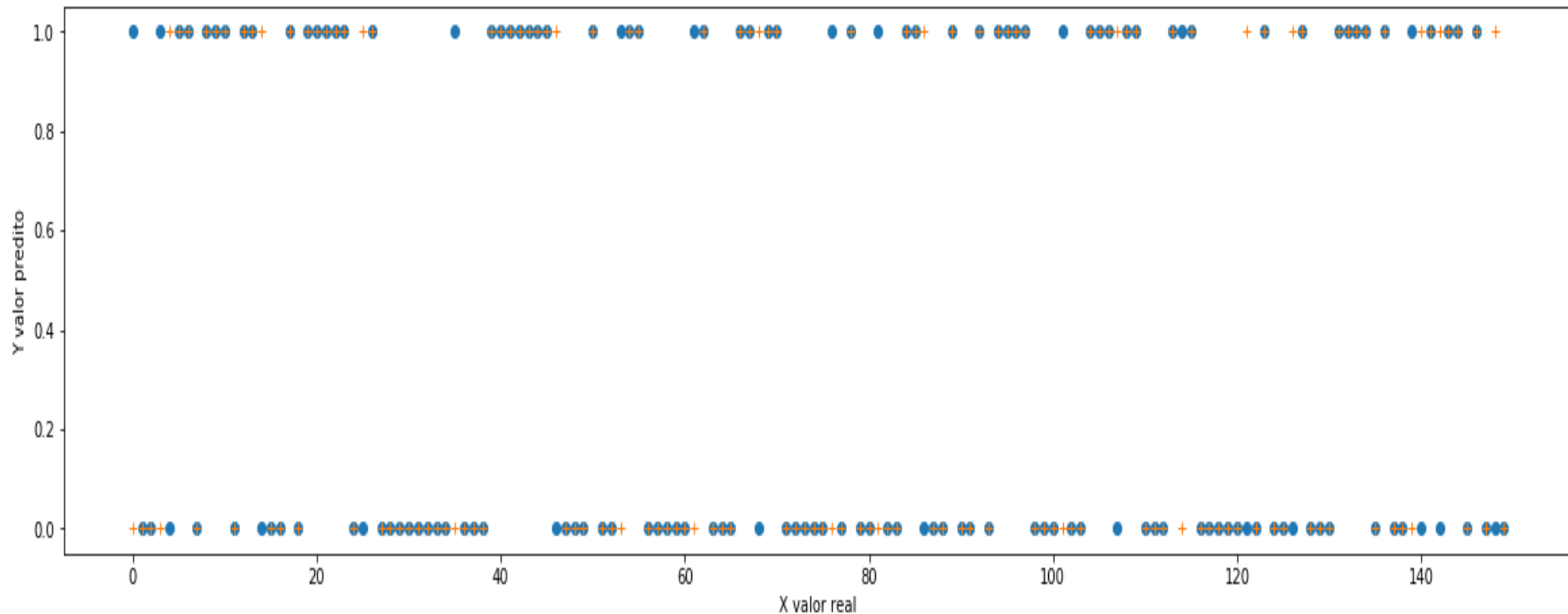
- O algoritmo de árvore decisão com a base2 97%,
- Cros Validation essa taxa teve um aumento de 3%.



DESENVOLVIMENTO

RESULTADO

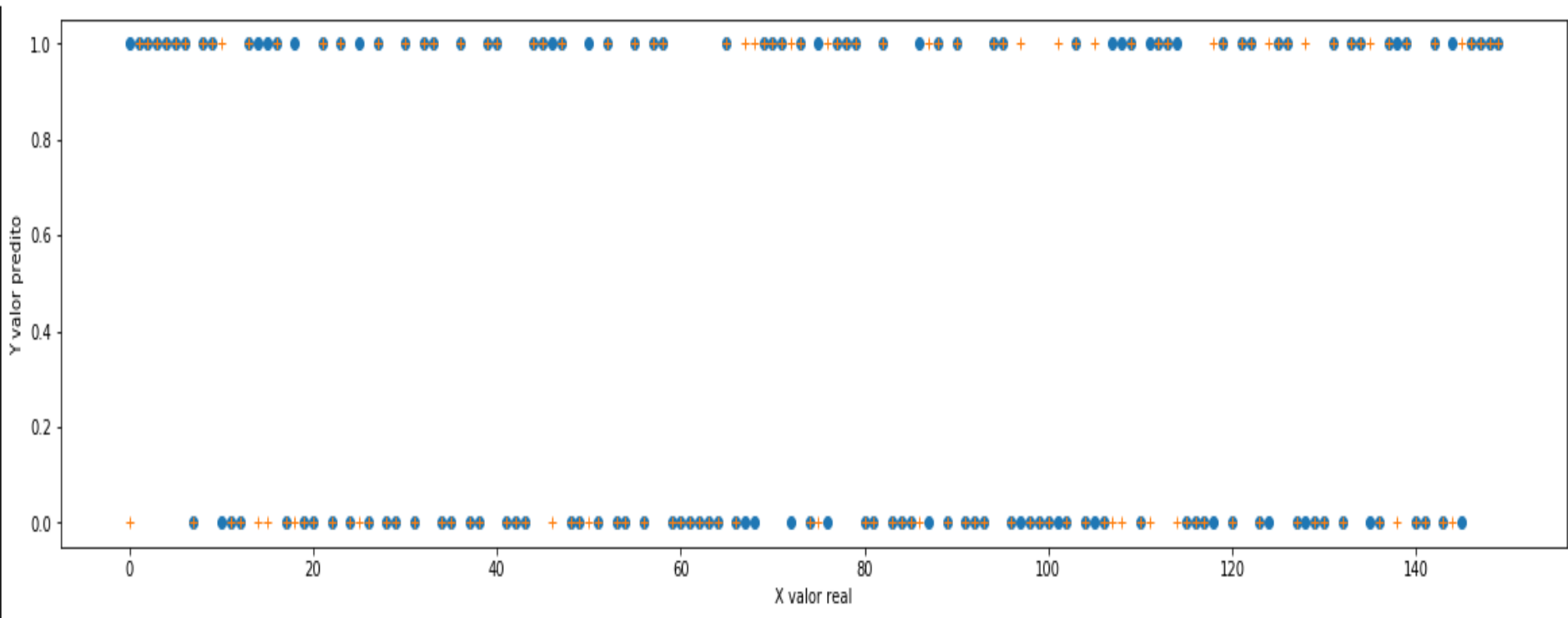
- O algoritmo de k-NN a base1 85%,
- Cros Validation essa taxa teve um aumento de 2.8%.



DESENVOLVIMENTO

RESULTADO

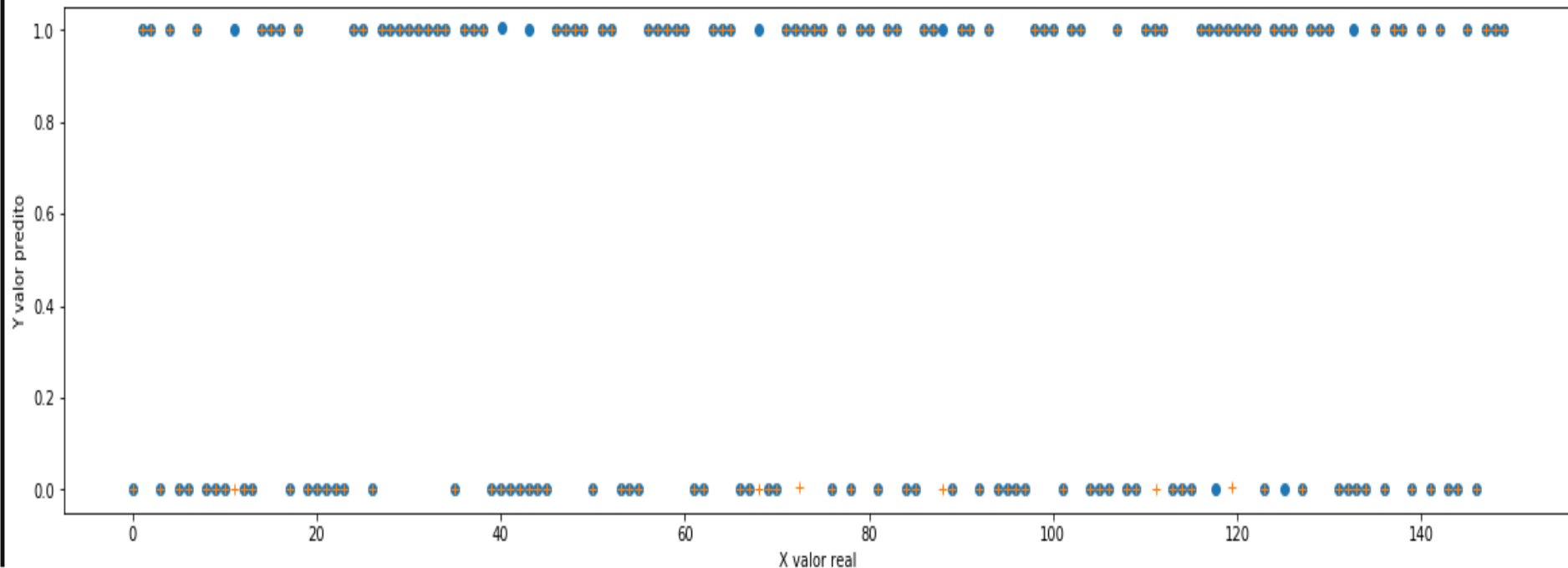
- O algoritmo de k-NN a base2 87%,
- Cros Validation essa taxa teve um aumento de 3.3%.



DESENVOLVIMENTO

RESULTADO

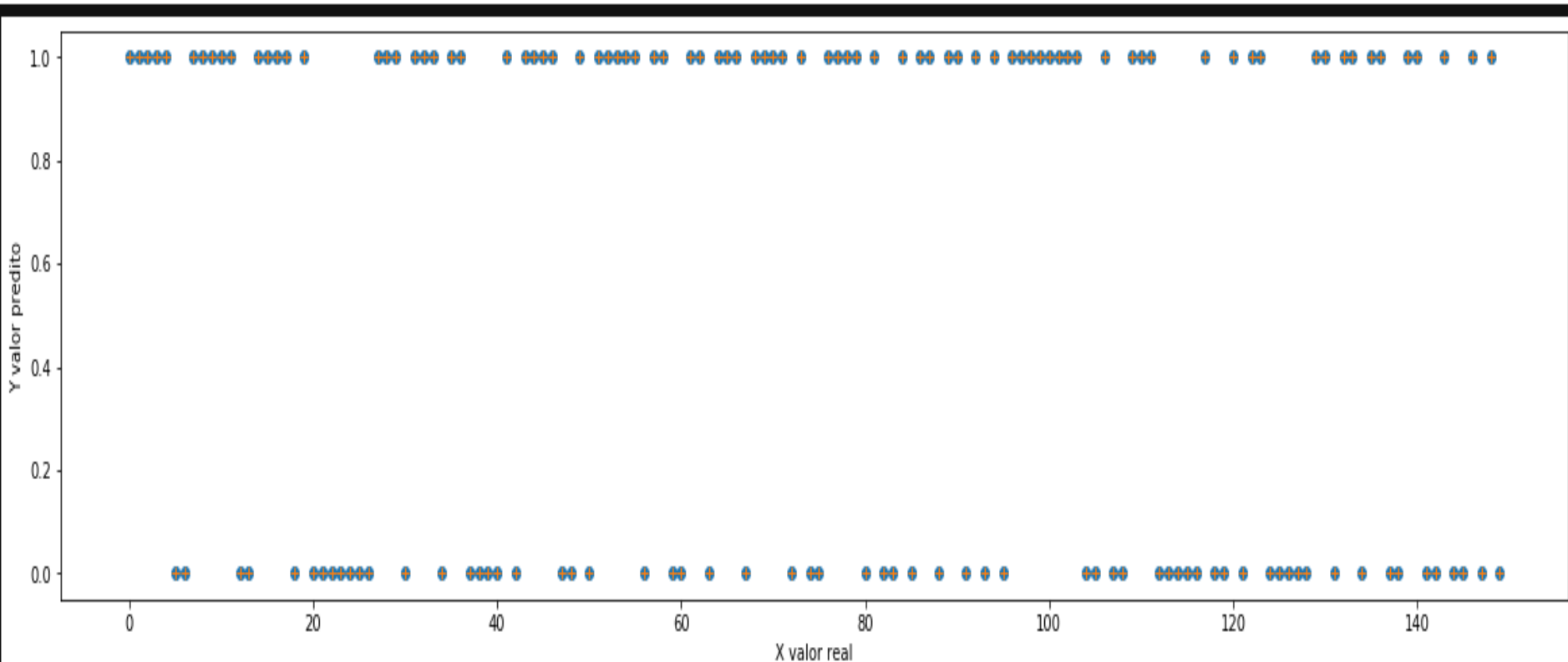
- O algoritmo de floresta aleatória com a base2 inicialmente foi testado com 10 arvores chegando a um resultado de 94% na taxa de acerto.



DESENVOLVIMENTO

RESULTADO

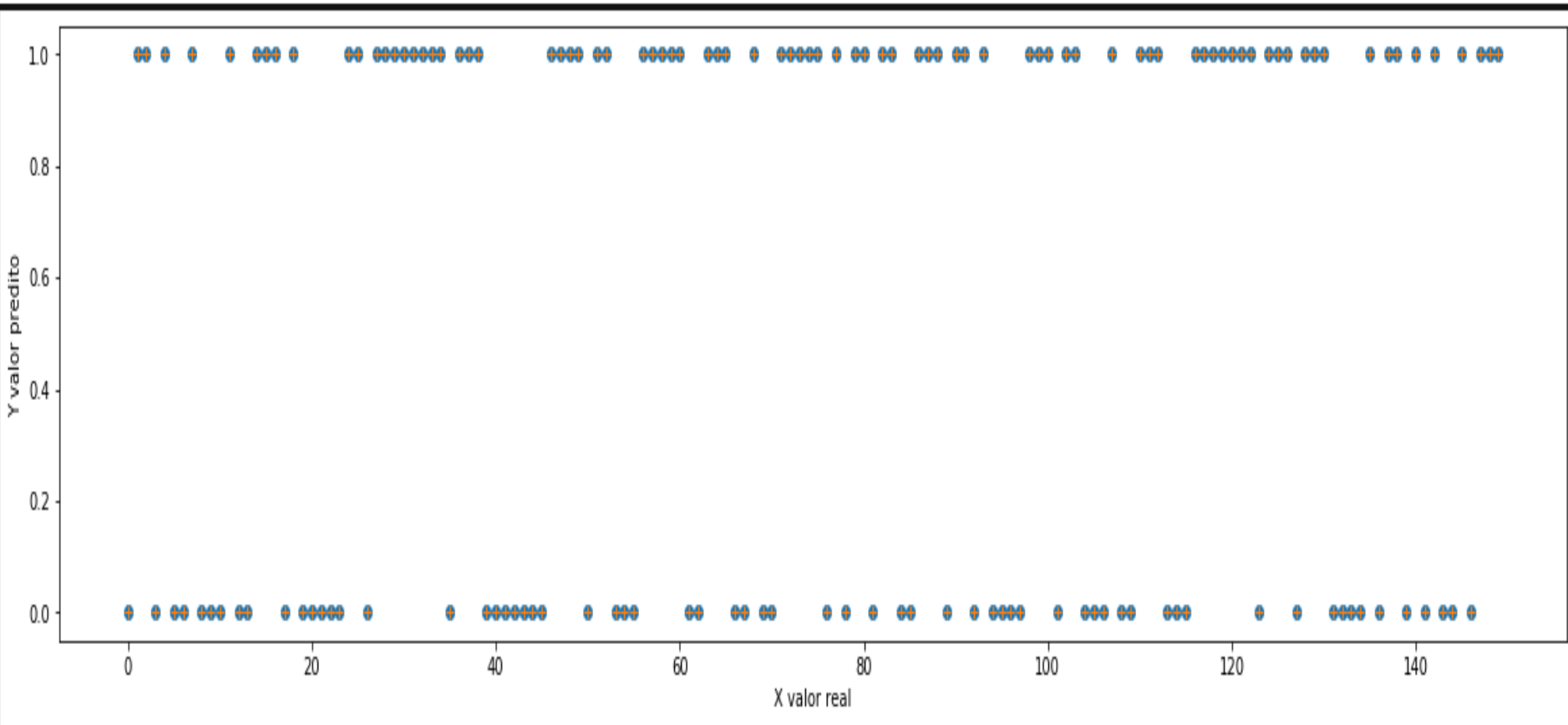
- O algoritmo de floresta aleatória com a base2 com 53 teve uma variação de 99% a 100%.



DESENVOLVIMENTO

RESULTADO

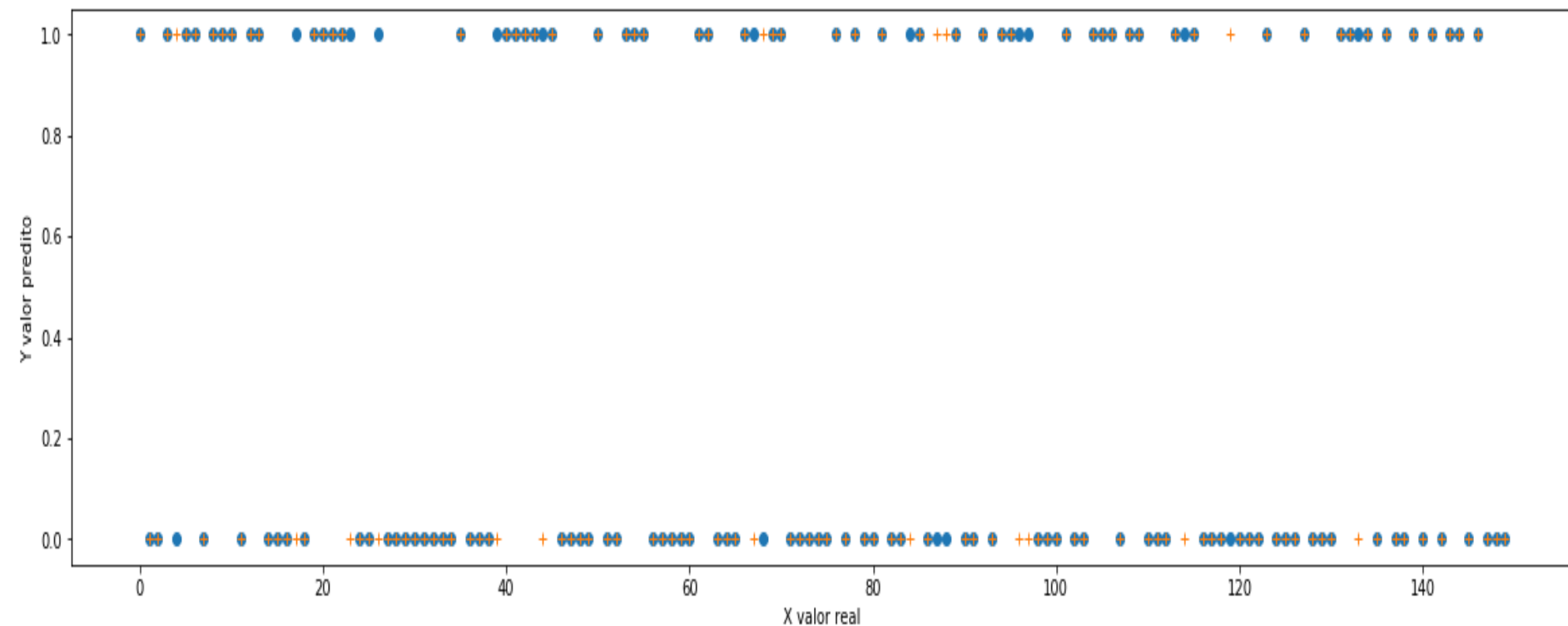
- regressão logística um taxa de acerto de 100%.



DESENVOLVIMENTO

RESULTADO

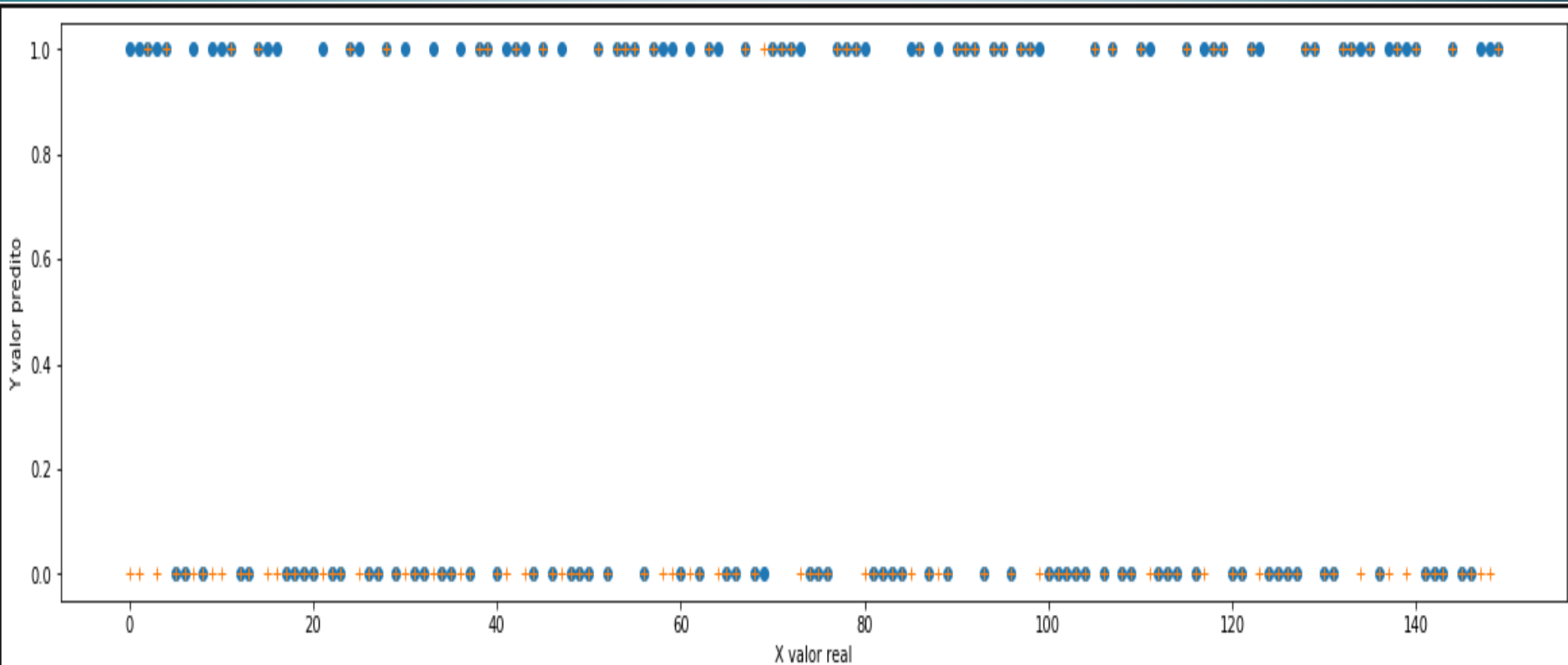
- Máquinas de vetores de suporte com a base1 teve um taxa de acerto de 80%



DESENVOLVIMENTO

RESULTADO

- Máquinas de vetores de suporte com a base2 teve um taxa de acerto de 87%



CONCLUSÃO

- 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 vitória ou seja, o valor da predição era de forma binária.
- O algoritmos árvore 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 árvore foi treinada.

CONCLUSÃO

- O algoritmo k-NN 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%.

CONCLUSÃO

- A regressão logística foi o algoritmo com a maior taxa de acertos, pois ele trabalha com os fatores binário de predição, sendo assim o oposto da regressão linear. Como os teste foram feitos para prever a derrota e a vitória.
- A máquinas de vetores trabalha definindo um limite linear logo para realizar a classificação ele separa os dados e os analisa para reconhecer padrões, assim que a uma entrada de um conjunto de dados e adicionada ele vai realizar a análise e dividir em duas classes, na qual as duas possíveis classes faz parte do classificador linear binário não probabilístico.

REFERENCIAS

- 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 PARA PREVISÃO DE JOGOS DE BASQUETE. [S.I.]: UNIVERSIDADE DO VALE DO TAQUARI, 2019.
- CASTRO, L. D. Introdução a Mineração De Dados: CONCEITOS BÁSICOS, ALGORITMOS E APLICAÇÕES SARAIVA EDITORA, 2016. ISBN 9788547200985. Disponível em:<https://books.google.com.br/books?id=7HxSvgAACAAji>.
- DEGENNARO, K. BWorld Robot Control Software. 2019. <https://news.sap.com/2017/08/corporate-sponsorships-reimagined-nba/i>. [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/notationaln analysis ofn sport.html](https://www.ebook.de/de/product/3473295/notationaln%20analysis%20of%20sport.html)

REFERENCIAS

- 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/jiaweinmann_michelines_kambers_jiann_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

REFERENCIAS

- 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/i>. [Online; accessed 20-Nov-2019].
- PANDAS. O projeto dos pandas. 2019. <https://pandas.pydata.org/i>. [Online; accessed 20-Nov-2019].
- PAPIC, 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.