Análise de Algoritmos de Aprendizagem de Máquina para Previsão de Precipitações para Utilização na Agricultura Familiar

Iago Magalhães de Mesquita[1], Francislane Teles Carneiro[1], Sarah Frota Alves[1], Leonardo Tabosa Albuquerque[1], Francisco Aldinei Pereira Aragão[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Estado do Ceará (IFCE) – Sobral – CE – Brasil.

{iago.magalhaes.mesquita61, francislane.teles.carneiro02, sarah.frota.alves08}@aluno.ifce.edu.br, {leonardo.tabosa, aldinei}@ifce.edu.br

Abstract. *This work proposes an analysis of data from meteorological elements such as: total precipitation (mm), atmospheric pressure at the station level (mb), air temperature - dry bulb (°c), relative humidity (%), direction (°gr), wind speed (m/s), seeking to predict precipitation and test Machine Learning models that best adapt to the data set used, namely: K-Nearest Neighbor (KNN), decision and the Multilayer Perceptron (MLP) neural network, with the objective of identifying the best model applied to rainfall forecasting applied in the region of Sobral-CE. In order to help small and medium agricultural producers, which are intrinsically related to the economic aspects of the region, which suffers from droughts and need to manage their water resources. In view of the models used, we had as a result that the decision tree obtained the best result with a hit rate of 99,995%, and the MLP of 99,693%, whereas the KNN obtained only 76,726%.*

Resumo. *Este trabalho tem como proposta uma análise de dados de elementos meteorológicos como: precipitação total (mm), pressão atmosférica ao nível da estação (mb), temperatura do ar - bulbo seco (°c), umidade relativa do ar (%), direção do vento (°gr), velocidade do vento (m/s), buscando prever a precipitação e testar modelos de Aprendizado de Máquina que melhor se adaptem ao conjunto de dados utilizados, sendo estes: K-Nearest Neighbor (KNN), Árvore de decisão e a rede neural Multilayer Perceptron (MLP), com o objetivo de identificar o melhor modelo aplicado a previsão de chuvas aplicados na região de Sobral- CE. A proposta tem como objetivo ajudar os pequenos e médios produtores agrícolas, que estão relacionados intrinsecamente aos aspectos econômicos da região, que sofre com estiagens e precisam gerir seus recursos hídricos. Diante dos modelos utilizados, a árvore de decisão obteve melhor resultado com uma taxa de acerto de 99.995%, e a MLP de 99.693%, já o KNN obteve apenas 76.726%.*

1. Introdução

A previsão da quantidade de precipitação a curto, médio e longo prazo têm sido objeto de estudo de várias pesquisas em todo o mundo, principalmente na área da agricultura. No Brasil, esse setor começou a se desenvolver junto ao processo do setor industrial, o que tornou o país um dos principais no ranking do agronegócio. Nesse contexto, as chuvas

são um dos meios de irrigação para as plantações e a principal fonte de reabastecimento de reservatórios na região nordeste do Brasil. A agricultura foi responsável por 26,6% do produto interno bruto (PIB) no ano de 2021 [ClimateFieldView, 2021], e ano após ano vem quebrando recordes de produção [Cepea 2022].

Muitas vezes agricultores familiares e pequenos produtores não têm acesso a grandes fontes de recursos hídricos em suas propriedades e dependem das chuvas para manter a lavoura irrigada e seus reservatórios em condições de uso. No Nordeste, devido ao seu clima semiárido, a região sofre com estiagens prolongadas e chuvas irregulares [Ramalho 2013].

Nos últimos anos, diversas obras foram feitas visando a redução do efeito da seca e a irregularidade das chuvas, objetivando apoiar os residentes da região. Tanto a transposição do rio São Francisco e construção de açudes ajudam nesta questão. Porém, o uso consciente e a economia de água ainda são ações primordiais para gerir bem os recursos hídricos.

Com base nos detalhes apresentados, o seguinte trabalho visa utilizar dados coletados por uma estação meteorológica e testar algoritmos de aprendizado de máquina (do inglês *Machine* Learning - ML) com o objetivo de selecionar o modelo que melhor se adeque ao conjunto de dados utilizado, buscando realizar previsões de precipitações na região em um período de até 2 anos. Assim seria possível auxiliar pequenos e médios cultivadores com relação ao uso de irrigação e agricultores familiares que dependem da chuva como única técnica de irrigação utilizados em seus cultivos. Trazendo como benefícios a redução dos gastos como energia elétrica, o maior aproveitamento de recursos hídricos e uma maior produtividade.

2. Revisão Bibliográfica

Muitos trabalhos relevantes podem ser apresentados seguindo diversas metodologias para realizar a previsão de precipitações em determinadas regiões. Muitas técnicas de ML são utilizadas para realizar previsões, principalmente de redes neurais.

Em [Rafaela *et al.* 2018], a proposta apresentada pelos autores foi a comparação de técnicas de ML para previsão de precipitações em Manaus utilizando registros de 65 anos da precipitação mensal da cidade. Utilizaram quatro abordagens: árvores de decisão, florestas aleatórias, redes neurais e vizinhos mais próximos. Nessas comparações, seus melhores resultados foram com as redes neurais e serviu como estratégia para reduzir os impactos ambientais. Obtendo uma acurácia de 50%.

Em [Pengcheng *et al.* 2018], a proposta apresentada pelos autores foi a de previsão de chuvas a curto prazo propondo a análise de alguns fatores definidos como principais, para entrada utilizando algoritmos de multicamadas com dados de 56 locais de meteorologia do mundo real na China. Além disso, foi proposto um algoritmo baseado em redes multicamadas chamado de DRCF, que chegou ao um valor de RMSE de 1.61.

Em [Zhang *et al.* 2020], os autores propõem utilizar diferentes técnicas de ML para a previsão de chuva com uma base de dados de 24 anos, onde o algoritmo SVM se mostrou promissor para realizar previsões anuais, já nos meses de janeiro, fevereiro, abril, outubro e novembro o SVR se destacou e nos demais a rede MLP foi mais promissora.

3. Material e Métodos

Segundo [Géron 2021], aprendizagem de máquina é a parte da ciência da computação em que possibilita que máquinas consigam aprender com dados sem serem explicitamente programados. O conceito de ML foi introduzido primeiramente para a resolução de problemas de reconhecimento de caracteres e posteriormente como filtro de SPAM. A figura 1 apresenta uma representação do funcionamento de aplicação de ML.

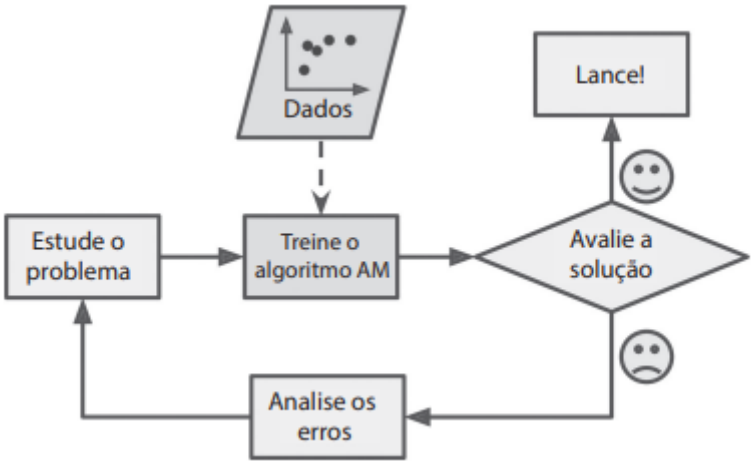

Figura 1. Adaptado de [Géron 2021].

A justificativa de sua utilização se dá por meio de características próprias do ML, como o fato de o programa ser muito menor e também ser possível realizar uma manutenção mais fácil em seu código fonte[Grus 2016].

Esta área abrange diversos algoritmos que funcionam de várias formas. Um conceito importante é como os algoritmos de ML podem ser treinados, com o aprendizado supervisionado e não supervisionado, onde o primeiro recebe algum tipo de supervisão durante o treinamento e o segundo recebe dados não rotulados e vai tentar aprender sem um professor[Géron 2021].

3.1. Base de Dados

A base de dados utilizada neste trabalho foi obtida junto ao Instituto Nacional de Meteorologia (INMET) no período de 2003 a 2020 na região de Sobral-CE. Os dados obtidos são de forma horária coletados por uma estação meteorológica situada na cidade. Uma representação da distribuição das precipitações ao longo dos anos presentes na base de dados pode ser visualizada na figura 2.

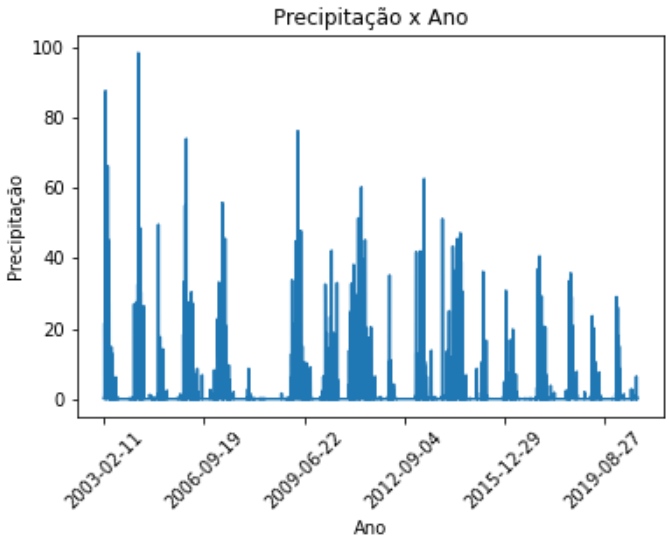

Figura 2. Base de dados.

3.2. Medidas de Avaliação

Para avaliar os resultados obtidos pelos modelos de aprendizagem de máquina, são utilizadas medidas estatísticas comumente utilizadas pela comunidade.

As métricas utilizadas nesse trabalho foram:

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

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

$$R2\ SCORE = \frac{\sum(\hat{y}_i - y_i)^2}{\sum(\hat{y}_i - y_i)^2} \quad (3)$$

Onde, o erro quadrático médio (*MSE*), com essa medida a diferença entre resultados previstos pelo modelo e os valores reais são elevados ao quadrado e posteriormente é retirada sua média. Sendo assim, quanto maior o seu valor, pior é o modelo. Já o *RMSE* ou raiz do erro quadrático médio, busca avaliar o erro do modelo em relação ao alvo, calculando a média do quadrado do erro antes de realizar o cálculo da raiz quadrada, o que possibilita a penalização de erros maiores. E o *R2 SCORE* retorna à proporção em relação ao valor da variável prevista e seu valor real. Seu valor é dado em porcentagem.

3.3. Árvore de Decisão

A árvore de decisão é um algoritmo que pode realizar diversas tarefas, uma delas é a regressão. Este modelo é capaz de moldar conjuntos complexos de dados. Seu funcionamento se dá por meio de nós, iniciando no nó raiz. Através desse nó é possível selecionar o próximo, que será o filho e assim por diante até chegar ao valor objetivado.

3.4. K-Nearest Neighbor

O algoritmo *K-Nearest Neighbor* (KNN) realiza a classificação e regressão baseado na distância entre algumas amostras, definida como k. Ele não possui parâmetros treináveis, pois o modelo mesmo propõe que a distância entre as variáveis é suficiente para a classificação ou regressão.

3.5. Multilayer Perceptron

A rede neural *Multilayer Perceptron* (MLP) é uma rede *feedforward* com uma ou mais camadas ocultas. Essas camadas podem possuir um número variável de neurônios. Através dessa arquitetura é possível resolver diversos problemas linearmente e não lineares separáveis. O treinamento se dá por meio de épocas, que realizam uma determinada quantidade de interações, em que as amostras são inseridas nas entradas e passam pelas camadas ocultas, atualizando os pesos e por fim, gerando uma saída que passa por uma função de ativação. Essa função pode gerar valores de saída em diferentes formatos, como probabilidade ou o próprio valor de classe.

Em seu treinamento é utilizado um algoritmo de aprendizado chamado *backpropagation*, que realiza a atualização dos pesos da rede de forma contrária da entrada dos dados, sendo da saída para a entrada. Existem outros modelos de algoritmos utilizados para treinamento, porém o backpropagation é o mais comum. Ele segue os seguintes passos para o treinamento da rede. O primeiro é a inicialização, onde os pesos das camadas são iniciados com valores aleatórios e pequenos uniformemente distribuídos. O segundo é a ativação, que realiza o cálculo dos valores dos neurônios da camada oculta e saída. O terceiro é treinar os pesos, que calcula os erros dos neurônios da camada oculta e saída, realiza a correção dos pesos e os atualiza. O último passo é a interação, que repete o processo a partir do segundo passo até atingir o valor de erro satisfatório.

4. Experimentos

A metodologia utilizada neste trabalho para realizar a previsão de precipitações foi de coletar dados de uma estação meteorológica próxima ao local onde será realizada a estimativa, logo após, é feito o processamento pelo algoritmo de ML que processará os dados e fornecerá o valor em mm da possível precipitação em um período de 24 horas. A figura 3 representa o fluxograma do processo realizado pela aplicação.

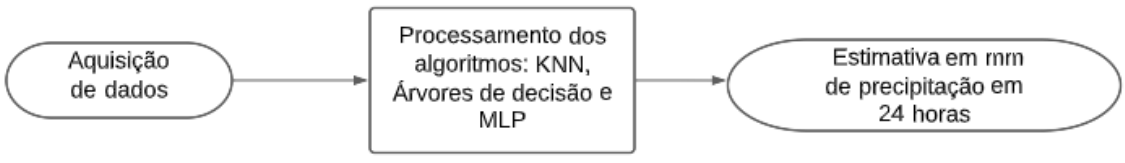

Figura 3. Fluxograma da aplicação.

Para realizar o treinamento dos modelos, foi utilizado a divisão dos dados em 70% para treino e 30% para teste. Também foi avaliado diversas configurações dos algoritmos testados e selecionados os melhores resultados de cada um. Além disso, as variáveis do conjunto de dados utilizadas pelos modelos foram: precipitação total (mm), pressão atmosférica ao nível da estação (mb), temperatura do ar - bulbo seco (°c), umidade

relativa do ar, direção do vento (°gr), velocidade do vento (m/s) buscando prever a precipitação.

Neste trabalho foram utilizados 3 algoritmos de ML para realizar a previsão. Sendo eles, o KNN, Árvore de decisão e a rede neural MLP. Todos foram treinados individualmente e avaliados a partir das variáveis citadas acima. Após o treinamento foi selecionado o modelo com melhor desempenho entre eles.

O algoritmo KNN seguiu a metodologia de avaliar entre as 3 medidas de cálculo de distância, sendo estas, a Manhattan, Euclidiana e Minkovski. Também foi testado o valor de k variando entre 1 e 30.

À árvore de decisão foram testadas apenas as configurações de variáveis de entrada no algoritmo.

Para a rede MLP foi realizado o teste com um número variável de camadas e neurônios, além de funções de ativação, como a tangente e a *ReLu*. Inicialmente foi testado com apenas 1 camada oculta e o número de neurônios variando de 100 a 1000, assim como as funções de ativação. A mesma metodologia foi empregada para testar com 2 camadas ocultas. Além disso, foram testadas as configurações de variáveis de entrada como citado anteriormente.

5. Resultados

Após a realização dos treinamentos dos 3 modelos, com os dados coletados foi possível verificar o algoritmo que melhor se adaptou ao conjunto de dados. A tabela 1 contém os resultados obtidos através das medidas de avaliação.

Tabela 1. Resultados das medidas de avaliação dos modelos.

Modelo	MSE	RMSE	R2 SCORE
KNN	0.0534063	0.2310980	76.726%
ÁRVORE DE DECISÃO	1.059602649006615e-05	0.0032551	99.995%
MLP	0.0007043	0.0265394	99.693%

Analisando cada modelo de forma individual pode-se notar que o algoritmo de Árvore de decisão apresentou os melhores resultados para a base de dados utilizada, assim como a MLP que também obteve ótimos resultados. Já o KNN se mostrou com capacidade de aprender com o conjunto utilizado.

Na figura 4 são exibidos gráficos com a realização de previsões em período de 2 anos, buscando verificar a eficiência dos modelos. A linha vermelha significa o valor real da precipitação em mm registrado pela estação meteorológica no dia em questão e a linha azul mostra a previsão do modelo.

Na figura 4 pode-se visualizar os resultados obtidos com o algoritmo KNN. O tempo de treinamento do algoritmo também foi levado em consideração e este modelo atingiu a marca de 0.1198 segundos com uma máquina de 12Gb de RAM e processador

Intel Core I7. Como é possível notar, o KNN até consegue entender as variações dos dados, porém, não o acompanha totalmente.

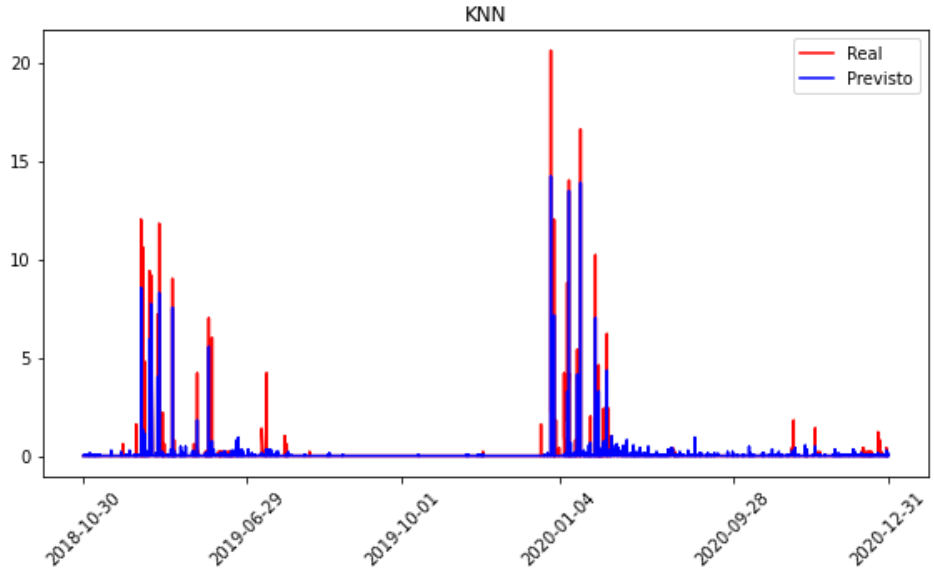

Figura 4. Resultados KNN.

Na figura 5 é possível visualizar os resultados da árvore de decisão. O tempo necessário para o treinamento do modelo foi de 0.38 segundos. Na figura é possível notar que o algoritmo se adaptou muito aos dados. Como é possível notar, os valores previstos quase sobrepõem totalmente os valores reais. Mostrando que o algoritmo atingiu resultados satisfatórios.

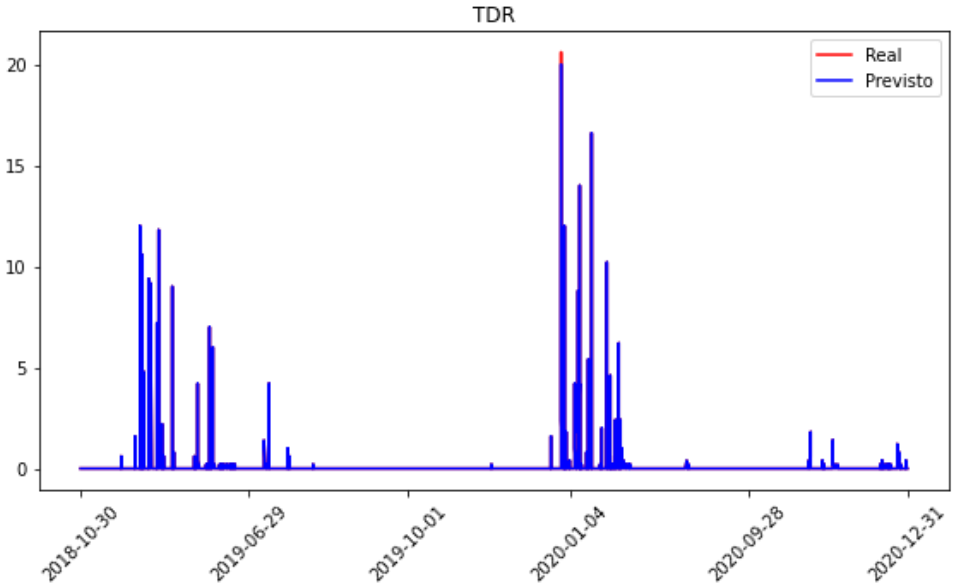

Figura 5. Resultados da árvore de decisão.

Já na figura 6 os resultados da MLP são apresentados. O tempo necessário para a convergência da rede foi de 30.49 segundos. Como é possível notar, os valores previstos sobrepõem totalmente os valores reais. Mostrando que o algoritmo atingiu resultados satisfatórios.

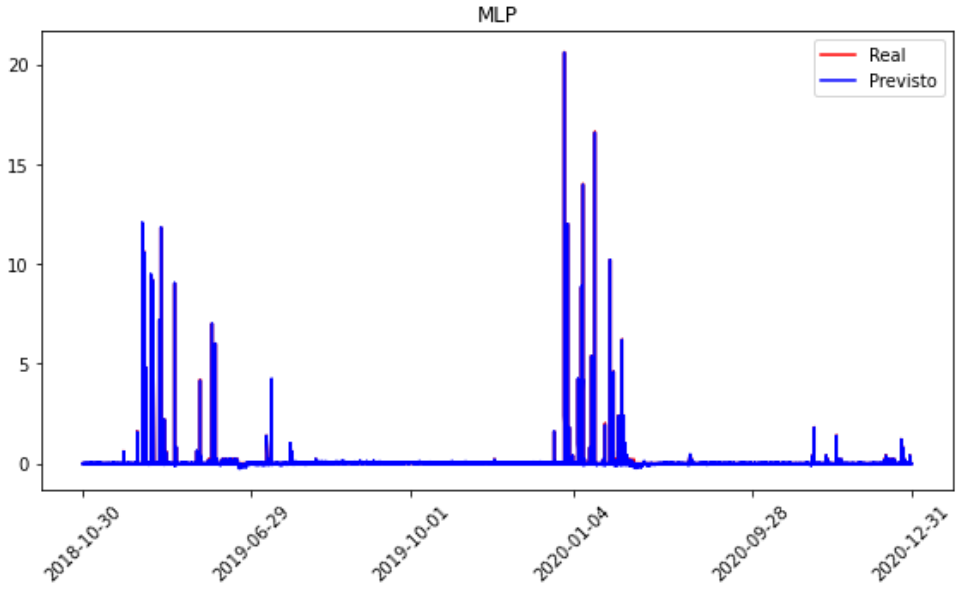

Figura 6. Resultados MLP.

Com os resultados obtidos, pode-se perceber que tanto o modelo de árvore de decisão quanto a rede neural MLP podem ser utilizadas para a realização de previsões de precipitações na região específica com uma precisão de até dois anos.

6. Conclusão

À vista das ponderações desse artigo apresentou-se uma abordagem para a realização de previsão de precipitações na região de Sobral-CE, com foco na utilização da agricultura de pequenos e médios produtores, que se utilizam da chuva como meio para irrigar e abastecer seus reservatórios. Diante da busca na literatura sobre a previsão de variáveis meteorológicas, com diferentes algoritmos de ML, concluímos que ainda existe uma carência de pesquisas que englobassem métodos semelhantes sobre a região de Sobral.

Em suma, diferentes técnicas podem solucionar problemas de previsão. Porém, à árvore de decisão se mostrou com melhores resultados, o que colaborou para previsão com prazo de até dois anos. Com esses resultados se torna previsível o momento adequado para a irrigação e garante o propósito de favorecer melhores condições no cultivo de diversas culturas.

Ante o exposto, o seguimento para trabalhos futuros pretende-se implementar essa metodologia em um software funcional, utilizando o modelo que apresentou melhor desempenho e desenvolver uma pequena estação meteorológica de baixo custo, utilizando ESP 32 que coletará dados de forma local, para maior precisão das informações da região.

Referencias

Qual a participação do agronegócio no PIB brasileiro?. [S. l.], Apr 06 2021. Disponível em: https://blog.climatefieldview.com.br/qual-e-a-participacao-do-agronegocio-no-pib-e-nas-exportacoes-brasileiras.

PIB do agronegócio brasileiro. [S. l.] Março de 2022. Disponível em: https://www.cepea.esalq.usp.br/br/pib-do-agronegocio-brasileiro.aspx.

Ramalho, Maria. A fragilidade ambiental do Nordeste brasileiro: o climasemiárido e as imprevisões das grandes estiagens. Sociedade e Território, Natal-RN, v. 25, n. 2, p. 104-115, 1 dez. 2013.

Rafaela dos Santos Sousa, Elloá B. Guedes, Maria Betânia Leal de Oliveira . Comparação de Técnicas de Aprendizagem de Máquinas para Previsão de Precipitações em Manaus . Em ENCOSIS 2018 , 2018.

Pengcheng Zhang, Yangyang Jia, Jerry Gao, Wei Song, Hareton Leung, "Short-term Rainfall Forecasting Using Multi-layer Perceptron", IEEE Transactions on Big Data, v. 6, p. 93 - 106, 19 Sep. 2018 .

Zhang, Xiaobo et al. Annual and Non-Monsoon Rainfall Prediction Modelling Using SVR-MLP: An Empirical Study From Odisha. IEEE Acess, [S. l.], v. 8, p. 30223-30233, 10 fev. 2020.

Géron, Aurélien. Mãos à obra: aprendizado de máquina com Scikit-Learn, Keras & TensorFlow: Conceitos, ferramentas e técnicas para a construção de sistemas inteligentes. [S. l.: s. n.], 2021.

Grus, Joel. Data Science do zero - Primeiras regras com o Python. [S. l.: s. n.], 2016.