Detecção de Anomalias em Dados Meteorológicos do Sertão de Pernambuco Utilizando Isolation Forest e DBSCAN

Anderson Rodrigues Cavalcante[1], **Victor Wanderley Costa de Medeiros**[1], **Glauco Estácio Gonçalves**[2]

[1]Universidade Federal Rural de Pernambuco (UFRPE) - Recife - PE - Brasil

[2]Universidade Federal do Pará (UFPA) - Belém - PA - Brasil

{anderson.rodrigues, victor.wanderley}@ufrpe.br, glaucogoncalves@ufpa.br

Abstract. *Anomalous values are some of the problems present in meteorological time series, which may appear due to defects, bad sensor configuration, and even extreme climate effects. Using non-supervised machine learning algorithms has become increasingly common for this type of problem. The present research intends to evaluate the usage of DBSCAN (Density Based Spatial Clustering of Application with Noise) and IF (Isolation Forest) for detecting anomalies present in the meteorological data on air temperature and relative humidity of Petrolina. Both Isolation Forest and DBSCAN, in their best hyperparameter settings, performed well. The IF had an accuracy of 98% and an F1 score of 95%. DBSCAN presented an accuracy of 97% and an F1 score of 94%. Both also got a revocation of 100%, which indicates that they did not classify values as false negatives, that is, no anomaly was considered normal.*

Resumo. *Valores anômalos são alguns dos problemas presentes em séries de dados meteorológicos, os quais podem aparecer por causa de defeitos, má configuração dos sensores e até mesmo efeitos climáticos extremos. O uso de algoritmos de aprendizado de máquina não supervisionado tem se tornado cada vez mais comum para este tipo de problema. Esta pesquisa avalia o uso do DBSCAN (Density Based Spatial Clustering of Application with Noise) e da IF (Isolation Forest) para detecção de anomalias presentes nos dados meteorológicos de temperatura e umidade relativa do ar de Petrolina. Tanto o Isolation Forest quanto o DBSCAN, em suas melhores configurações de hiperparâmetros, apresentaram bom desempenho. O IF apresentou uma acurácia de 98% e uma pontuação F1 de 95%. Já o DBSCAN apresentou uma acurácia de 97% e uma pontuação F1 de 94%. Ambos também obtiveram uma revocação de 100%, o que indica que não classificaram valores como falsos negativos, ou seja, nenhuma anomalia foi considerada normal.*

1. Introdução

O setor agrícola brasileiro cresceu 24,2% no ano de 2020, segundo o Centro de Estudos Avançados em Economia Aplicada (CEPEA, 2022). O agronegócio, em geral, alcançou o patamar de 26,6% do PIB. Estima-se que a área plantada terá um crescimento de 82 milhões de hectares em 2021 para 93,3 milhões em 2031, conforme o Ministério da Agricultura, Pecuária e Abastecimento (MAPA, 2021).

Doblas-Reyes *et al.* (2010) mostra que o clima tem grande influência na

produção agrícola, na demanda hídrica, no controle de pestes e doenças e na necessidade do emprego de fertilizantes. Chergui, Kechadi e McDonnell (2020) mostram que a análise de dados pode colaborar com a agricultura de diferentes formas como: no monitoramento do clima, de ervas daninhas, do rendimento da colheita e da irrigação.

Sabe-se ainda que as temperaturas diárias muito altas ou muito baixas podem prejudicar a taxa de crescimento da planta, a duração reprodutiva e o potencial de rendimento das culturas Hatfield e Prueger (2015). A mesma preocupação vale para a umidade relativa do ar. Segundo o Ministério da Agricultura da Colúmbia Britânica (2015), níveis anormais de umidade causam crescimento de organismos patogênicos e a não evaporação da água nas folhas das plantas caso ela esteja muito elevada. Assim, a tomada de decisão em processos agrícolas dependem de dados meteorológicos confiáveis. Sabe-se também que esses dados podem apresentar leituras anômalas, por falhas em sensores ou outras etapas do processo de coleta.

Anomalias são ocorrências incomuns em um conjunto não se encaixando nos padrões gerais formados pelos demais dados. Dasgupta e Forrest (1996) definem anomalias como qualquer novidade que exceda uma variação permitida. Agarwal e Gupta (2021) dizem que um outlier é uma observação que está longe do resto dos pontos em um conjunto de dados. Portanto, anomalia, novidade, ponto discrepante e outlier são termos empregados para definir pontos com características que não se encaixam nos padrões gerais do conjunto de dados.

As técnicas de identificação de anomalias em conjuntos de dados meteorológicos podem ser usadas para solucionar diferentes problemas. Angiulli e Fassetti (2006), por exemplo, as utilizaram para detectar anomalias e relacioná-las aos fenômenos climáticos El Niño e La Niña.

Arvor *et al.* (2008) demonstra a importância de se coletar dados confiáveis para análise, pontuando que dados climáticos possuem anomalias geradas por falhas em sensores e eventos meteorológicos extremos. Yuxiang *et al.* (2005) utilizou dados climáticos espaço-temporais do Sul da China para detectar valores discrepantes em certas épocas do ano em diferentes áreas.

Zemicheal e Dietterich (2019) pesquisaram como controlar a qualidade de valores climáticos vindos de sensores ausentes ou defeituosos. Já Wibisono *et al.* (2021) utilizou o DBSCAN para a detecção de anomalias em um conjunto de variáveis meteorológicas, entre elas estava a temperatura máxima e a umidade relativa do ar. Celik, Dadaser e Dokuz (2011) utilizaram o DBSCAN para detectar anomalias em dados de temperatura mensal e o seu desempenho foi avaliado comparando-o com o método estatístico que utiliza média e desvio padrão.

Sabendo da importância da temperatura e umidade relativa na atividade agrícola, este trabalho tem como principal objetivo avaliar, por meio de experimentação, algoritmos de aprendizado de máquina não supervisionados nas séries temporais dessas variáveis. Os métodos escolhidos nesta avaliação foram o *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) e *Isolation Forest* (IF). Estes algoritmos são bastante utilizados para detecção de anomalias em dados meteorológicos como fez Eze *et al.* (2022) e Gupta *et al.* (2021).

2. Referencial Teórico

2.1. *Density-Based Spatial Clustering of Applications with Noise*

O DBSCAN é um algoritmo de agrupamento cujo objetivo é formar grupos dos dados de entrada, separando-os em regiões de pontos a partir de um determinado raio. Assim, os valores escolhidos que formam círculos com baixa densidade de pontos são considerados anomalias.

Os principais parâmetros do algoritmo são: o número mínimo de amostras (*min_samples*) que cada grupo deve ter e o raio máximo (ε) que a área que cada subgrupo de valores deve possuir. O grupo que possuir um número de amostras menor do que o *min_samples* é considerado anomalia pois difere dos outros que possuem ao menos a quantidade mínima definida no modelo.

A Figura 1 ilustra o funcionamento do DBSCAN: os pontos-chave (vermelhos) são pontos que satisfazem os critérios para formar o grupo. Os pontos chamados de bordas (azuis) são os pontos que não satisfazem o critério do modelo, mas considerados dados normais porque conseguem estar em um grupo. Por fim, os pontos que são as anomalias (preto), não satisfazem nenhum dos casos anteriores.

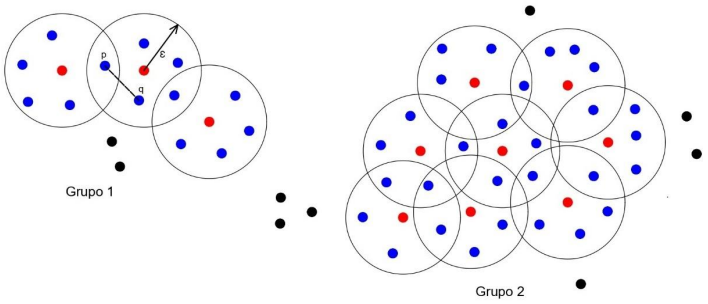

Figura 1 - DBSCAN e dois grupos de dados normais criados.

2.2. *Isolation Forest*

O ***Isolation Forest*** é um algoritmo de aprendizado não supervisionado baseado em árvores de decisão Liu, Ting e Zhou (2008). O conjunto de dados é dividido em árvores até que todos os pontos estejam isolados em uma folha. Os pontos que forem mais facilmente isolados, ou seja, folhas com menor profundidade, estão distantes dos outros e são, consequentemente, classificados como anomalias. A Figura 2 ilustra como o IF faz uso das árvores para sua classificação.

Essa floresta criada segue o parâmetro chamado *n_estimators*, que é o número de árvores binárias que serão criadas e terão a quantidade de valores conforme o fator *max_samples*. A contaminação (*contamination*) é a porcentagem aproximada de quantas anomalias existem no conjunto de dados. Este hiperparâmetro funciona como um limiar para decidir se uma determinada entrada será classificada como anomalia ou não.

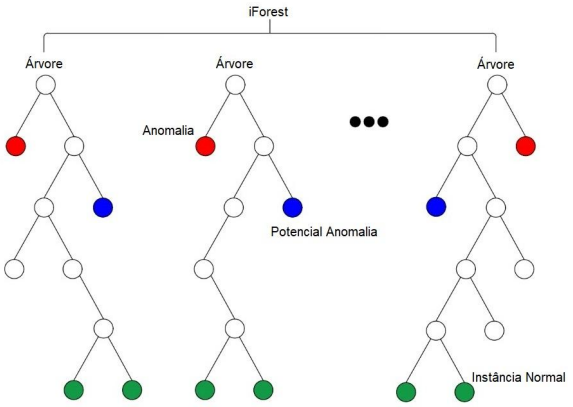

Figura 2 - Isolation Forest através de árvores. Fonte: Autor.

3. Materiais e Métodos

Para realização da avaliação proposta neste trabalho, utilizou-se a série temporal da **temperatura máxima** e da **umidade média relativa do ar**. Foram coletados dados diários no intervalo de tempo entre 01/01/2020 e 31/12/2021, obtidos da estação A307 do Instituto Nacional de Meteorologia (INMET), localizada no município de Petrolina, sertão de Pernambuco, com latitude de -9.388323° e longitude de -40.523262°.

Nesta fase, foi retirada a sazonalidade da série temporal para que ela ficasse livre de tendências utilizando uma função de auto regressão (AutoReg) com lag 1 da biblioteca Stats Models. Após isso, retirou-se uma amostra aleatória de 5% de dados (com probabilidade uniforme) dentro da série temporal. Nesse subconjunto, para aferir a eficácia dos algoritmos, o convertemos em anomalias empregando o método utilizado por Basu e Meckesheimer (2007) e Y. Lu *et al.* (2018) que foram inseridas nas mesmas posições de onde foram retirados os dados reais.

Para geração das anomalias utilizou-se o esquema abaixo:

$$Y = Y + \text{sign}(Y)\ X\sigma_{\epsilon}, \quad (1)$$

$$X = \exp(3 \times \text{abs}(x_1)) + 3, \quad (2)$$

Na Equação 1, o desvio padrão da série temporal até o index do valor (temperatura ou umidade) é calculado e em seguida multiplicado pelo valor da Equação 2, onde que x_1 é um número gerado aleatoriamente entre -0,5 e 0,5. Essa multiplicação da Equação 1 recebe o *sign*[1] do valor atual e altera a operação para soma ou subtração, originando o dado anômalo.

Ao final deste processo tem-se a série temporal com anomalias, cujo resultado pode ser visto na Figura 3, onde as linhas azul denotam os valores da série temporal sem a componente de tendência e os pontos em preto são as anomalias inseridas.

[1] *sign* - Sinal do inteiro obtido pela função *np.sign,* da biblioteca numpy. Se o valor for menor que 0, será retornado -1. Se for maior, retornará 1.

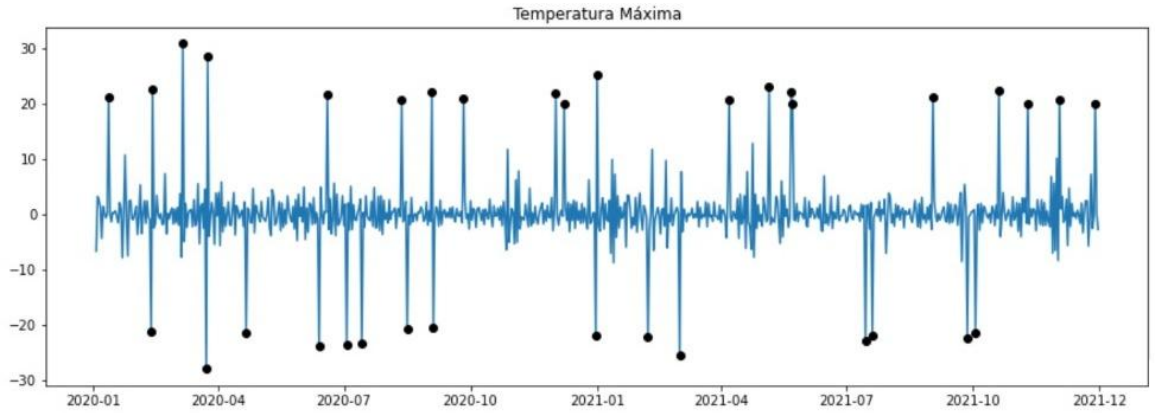

Figura 3 - Série temporal da temperatura máxima com anomalias inseridas.

O experimento foi implementado na linguagem Python e utilizou-se a implementação dos algoritmos disponíveis na biblioteca *Scikit-learn e Stats Models.* Os fatores utilizados no experimento estão dispostos na Tabela 1 e seus níveis foram baseados nos valores empregados em trabalhos anteriores Celik, Dadaser e Dokuz (2011), Liu, Ting e Zhou (2008) bem como nos valores padrão da biblioteca, os quais são por vezes, utilizados na prática sem maior consideração.

Algoritmos	**Fatores**	**Níveis**
DBSCAN	eps	0.3, 0.5 (Padrão)
	min_samples	5, 8, 10
	metric	‘euclidean’ (padrão)
	algorithm	‘auto’ (padrão)
Isolation Forest	n_estimators (número de árvores)	100
	max_samples	‘auto’ (padrão)
	contamination	‘auto’ (padrão), 0.05, 0.08, 0.1

Tabela 1 - Fatores utilizados na avaliação de cada algoritmo.

Para a comparação dos modelos avaliados neste trabalho utilizou-se as mesmas métricas apresentadas em Luo, Jia e Zhang (2019): revocação, precisão, acurácia, e pontuação F1. Nossos dados serão classificados em anomalia e não anomalia. Uma anomalia fará parte da classe Positiva (P) e um valor normal da classe Negativa (N). O resultado pode ser Verdadeiro (T), em caso de acerto, ou Falso (F).

4. Resultados e Discussões

A seguir são apresentados os resultados da avaliação em duas seções, uma para o IF (Seção 4.1) e outra para o DBSCAN (Seção 4.2).

4.1. Isolation Forest

Na Figura 4 o algoritmo classifica como anomalia os pontos em preto e como dados normais os pontos em vermelho.

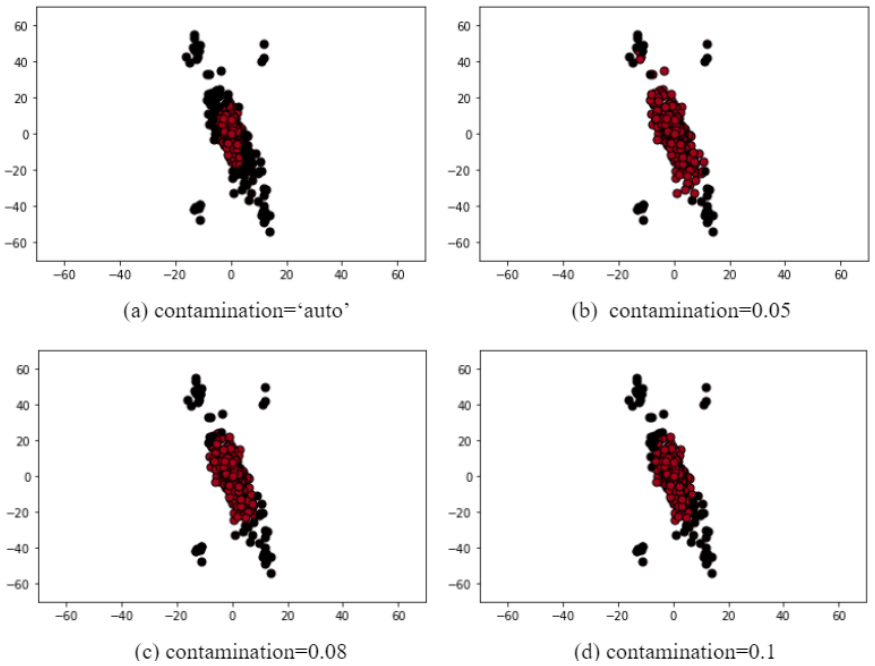

Figura 4 - Resultados gráficos do Isolation Forest.

Os resultados do IF podem ser visualizados na Tabela 2. Todos os campos da revocação tiveram resultado 1, pois nenhum falso negativo foi gerado nos resultados, ou seja, em todos os casos não houve nenhum registro de uma anomalia que foi considerado um valor normal.

Apesar da acurácia ter ficado acima de 90% em todos os casos, não podemos nos basear isoladamente em uma única métrica pois vemos que alguns valores mesmo com a acurácia alta a precisão está baixa devido ao alto número de Falsos Positivos. Um exemplo disso é a contaminação determinada como 'auto', Figura 7 (a), que atingiu uma acurácia de 91%, mas uma precisão de 36%, tendo um alto índice de dados normais sendo considerados anomalias. É possível observar o mau desempenho do hiperparâmetro *contamination* configurado para 'auto' ao comparar a Figura 4(a) com as Figuras 4(b), 4(c) e 4(d). O número de pontos pretos indicando anomalias é visivelmente superior.

Árvores	Contaminação	Precisão	Acurácia	Revocação	F1
100	auto	0.36	0.91	1	0.53
100	**0.05**	**0.94**	**0.98**	**1**	**0.95**
100	0.08	0.61	0.96	1	0.75
100	0.1	0.49	0.94	1	0.66

Tabela 2 - Resultados obtidos pelo Isolation Forest.

O parâmetro *contamination* mostrou-se bastante sensível, já que mesmo a série temporal tendo apenas 5% de anomalias, um valor mais alto de *contamination* faz com que o IF busque por mais anomalias. Dessa forma, um valor de 8% ou 10% leva a mais Falsos Positivos. Este fato é evidenciado pela alta acurácia (a maioria das anomalias foi identificada corretamente) e pelo baixo valor de precisão e pontuação F1 para os valores de *contamination* iguais a 0,08 e 0,1.

Os resultados gerados pela *contamination* de 5%, Figura 4(b) foram os melhores para o IF. Isso mostra que, se for conhecida, ainda que aproximadamente, a proporção de dados anômalos do conjunto de dados, o *Isolation Forest* pode ter resultados com precisão, pontuação F1 e revocação acima de 90%.

4.2. DBSCAN

O DBSCAN teve seu melhor resultado (Tabela 3) utilizando um ε (eps) de 0,5 e um número mínimo de amostras de 8 pontos, Figura 5(e). Todas as anomalias foram detectadas, o que fez a revocação ser de 100%, sendo o melhor resultado do algoritmo.

O mesmo bom desempenho não acontece com o ε de 0,3 e número mínimo de amostras de 8, Figura 5(b), onde a taxa de precisão caiu para 55% e a pontuação F1 para 71%, mesmo que acurácia tenha sido de 96%. Ou seja, muitos pontos de valores normais (classe negativa) foram considerados anomalias, resultando numa elevada taxa de Falsos Positivos.

Utilizando ε = 0,3 e número mínimo de amostras 10, Figura 5(c), a precisão foi ainda menor, apenas 44%. O algoritmo está sendo menos tolerante no que deve ser uma anomalia, ou seja, cada círculo deve ser formado por 10 amostras, que é um número maior, e 0,3 de raio, que é uma área menor, sendo difícil de atender a esses requisitos, visto que é um número maior de amostras em um círculo com raio menor. Como citado, pela precisão de 44% e a F1 de 61%, muitos valores normais foram considerados anomalias, por ser mais difícil atender a esses critérios.

eps	min_samples	Precisão	Acurácia	Revocação	F1
0.3	5	0.85	0.99	1	0.92
0.5	5	0.92	0.98	0.8	0.85
0.3	8	0.55	0.96	0.55	0.71
0.5	**8**	**0.93**	**0.97**	**1**	**0.94**
0.3	10	0.44	0.93	1	0.61
0.5	10	0.78	0.98	1	0.88

Tabela 3 - Resultados obtidos pelo DBSCAN. Fonte: Autor

Podemos observar na Figura 5 que o DBSCAN forma grupos (como mostrado na Figura 1), destacados em cores distintas. A Figura 5(c) (min_samples=10, ε=0,3), com uma precisão de apenas 44% e F1 de 61%, formou um grupo. Isso porque os outros grupos da classe negativa (valores normais) foram considerados como anomalias. Diferentemente do que aconteceu na Figura 5(a) com o min_samples=5 e ε=0,3, por exemplo, que obteve bons resultados (precisão=85%, F1=92%) e formou três grupos de valores não anômalos.

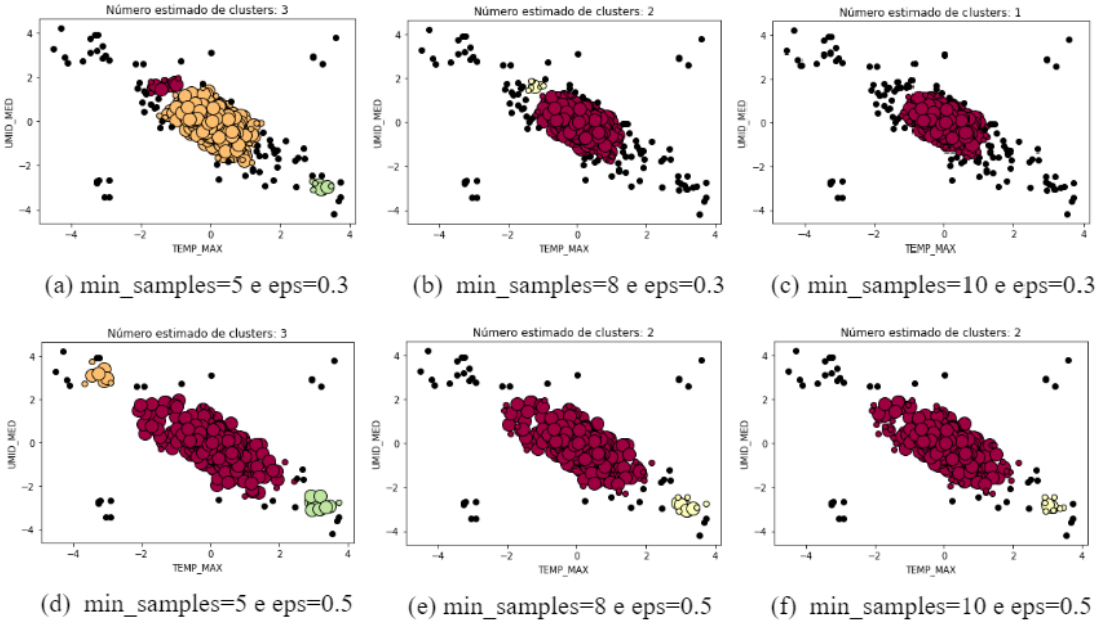

Figura 5. Resultados gráficos do DBSCAN. Fonte: Autor

5. Considerações Finais

Dado o cenário de crescimento do setor agrícola, este trabalho avaliou dois algoritmos de aprendizagem de máquina não supervisionada para detecção de anomalias em dados meteorológicos. A principal contribuição é avaliar como estes algoritmos podem ser usados com este tipo de dados e como seus parâmetros impactam sua eficiência.

A acurácia geral foi superior a 90% em todos os testes. Contudo, analisando outras métricas de desempenho, observamos que, de forma geral, apenas algumas configurações dos algoritmos conseguiram atingir valores acima de 90% em todas as métricas: acurácia, precisão, revocação e pontuação F1.

Frente ao DBSCAN, o algoritmo IF demonstrou melhor precisão e acurácia na identificação de anomalias, com uma taxa de Falsos Negativos de zero. Assim, para temperatura máxima e umidade média relativa do ar, o IF possui uma boa vantagem para casos em que não se pode haver Falsos Negativos. Contudo, mostrou-se sensível ao *contamination*, o qual impacta diretamente na taxa de falsos positivos do algoritmo.

Já o DBSCAN mostrou resultados consistentes mesmo utilizando os valores padrão da biblioteca *scikit learn* para os hiperparâmetros do algoritmo, ou seja, seria possível obter bons resultados sem realizar uma exploração dos hiperparâmetros. Isso pode indicar que o DBSCAN seja uma boa escolha quando se deseja utilizar um método de detecção de anomalias sem realizar ajustes finos.

Estes resultados mostram que a escolha do algoritmo para detecção de anomalias deve considerar diferentes aspectos, que variam com a experiência do usuário do algoritmo, o conhecimento que este tem sobre os dados manipulados e a aplicação que os dados terão.

Trabalhos futuros nesta área envolvem a avaliação de outras técnicas de detecção de anomalias, como a Análise Topológica de Dados Santos *et al.* (2019), que faz uso da

topologia algébrica para encontrar estruturas nos dados, como a sua forma e conectividade, também sendo útil para a detecção de anomalias.

5. Agradecimentos

Ao Juá Labs, que tem sido um importante meio de fomento à pesquisa científica na Universidade Federal Rural de Pernambuco.

Referências

Agarwal, A., & Gupta, N. (2021). Comparison of Outlier Detection Techniques for Structured Data. arXiv preprint arXiv:2106.08779.

Angiulli, F., & Fassetti, F. (2007). Detecting distance-based outliers in streams of data. In Proceedings of the 16^{th} ACM conference on Conference on information and knowledge management (pp. 811-820).

Arvor, D., Jonathan, M., Meirelles, M. S. P., & Dubreuil, V. (2008). Detecting outliers and asserting consistency in agriculture ground truth information by using temporal VI data from modis. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, v. 37, pt. B7, p. 1031-1036, 2008. Edition of Proceedings of 11^{th} ISPRS Congress, Beijing, Jul. 2008.

Basu, S., & Meckesheimer, M. (2007). Automatic outlier detection for time series: an application to sensor data. Knowledge and Information Systems, 11, 137-154.

Celik, M., Dadaşer-Celik, F., & Dokuz, A. S. (2011). Anomaly detection in temperature data using DBSCAN algorithm. In 2011 international symposium on innovations in intelligent systems and applications (pp. 91-95). IEEE.

Centro de Estudos Avançados em Economia Aplicada (2020) Metodologia - PIB do Agronegócio, 2020, Disponível em: <https://cepea.esalq.usp.br/upload/kceditor/files/sut.pib_dez_2020.9 mar2021.pdf>.

Chergui, N., Kechadi, M. T., & McDonnell, M. (2020). The impact of data analytics in digital agriculture: A review. In 2020 International Multi-Conference on: "Organization of Knowledge and Advanced Technologies"(OCTA) (pp. 1-13). IEEE.

Dasgupta, D., & Forrest, S. (1996). Novelty detection in time series data using ideas from immunology. In Proceedings of the international conference on intelligent systems (pp. 82-87).

Doblas-Reyes, F., Garcia, A., Hansen, J., Mariani, L., Nain, A., Ramesh, K. & Venkataraman, R. (2003). Weather and climate forecasts for agriculture. Guide to agricultural, meteorological practices.

Eze, C., Okeke-Uzodike, O. E., Emmanuel, E. I., & Mkpojiogu, E. O. (2022). Emotional Intelligence as a Predictor of Success in e-Learning Engagement During COVID-19: A Case of Veritas University Abuja, Nigeria. In ICT Infrastructure and Computing: Proceedings of ICT4SD 2022 (pp. 275-286). Singapore: Springer Nature Singapore.

Gupta, R., Nahrstedt, K., Suri, N., & Smith, J. (2021). Svad: End-to-end sensory data

analysis for iobt-driven platforms. In 2021 IEEE 7th World Forum on Internet of Things (WF-IoT) (pp. 903-908). IEEE.

Hatfield, J. L., & Prueger, J. H. (2015). Temperature extremes: Effect on plant growth and development. Weather and climate extremes, 10, 4-10.

INMET (2011). Rede de estações meteorológicas automáticas do INMET. Relatório Técnico, Instituto Nacional de Meteorologia.

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. In 2008 8^{th} IEEE international conference on data mining (pp. 413-422). IEEE.

Lu, Y., Kumar, J., Collier, N., Krishna, B., & Langston, M. A. (2018). Detecting outliers in streaming time series data from ARM distributed sensors. In 2018 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 779-786). IEEE.

Luo, H., Jia, S., & Zhang, W. (2019). Hierarchical temporal memory based anomaly detection for hydrological monitoring of unmanned surface vehicle. In 2019 IEEE 2^{nd} International Conference on Information Communication and Signal Processing (ICICSP) (pp. 420-424). IEEE.

Ministério da Agricultura, Pecuária e Abastecimento (2021). Projeções do Agronegócio 2020-2021 a 2030-2031. Disponível em: <https://www.gov.br /agricultura/pt-br/assuntos/politica-agricola/todas-publicacoes-de-politica-agricola/projecoes-do-agronegocio/projecoes-do-agronegocio-2020-2021-a-2030-2031.pdf/view >

Ministry of Agriculture of British Columbia (2015) Understanding Humidity Control in Greenhouses, Disponível em: <https://www2.gov.bc. ca/assets/gov/farming-natural-resources-and-industry/agriculture-and-seafood/animal -and-crops/crop-production/understanding_humidity_control.pdf>.

Santos, M. F., Oliveira, W. R., Amorim, M., & Stosic, T. (2019). Análise topológica de dados para caracterização de periodicidade em séries temporais de dados pluviométricos. Em Revista Mundi Engenharia, Tecnologia e Gestão (ISSN: 2525-4782). 4.

Wibisono, S., Anwar, M. T., Supriyanto, A., & Amin, I. H. A. (2021). Multivariate weather anomaly detection using DBSCAN clustering algorithm. In Journal of Physics: Conference Series (Vol. 1869, No. 1, p. 012077). IOP Publishing.

Yuxiang, S., Kunqing, X., Xiujun, M., Xingxing, J., Wen, P., & Xiaoping, G. (2005). Detecting spatio-temporal outliers in climate dataset: A method study. In Proceedings. 2005 IEEE International Geoscience and Remote Sensing Symposium, 2005. IGARSS'05. (Vol. 2, pp. 4-pp). IEEE.

Zemicheal, T., & Dietterich, T. G. (2019). Anomaly detection in the presence of missing values for weather data quality control. In Proceedings of the 2^{nd} ACM SIGCAS Conference on Computing and Sustainable Societies (pp. 65-73).