Anderson Rodrigues Cavalcante

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

Recife

Junho 2021

Anderson Rodrigues Cavalcante

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

Artigo apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Universidade Federal Rural de Pernambuco – UFRPE
Departamento de Estatística e Informática
Curso de Bacharelado em Sistemas de Informação

Orientador: Victor Wanderley Costa de Medeiros

Recife

Junho 2021

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

C377d Cavalcante, Anderson Rodrigues
Detecção de anomalias em dados meteorológicos do sertão de Pernambuco utilizando Isolation Forest e DBSCAN / Anderson Rodrigues Cavalcante. - 2022.
20 f. : il.

Orientador: Victor Wanderley Costa de Medeiros.
Inclui referências.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Sistemas da Informação, Recife, 2022.

1. DBSCAN. 2. Isolation Forest. 3. Detecção de anomalias. 4. Meteorologia. I. Medeiros, Victor Wanderley Costa de, orient. II. Título

CDD 004

ANDERSON RODRIGUES CAVALCANTE

DETECÇÃO DE ANOMALIAS EM DADOS METEOROLÓGICOS DO SERTÃO DE PERNAMBUCO UTILIZANDO ISOLATION FOREST E DBSCAN

Artigo apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

Aprovada em: 02 de Junho de 2022.

BANCA EXAMINADORA

Victor Wanderley Costa de Medeiros (Orientador)
Departamento de Estatística e Informática
Universidade Federal Rural de Pernambuco

Glauco Estácio Gonçalves
Instituto de Tecnologia
Universidade Federal do Pará

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]

[1]Departamento de Estatística e Informática – Universidade Federal Rural de Pernambuco Rua Dom Manuel de Medeiros, s/n, - CEP: 52171-900 – Recife – PE – Brasil

[{anderson.rodrigues, victor.wanderley}@ufrpe.br]

Resumo. *Valores anômalos são uns dos problemas presentes na era do Big Data. São necessárias técnicas robustas para a manipulação de informações corretas e incorretas que a cada instante são geradas. O uso de algoritmos de aprendizado de máquina não supervisionados dá a confiança de um bom desempenho nos resultados finais. Esta pesquisa utilizará dados meteorológicos de temperatura e umidade relativa do ar vindos do Instituto Nacional de Meteorologia, de Petrolina, com o DBSCAN (Density Based Spatial Clustering of Application with Noise) e o IF (Isolation Forest) implementados para detectar anomalias presentes nos dados, visto que anomalias meteorológicas podem aparecer por meio de defeitos, má configuração dos sensores e até mesmo efeitos climáticos extremos.*

Palavras-chave: DBSCAN, Isolation Forest, Detecção de anomalias, Meteorologia.

Abstract. Anomalous values are one of the problems present in the Big Data age. Robust techniques are required to manipulate correct and incorrect information that is generated at each time. Using non-supervised machine learning algorithms gives the confidence of good performance in the final results. This research will use meteorological data on air temperature and relative humidity from the Instituto Nacional de Meteorologia, of Petrolina, with DBSCAN (*Density Based Spatial Clustering of Application with Noise*) and IF *(Isolation Forest)* implemented to detect anomalies present in the data, since weathering meteorological anomalies may appear through defects, bad sensor configuration and even extreme climate effects.

Keywords: DBSCAN, Isolation Forest, Anomaly Detection, Meteorology.

1 Introdução

O setor agrícola brasileiro chegou a 18,4% do PIB no ano de 2020, segundo o Centro de Estudos em Economia Aplicada (CEPEA) [23]. Em 2021, houve um crescimento de 2,4%, elevando este patamar a 20,8% do PIB. Estima-se que a área plantada terá um crescimento de 82 milhões de hectares em 2021 para 93,3 milhões em 2031, de acordo com o Ministério da Agricultura, Pecuária e Abastecimento [37].

A Sociedade Internacional de Meteorologia para a Agricultura [24] mostra que o clima tem grande influência na produção agrícola, na incidência de pestes e doenças, na demanda hídrica e na necessidade do emprego de fertilizantes. A análise de dados

é uma ferramenta útil para mitigar esses efeitos adversos. Nabila, Kechadi e McDonnell [39] trazem em sua pesquisa pontos onde a análise de dados pode atuar na agricultura: no monitoramento do clima, de pestes e ervas daninhas, do rendimento da colheita e da irrigação.

A Open Weather é um instituto de ciência de dados focado em meteorologia e afirma que as temperaturas diárias muito altas ou muito baixas podem prejudicar o curso dos processos bioquímicos nas células e causar a morte das plantas [25]. A taxa de crescimento da planta, a duração reprodutiva e o potencial de rendimento têm uma forte correlação com o monitoramento adequado da temperatura [26].

A mesma preocupação vale para a umidade relativa do ar. Segundo o Ministério da Agricultura da Colúmbia Britânica [27], níveis anormais de umidade podem promover o crescimento de organismos patogênicos. Um outro agravante é que as plantas não podem evaporar a água de suas folhas caso a umidade esteja muito elevada.

Um ponto importante também é ter a garantia de que os sensores empregados no monitoramento estão em bom funcionamento e não estão apresentando valores errôneos de medição. Problemas desta natureza podem refletir num manejo agrícola inadequado.

Por esses motivos, é necessária atenção às variáveis de temperatura e umidade relativa para melhor monitorar eventos climáticos extremos, como demonstrado pela Fapesp [28] nas regiões brasileiras e ao redor do mundo pelo Climate Centre [29].

Sabendo da importância das variáveis meteorológicas, temperatura e umidade relativa na atividade agrícola, este trabalho tem como principal objetivo avaliar, através de experimentação, algoritmos de aprendizado de máquina não supervisionados na detecção de anomalias em séries temporais destas variáveis. Os métodos escolhidos nesta avaliação foram o *Density-Based Spatial Clustering of Applications with Noise (*DBSCAN) e *Isolation Forest* (IF). O DBSCAN funciona agrupando os dados e verificando a densidade desse grupo. O Isolation Forest por sua vez trás uma abordagem que utiliza árvores binárias e isola as anomalias. Estes dois são da classe dos algoritmos não supervisionados, que são os mais indicados para este tipo de cenário, como mostrado no Referencial Teórico.

2 Referencial Teórico

Para Foorthuis [30], no contexto de *data science*, anomalias são ocorrências incomuns em um conjunto de dados, não se encaixando nos padrões gerais. Dasgupta e Forrest [31] chamam anomalias de novidades e as caracterizam como qualquer desvio que exceda uma variação permitida, chamando também de eventos anormais. Em [32] Agarwal e Gupta dizem que um *outlier* é uma observação que está longe do resto dos pontos em um conjunto de dados. Em seu trabalho, um *outlier* também é chamado de “ponto discrepante”. Portanto, anomalia, novidade, ponto discrepante e *outlier* são termos empregados para definir dados com características diferentes do resto dos valores - pontos que não se encaixam nos padrões gerais do conjunto de dados.

As técnicas de identificação de anomalias em conjuntos de dados possui aplicação em diversas áreas e com termos diferentes utilizados. Gidea e Katz [22] implementaram a análise de dados topológicos para identificar quedas que possam indicar crise no mercado financeiro. Essas quedas anormais são conhecidas nas bolsas de valores por *crashes*. Angiulli e Fassetti [12] chamam anomalias de *outliers* para detectá-los dentro dos fenômenos El Niño e La Niña. Em [11], foram utilizadas redes Bayesianas dinâmicas para identificar erros de medição em sensores ambientais. Esses erros são chamados de anomalias. Moschini *et al.* [13] foi utilizado o método ARIMA para identificar fraudes em cartões de crédito. Esses comportamentos anormais nas compras dos usuários são chamados tanto de outliers como de anomalias.

2.1 Aprendizagem de máquina

A aprendizagem de máquina supervisionada é quando os dados já possuem rótulos estabelecidos. O algoritmo precisa saber previamente através de conjuntos de treinamento, os valores corretos para os dados e após o treinamento receber o conjunto de dados de testes para fazer as previsões. Estes algoritmos são indicados para problemas de regressão e classificação.

A aprendizagem não supervisionada é utilizada quando não temos as informações prévias. Problemas de agrupamento, sobretudo para detecção de anomalias [33], são mais indicados a utilizarem algoritmos não supervisionados, pois eles irão agrupar os dados de acordo com características inerentes aos valores sem previamente ensinar ao algoritmo os rótulos deles. O algoritmo deve encontrar relações de proximidade nos dados para poder separá-los em grupos semelhantes.

Como mostrado por Nassif *et al.* [34], algoritmos supervisionados para detecção de anomalias possuem dois problemas: a quantidade de anomalias no conjunto de testes é bem menor do que as instâncias normais e é difícil de rotular um valor como discrepante. Por estas razões, este trabalho irá focar no emprego de algoritmos de aprendizado não supervisionado na detecção de anomalias.

2.2 *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN)

O DBSCAN foi proposto por Ester *et al.* [35] e sua característica baseada em vizinhança pode ser resumida da seguinte maneira:

Definição 1: (Vizinhança de ε de um ponto), dado um conjunto de valores D, a vizinhança de um objeto p com raio ε é chamada de Vizinhança-ε de p é dada por: $V\varepsilon(p) = \{q \text{ em } D \mid dis(p,q) \leq \varepsilon\}$.

Este é um algoritmo de clusterização que tem como objetivo o agrupamento dos dados separando-os em regiões de grupos de pontos em um determinado raio. Com isso, os valores escolhidos e que formam círculos com baixa densidade de pontos são considerados anomalias. Uma característica importante dos algoritmos de agrupamento é a alta adaptabilidade à mudança nos dados e a baixa sobrecarga no modelo, diferentemente dos algoritmos de classificação [14].

Os principais fatores de configuração 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. Vários subgrupos de dados com n *min_samples* são formados com uma área circular (ε) feita ao redor de cada valor. O grupo que possuir um número de amostras menor do que o *min_samples* é considerado anomalia pois difere dos outros que possuem o *min_samples* com a quantidade mínima definida no modelo.

Para melhor entendimento da explicação, vejamos a Figura 1.

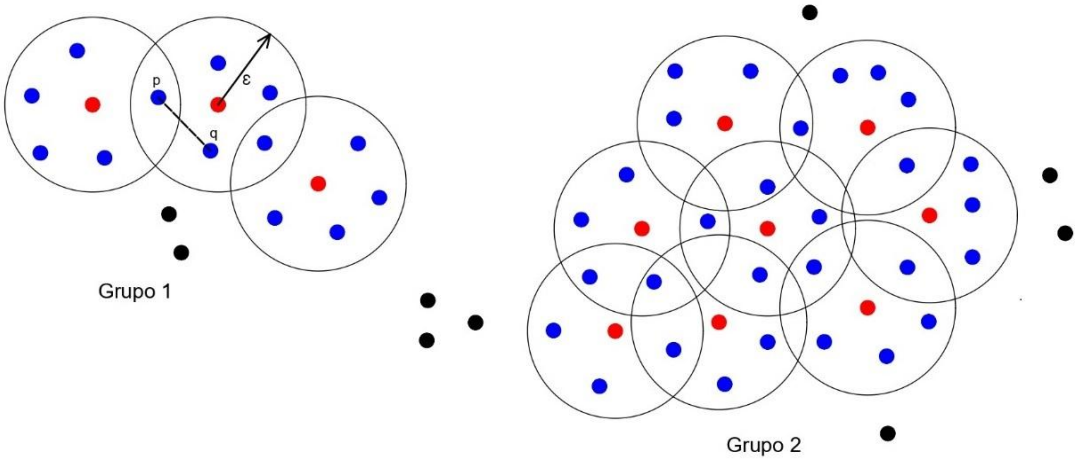


Figura 1 - DBSCAN e dois grupos de dados normais criados. **Fonte:** Autor.

A Figura 1 mostra o DBSCAN criando dois grupos com o *min_samples* = 5. Existem três tipos de pontos: os chamados pontos chave (vermelhos), que são pontos que conseguem alcançar os fatores do modelo, ou seja, satisfazem os critérios para formar o grupo; os pontos chamados de bordas (azuis), que são os pontos que não satisfazem o critério do modelo, mas que são atestados como dados normais porque conseguem estar dentro de um grupo formado pelo ponto vermelho; e os pontos que são as anomalias (preto), que não fazem parte de nenhum grupo [15].

O DBSCAN utiliza duas regras: pontos no raio de pesquisa de um ponto chave fazem parte do grupo deste ponto chave. Os círculos que possuem os mesmos valores de borda fazem parte do mesmo grupo.

2.3 *Isolation Forest*

O ***Isolation Forest*** é um algoritmo de aprendizado não supervisionado baseado em árvores de decisão, em [36]. O conjunto de dados é particionado e 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, provavelmente estão distantes dos outros e são consequentemente classificados como anomalias.

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 de acordo com o fator *max_samples*. A contaminação (*contamination*) é a porcentagem aproximada de quantas anomalias existem no conjunto de dados.

Cada ponto é classificado com uma nota entre 0 e 1. Aqueles considerados como normais são pontuados com um número abaixo de 0,5 e os discrepantes são pontuados com notas mais próximas de 1, ou seja, anomalias. Baseado na

quantidade aproximada de anomalias (*contamination*), o algoritmo irá pontuar a quantidade de dados condizente com a contaminação aproximada. A Figura 2 ilustra através de um grafo como a caracterização do IF funciona.

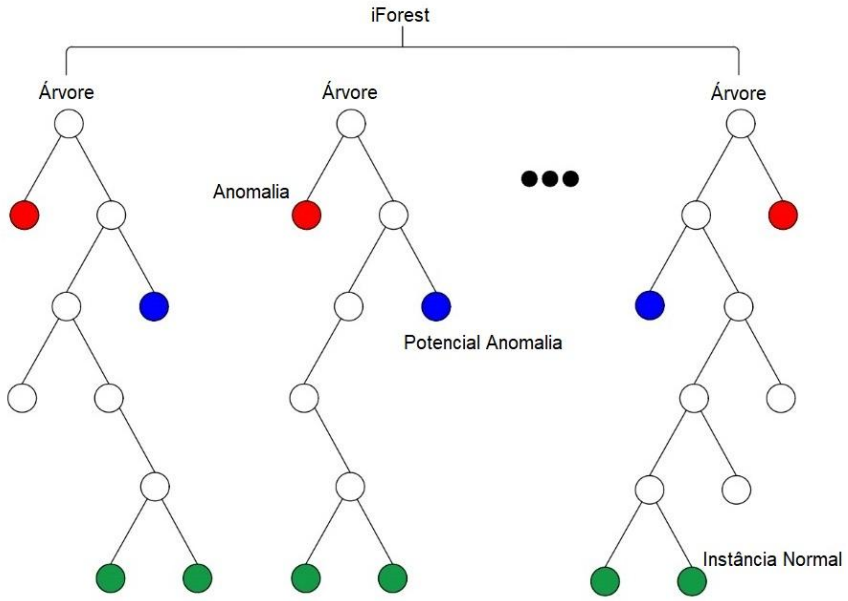


Figura 2 - Isolation Forest através de árvores. **Fonte:** Autor.

A pontuação das anomalias é dada por uma fórmula de definição do *score* como definido na Equação 1.

$$s(x) = 2^{-\frac{E(h(x))}{c(n)}}, \quad (1)$$

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n}, \quad (2)$$

$$H(i) = ln(i) + 0.5772156649, \quad (3)$$

- Onde *E(h(x))* é a média da profundidade que o valor x possui em todas as árvores construídas.
- *c(n)*, Equação 2, é o comprimento médio do caminho de uma pesquisa mal sucedida na árvore de pesquisa binária (BST) em um conjunto de tamanho n [36].
- $H(i)$, Equação 3, é o número harmônico e pode ser estimado por *ln(i)* + 0,5772156649 (constante de Euler) [36], em que i é o nível da árvore.

Como as árvores do IF têm uma estrutura equivalente à BST, a estimativa da média *h(x)* é a mesma que a pesquisa mal sucedida na BST.

3 Trabalhos relacionados

De acordo com Gupta *et al.* [10], a detecção de anomalias é uma área da mineração de dados que têm sido amplamente estudada nos últimos anos. O crescimento acelerado do número de dispositivos de hardware, capazes de coletar

dados diversos, proporcionou um aumento expressivo na quantidade e na qualidade de dados disponíveis para análise. Dados do mercado financeiro, biológicos, climáticos, sistemas de diagnósticos, por exemplo, podem ser minerados para auxiliar na detecção de anomalias.

Dasgupta e Forrest [31] apresentam um método inspirado no sistema imunológico utilizando algoritmo baseado no princípio de seleção negativa para detectar anomalias dentro de janelas deslizantes em séries temporais variadas. Um dos conjuntos de dados foi gerado a partir da equação de Mackey-Glass. Um segundo conjunto veio de sensores de máquinas de moagem para detectar falhas em equipamentos para prevenir possíveis quebras. Em [1] mostra como a detecção de anomalias pode ser importante ao demonstrar em sua pesquisa dados anômalos em eletrocardiogramas, que podem indicar contração ventricular prematura. Em [16] o Isolation Forest é utilizado para detectar anomalias em nuvens de pontos de imagens. Esta técnica permitiu a limpeza das imagens 3D geradas eliminando pontos estranhos dentro da nuvem.

Arvor *et al.* [2] 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. Em Yuxiang *et al.* [9] são utilizados dados climáticos espaço-temporais do Sul da China para detectar valores discrepantes em certas épocas do ano em diferentes áreas.

Zemicheal e Dietterich [40] pesquisaram como controlar a qualidade de valores climáticos vindos de sensores que estivessem quebrados, temporariamente ou permanentemente removidos. Cinco métodos (Imputação Média, Imputação MAP, Redução, Marginalização e Distribuição Proporcional) foram comparados entre si e para confirmar sua eficácia utilizaram alguns algoritmos de detecção de anomalias: Isolation Forest, LODA e EGMM.

Wibisono *et al.* [41] foi utilizado 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. Seu trabalho utilizou dados de Seramang, província de Java Central, na Indonésia. Antes da execução do algoritmo os dados foram normalizados para que nenhum parâmetro fosse dominante devido a diferença nas escalas.

Celik, Dadaser e Dokuz [7] utilizou-se 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. O DBSCAN apresentou melhores resultados pois as anomalias não são apenas pontos extremos, que se descolam muito da média e desvio padrão, mas também podem ser pontos bem próximos de um valor não anômalo.

4 Materiais e Métodos

Para realização do experimento utilizou-se a série temporal da **temperatura máxima** e da **umidade média relativa do ar**. A escolha dessas variáveis se deu pelo tamanho da sua importância dentro do cenário agrícola.

O Ministério da Agricultura da Colúmbia Britânica [27] mostra que umidade relativa do ar em números anormais promove crescimento de patógenos nas plantas e

atrapalha a evaporação de água em suas folhas. A Open Weather [25] mostra que temperaturas extremas ou abaixo do normal prejudicam os processos bioquímicos dos vegetais, sendo também importante no desenvolvimento e reprodução deles [26].

Esses são dados diários no intervalo de tempo entre 01/01/2020 e 31/12/2021, obtidos da estação A307 do Instituto Nacional de Meteorologia (INMET) [38], localizada no município de Petrolina, sertão de Pernambuco, com latitude de -9.388323° e longitude de -40.523262°.

Para os algoritmos serem avaliados iremos incluir uma quantidade de anomalias na série temporal empregando o método utilizado por B. Sabyasachi [20] e Y. Lu *et al.* [3]. Nesta fase, foi retirada a sazonalidade da série temporal para que ela ficasse livre de tendências. Após isso, retirou-se uma amostra aleatória de 5% de dados igualmente distribuídos dentro da série temporal. Nesse subconjunto, aplicou-se a fórmula de [20] para geração das anomalias e depois inseridos novamente nas mesmas posições de onde foram retirados.

Dessa forma, obteve-se um conjunto de dados com valores comuns e valores discrepantes permitindo a análise comparativa entre os algoritmos DBSCAN e IF. Abordagem semelhante foi adotada por Thang e Kim em [5].

Os modelos foram implementados na linguagem Python e utilizaram a biblioteca *scikit-learn.* Os fatores utilizados nos modelos estão dispostos na Tabela 1 e estão baseados nas propostas de [7, 8, 36]. Foram escolhidos níveis baseados nessa literatura e também a opção em que fosse possível não passar valor algum, e o próprio algoritmo utilizar os valores padrão do modelo, isso porque em [19] e [8] é mostrado que pode ser difícil encontrar as métricas ótimas para esses algoritmos, então uma opção é utilizar os valores padrão do próprio modelo.

Por exemplo, se nenhum eps for atribuído ao DBSCAN, ele irá considerar o valor padrão de 0.5. O mesmo vale para a *contamination* do Isolation Forest, em que se o valor ‘auto’ for escolhido, o próprio algoritmo vai tentar descobrir qual a proporção aproximada de anomalias que o conjunto de dados possui.

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 execução de cada algoritmo. **Fonte:** Autor

O conjunto de dados foi organizado de modo que fossem inseridas anomalias para que a precisão dos algoritmos pudesse ser testada. Podemos ver o conjunto de dados a partir da Figura 3.

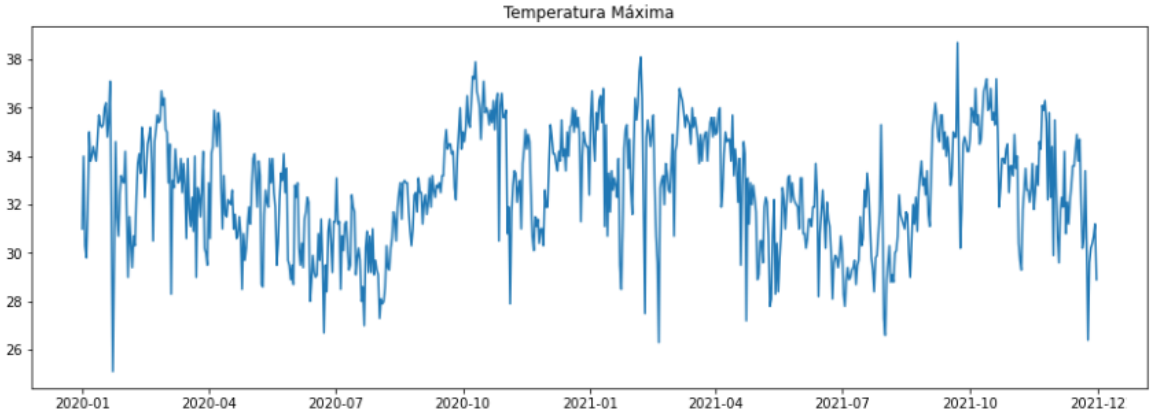


Figura 3 - Série temporal original da temperatura. **Fonte:** Autor.

O método de geração dos anomalias que serão inseridos na série temporal foi semelhante ao utilizado em [20], onde utilizou-se um método de auto regressão com lag 1 chamado AutoReg, da biblioteca *scikit-learn,* para remover a sazonalidade e a série temporal ficar limpa de tendências.

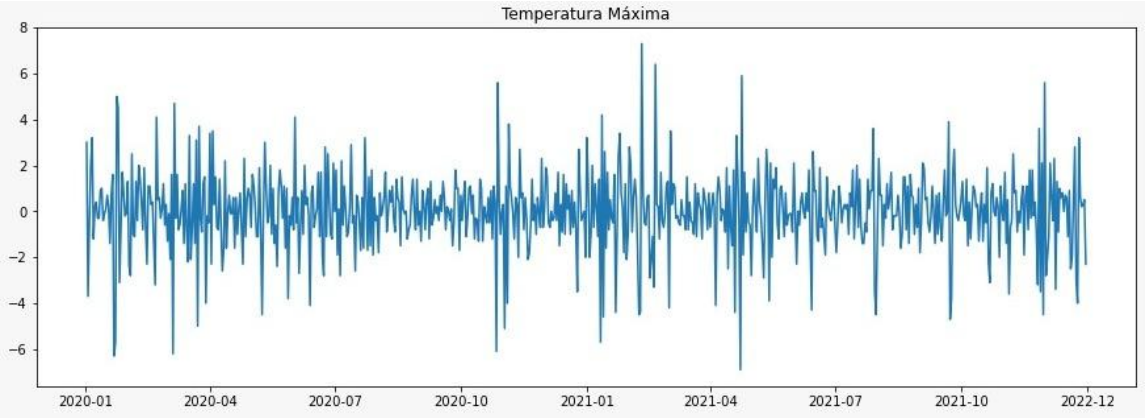


Figura 4 - Série temporal de temperatura sem sazonalidade. **Fonte:** Autor.

Segue a fórmula utilizada para geração das anomalias :

$$Y = Y + \text{sign}(Y)\, X\, \sigma_{\epsilon}, \tag{4}$$

$$X = \exp(3 \times \text{abs}(x_1)) + 3, \tag{5}$$

Na Equação 4, o desvio padrão da série temporal até o instante do valor (temperatura ou umidade) é calculado e em seguida multiplicado pelo valor referente

a Equação 5, em que x_1 é um número gerado aleatoriamente entre -0.5 e 0.5. Esta multiplicação recebe o *sign*[1] do valor da série temporal, que altera o valor gerado pela Equação 5 para positivo ou negativo, originando o dado anômalo.

O resultado da implementação da fórmula anterior, com as anomalias inseridas no conjunto de dados pode ser visto na Figura 4.

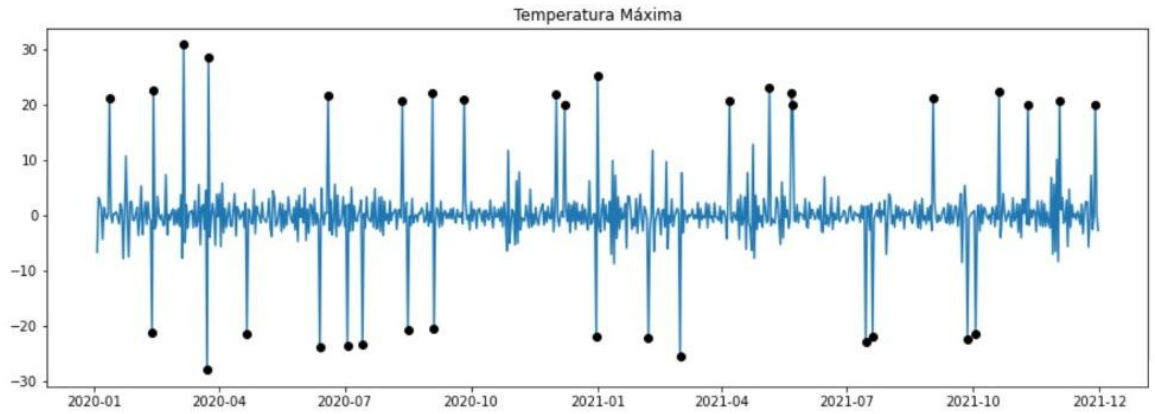


Figura 5 - Série temporal com anomalias inseridas. **Fonte:** Autor.

Para a comparação dos modelos avaliados neste trabalho utilizou-se as mesmas métricas utilizadas em [17], revocação, precisão, acurácia, e pontuação F1. Nossos dados serão classificados em anomalia e não anomalia (valor normal). Uma anomalia fará parte da classe Positiva (P) e um valor normal da classe Negativa (N). Uma vez detectada, o resultado pode ser Verdadeiro (T), caso acertou em classificar se é uma valor anômalo ou não, ou Falso (F), em caso de erro em classificar. Temos as métricas calculadas como:

$$Revocação = \frac{TP}{TP+FN}, \tag{6}$$

$$Precisão = \frac{TP}{TP+FP}, \tag{7}$$

$$Acurácia = \frac{TP+TN}{TP+TN+FP+FN}, \tag{8}$$

$$Pontuação\ F1 = \frac{2TP}{2TP+FP+FN}, \tag{9}$$

5 Resultados e Discussões

Os gráficos gerados mostram as anomalias representadas como pontos em preto. Ambos os algoritmos conseguiram atingir em seus melhores resultados os valores acima de 90% nas quatro métricas calculadas. Nas figuras geradas, os pontos em preto são anomalias e a execução do algoritmo considera que elas estão presentes na temperatura máxima e na umidade média relativa do ar.

[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.

A Acurácia foi a única métrica que teve desempenho de 90% em todos os testes. Esse resultado pode levar a conclusões precipitadas, por isso a importância de analisar mais de uma métrica de desempenho.

5.1 DBSCAN

A Tabela 2 apresenta os resultados obtidos em todas as métricas estabelecidas. O algoritmo conseguiu ter uma taxa de precisão de 93%, acurácia de 97% e F1 de 94% utilizando um ε de 0.5 de raio e um número mínimo de amostras de 8 pontos, Figura 6 (e). Todas as anomalias foram detectadas, o que fez a Revocação ser de 100%, que foi o seu melhor resultado.

O mesmo bom desempenho não acontece com o ε de 0.3 e número mínimo de amostras de 8, Figura 6 (b), onde a taxa de precisão caiu para 55% e a F1 de 71%, mesmo que acurácia tenha sido de 96%. Ou seja, um grande número de pontos de valores normais (classe negativa) foram considerados como anomalias, resultando numa elevada taxa de Falsos Positivos.

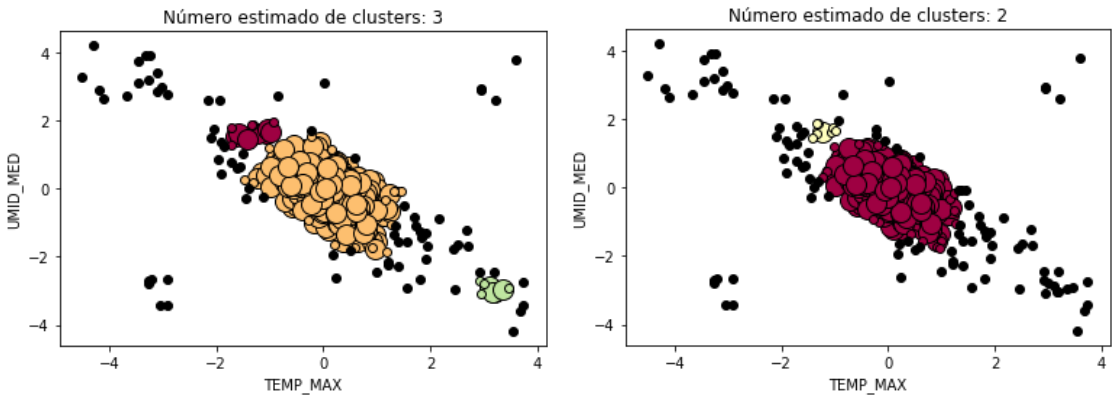
Quando foram utilizados os fatores de ε = 0.3 e número mínimo de amostras 10, Figura 6 (c), a precisão foi ainda menor, apenas 44%. Isso mostra que estes fatores estão sendo menos tolerantes no que deve ser uma anomalia, ou seja, cada círculo deve ser formado por 10 amostras, que é um número maior, e 0.3 de raio, que é uma área menor, é difícil atender a esses requisitos, visto que é um número maior de amostras dentro de um círculo com raio menor. Uma elevada quantidade de valores normais foram considerados como anomalias porque o algoritmo não conseguiu formar mais círculos com esses valores. Isso pode ser visto, como citado, pela precisão de 44% e a F1 de 61%.

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 2 - Resultados obtidos pelo DBSCAN. **Fonte:** Autor

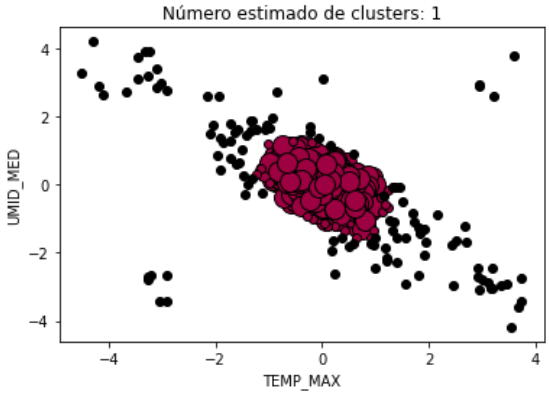
Podemos observar na Figura 6 que o DBSCAN forma o que se chama de grupo. Quando os subconjuntos de dados normais são formados escolhemos plotá-los em cores diferentes, destacando os grupos, assim como mostrado na Figura 1. Porém, à medida que os fatores vão sendo alterados estes grupos podem ser eliminados pois seus dados não satisfazem mais aos requisitos de ε e número mínimo de amostras.

Todos os outros exemplos possuem mais de um grupo. Apenas a Figura 6 (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, que obteve bons resultados (precisão=85%, F1=92%), formou três grupos nos valores não anômalos.

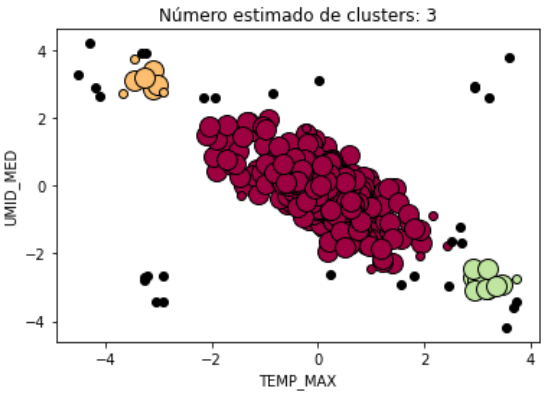


(a) min_samples=5 e eps=0.3

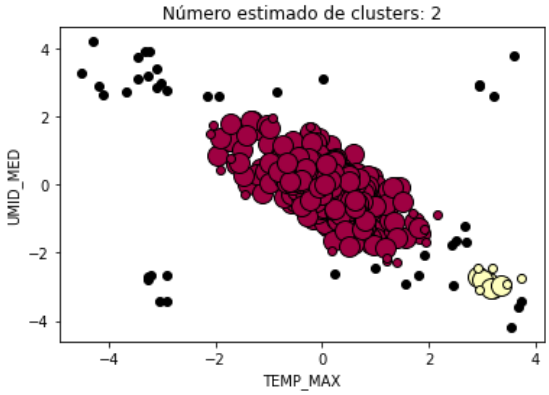
(b) min_samples=8 e eps=0.3



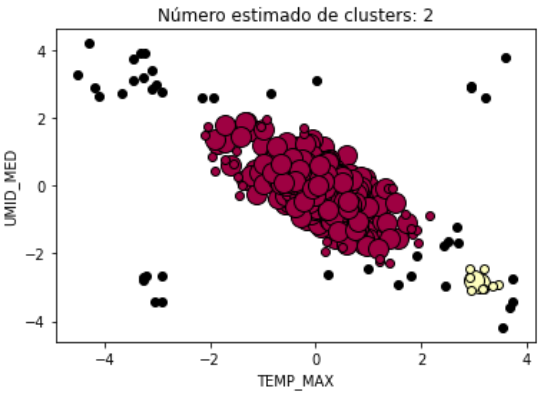
(c) min_samples=10 e eps=0.3



(d) min_samples=5 e eps=0.5



(e) min_samples=8 e eps=0.5



(f) min_samples=10 e eps=0.5

Figura 6 - Resultados gráficos do DBSCAN. **Fonte:** Autor

5.2 Isolation Forest

Na Figura 7 podemos ver que o algoritmo classifica como anomalia os pontos em preto e como dados normais os pontos em vermelho. Ao contrário do que acontece com o DBSCAN, o IF não cria grupos bem definidos nos valores normais, invés disso, são criadas diversas árvores que servem para isolar os valores anômalos.

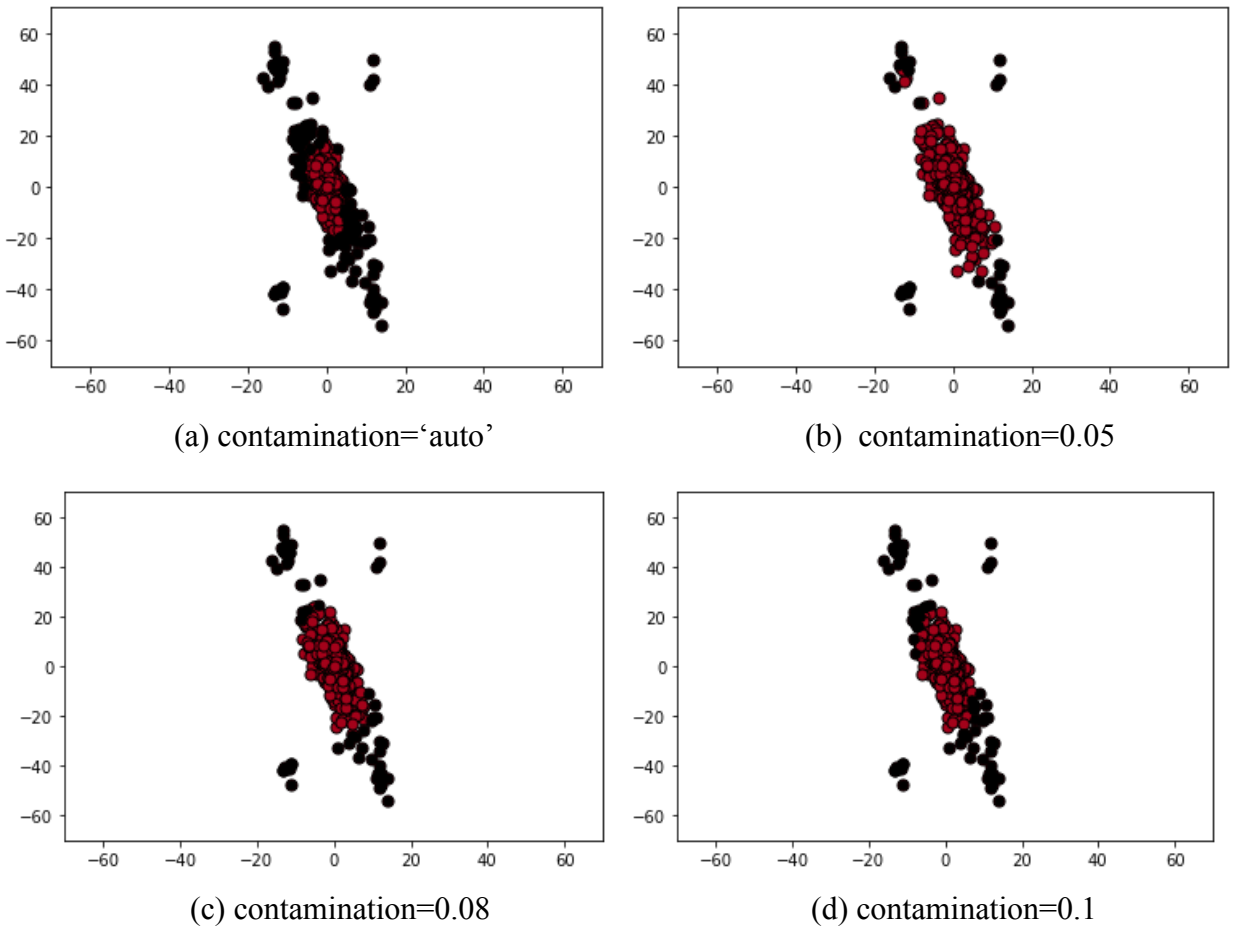


Figura 7 - Resultados gráficos do Isolation Forest. **Fonte:** Autor

Em todos os casos utilizamos 100 unidades de árvores criadas, pois como dito em [16] é um número já considerado adequado de testagem do próprio algoritmo. O número de amostras utilizadas ficaram com o tamanho da própria série temporal e o fator *contamination* foi o determinante para uma melhora ou piora no desempenho do algoritmo.

Os resultados do IF podem ser visualizados na Tabela 3. 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.

Em todos os casos podemos notar que a Acurácia foi acima de 90%. Porém, sabendo que não podemos nos basear isoladamente em uma única métrica, vemos que alguns valores mesmo com a acurácia alta a precisão está baixa devido ao alto número de Falsos Positivos (valores que não são mas que foram considerados como anomalias).

Um exemplo disso é a contaminação determinada como 'auto', Figura 7 (a), que atingiu uma acurácia de 91%, mas uma precisão de 36%, isso porque houve um alto índice de não anomalias sendo consideradas como anomalias. Isso fica mais evidente visualmente na diferença das anomalias detectadas com a *contamination*='auto' e as outras imagens tiveram bem menos preenchimento de anomalias.

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 3 - Resultados obtidos pelo Isolation Forest. **Fonte:** Autor

O parâmetro *contamination* faz com que mesmo a série temporal tendo 5% de seus dados como anomalias, o *contamination* com níveis de 8% e 10% faz o IF em sua execução busca por um valor aproximado de anomalias a esses escolhidos. Dessa forma, o algoritmo estava procurando por 8% e 10% de anomalias na série temporal. Isso gerou erro de Falsos Positivos, pois mesmo não havendo essa proporção de anomalias o IF tenta encontrar essa quantidade. Esse comportamento pode ser percebido através das Figuras 7 (c) e (d), que gerou uma alta acurácia, detectou todas anomalias, mas a Precisão e F1 mostram a quantidade de Falsos Positivos gerados.

Os resultados gerados pelo *contamination* de 5%, Figura 7 (b) foram os melhores do IF. Isso mostra que se for conhecido aproximadamente quanto de anomalia tem o conjunto de dados, o Isolation Forest pode ter resultados com precisão, F1 e Revocação acima de 90%.

5 Conclusões

Com a abordagem de pesquisa conseguimos trazer uma visão de como o DBSCAN e o Isolation Forest funcionam na detecção de anomalias em um contexto de meteorologia.

O Isolation Forest foi o que demonstrou melhor precisão e acurácia para encontrar os pontos estranhos. Por mais alta que possa ser as métricas marcadas em negrito no DBSCAN e IF, pode ser difícil encontrar os fatores a serem utilizados em cada um deles. Por isso, nossa metodologia utilizou diversos parâmetros para conseguirmos encontrar o melhor modelo para ambos os algoritmos.

Um ponto positivo do Isolation Forest foi que a taxa de Falsos Negativos foi de zero. Então, pode-se afirmar que para este cenário de temperatura máxima e umidade média relativa do ar, o IF possui uma boa vantagem para casos em que não se pode haver tolerância para este tipo de erro onde uma anomalia é considerada um valor normal. Pode-se tomar como exemplo em [1], onde é feita a detecção de pacientes com arritmias. É preferível que hajam erros de Falsos Positivos, onde pacientes sem arritmias sejam considerados portadores dessa condição, ao algoritmo dizer que um paciente com arritmia seja uma pessoa saudável. A segunda opção traz consequências piores. Em nosso caso, é melhor o algoritmo dizer que um sensor está com defeito e se descobrir que não está, a não informar que ele está com defeito e continuar reportando informações incorretas ao produtor.

A característica do DBSCAN em analisar a quantidade de vizinhos dentro de uma área definida no modelo em comparação às árvores criadas pelo Isolation

Forest, parece produzir resultados inferiores ao IF, devido à natureza de análise mais individual de cada ponto, onde eles são escolhidos aleatoriamente e os valores localizados em menores níveis podem ser acusados como anomalias por estarem em um ponto mais isolado de outros. Na Tabela 4 podemos ver um comparativo entre os dois.

Ambos os algoritmos possuem os fatores com valores padrão, porém o IF não obteve bons resultados. O Isolation Forest teve boa performance à medida que a taxa da contaminação do modelo foi ficando mais próxima com a contaminação real que inserimos inicialmente na série, visto que possuímos muitas anomalias bem no limite de um valor real, o que fez com que usássemos os exemplos semelhantes a outras pesquisas [7, 8, 14, 16] que utilizaram os mesmos algoritmos.

O DBSCAN mostrou boa taxa de acertos sem escolha dos fatores, ou seja, sem customização do modelo. Isso pode mostrar que o DBSCAN pode ter boa performance para um uso mais simples se não quisermos utilizar o método para encontrar seus parâmetros implementado em [5].

Algoritmo	Precisão	Acurácia	Revocação	F1
DBSCAN	0.93	0.97	1	0.94
IF	0.94	0.98	1	0.97

Tabela 4 - Melhor resultado do DBSCAN e do Isolation Forest. **Fonte:** Autor

Estima-se que o setor agrícola apresentará um grande crescimento nos próximos 10 anos, tendo hoje 95% dos produtores com interesse em utilizar alguma tecnologia digital para seu auxílio e aproximadamente 60% já os utiliza na lavoura, segundo a Embrapa [4]. Dado este cenário de crescimento, este trabalho apresenta uma contribuição relevante no processo de transformação digital da agricultura brasileira.

6 Trabalhos Futuros

Este trabalho teve algoritmos de aprendizado de máquina não supervisionado como sua base de estudo. Sabendo do seu desempenho através do DBSCAN e do IF, é interessante fazermos um estudo de uma técnica chamada de Análise de Dados Topológicos, que faz uso da topologia algébrica para encontrar certas estruturas nos dados, como a sua forma e conectividade, trazendo os dados mais para dentro dos contextos deles para uma melhor interpretação [18]. Seria interessante comparar os dois métodos já que a análise de dados topológicos também é robusta trabalhando em abordagens semelhantes às utilizadas com aprendizado de máquina tradicional, como encontrar comportamentos extremos em bolsas de valores [22] e dados meteorológicos, como fez Santos *et al.* [21].

7 Agradecimentos

Ao DEINFO UFRPE, que me proporcionou este bacharel. Ao Professor Dr. Victor Medeiros, por me orientar a redigir um projeto no formato de artigo científico. Ao Professor Dr. Glauco Gonçalves, que me trouxe ao mundo da pesquisa

acadêmica. Ao futuro intitulado mestre Wellington Antônio, que me ajudou no desenvolvimento prático inicial deste projeto.

Referências

[1] E. Keogh, J. Lin, and A. Fu, “HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence,” in Proc. of the 5 th IEEE Intl. Conf. on Data Mining (ICDM), 2005, pp. 226–233.

[2] Arvor, Damien & Jonathan, Milton & Simoes, Margareth & Dubreuil, Vincent. (2008). Detecting outliers and asserting consistency in agriculture ground truth information by using temporal VI data from MODIS. XXI Congress “International Society for Photogrammetry and Remote Sensing”, Beijing.

[3] Y. Lu, J. Kumar, N. Collier, B. Krishna and M. A. Langston, "Detecting Outliers in Streaming Time Series Data from ARM Distributed Sensors," 2018 IEEE International Conference on Data Mining Workshops (ICDMW), 2018, pp. 2, doi: 10.1109/ICDMW.2018.00117.

[4] Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), 10/08/2020, “Pesquisa mostra o retrato da agricultura digital brasileira”, <https://www.embrapa.br/busca-de-noticias/-/noticia/54770717/pesquisa-mostra-o-retrato-da-agricultura-digital-brasileira>.

[5] T. M. Thang and J. Kim, "The Anomaly Detection by Using DBSCAN Clustering with Multiple Parameters," *2011 International Conference on Information Science and Applications*, 2011, pp. 4, doi: 10.1109/ICISA.2011.5772437

[6] https://portal.inmet.gov.br/

[7] Celik, Mete & Dadaser-Celik, Filiz & Dokuz, Ahmet. (2011). Anomaly Detection in Temperature Data Using DBSCAN Algorithm. INISTA 2011 - 2011 International Symposium on Innovations in Intelligent SysTems and Applications. 10.1109/INISTA.2011.5946052.

[8] Chesnokov, M.. (2019). Time Series Anomaly Searching Based on DBSCAN Ensembles. Scientific and Technical Information Processing. 46. 299-305. 10.3103/S0147688219050010.

[9] Sun Yuxiang, Xie Kunqing, Ma Xiujun, Jin Xingxing, Pu Wen and Gao Xiaoping, "Detecting spatio-temporal outliers in climate dataset: a method study," *Proceedings. 2005 IEEE International Geoscience and Remote Sensing Symposium, 2005. IGARSS '05.*, 2005, pp. 4 pp.-, doi: 10.1109/IGARSS.2005.1525218.

[10] M. Gupta, J. Gao, C. C. Aggarwal and J. Han, "Outlier Detection for Temporal Data: A Survey," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 9, pp. 2250-2267, Sept. 2014, doi: 10.1109/TKDE.2013.184.

[11] D. J. Hill, B. S. Minsker, and E. Amir, “Real-time Bayesian anomaly detection for environmental sensor data,” in Proc. 32nd Conf. IAHR, 2007.

[12] Angiulli, Fabrizio & Fassetti, Fabio. (2007). Detecting distance-based outliers in streams of data. International Conference on Information and Knowledge Management, Proceedings. 811-820. 10.1145/1321440.1321552.

[13] Moschini, Giulia, R'egis Houssou, J'erome Bovay and Stephan Robert-Nicoud. “Anomaly and Fraud Detection in Credit Card Transactions Using the ARIMA Model.” *ArXiv* abs/2009.07578 (2021).

[14] D. Deng, "DBSCAN Clustering Algorithm Based on Density," *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, 2020, pp. 949-953, doi: 10.1109/IFEEA51475.2020.00199.

[15] Difrancesco, Paul-Mark & Bonneau, David & Hutchinson, D.. (2020). The Implications of M3C2 Projection Diameter on 3D Semi-Automated Rockfall Extraction from Sequential Terrestrial Laser Scanning Point Clouds. Remote Sensing. 12. 1885. 10.3390/rs12111885.

[16] Regaya, Yousra & Fadli, Fodil & Amira, Abbes. (2021). Point-Denoise: Unsupervised outlier detection for 3D point clouds enhancement. Multimedia Tools and Applications. 80. 1-17. 10.1007/s11042-021-10924-x.

[17] Luo, Hao & Jia, Shuli & Zhang, Wenxuan. (2019). Hierarchical Temporal Memory Based Anomaly Detection for Hydrological Monitoring of Unmanned Surface Vehicle. 420-424. 10.1109/ICICSP48821.2019.8958534.

[18] Wasserman, Larry. (2016). Topological Data Analysis. Annual Review of Statistics and Its Application. 5. 10.1146/annurev-statistics-031017-100045.

[19] Karami, Amin & Johansson, Ronnie. (2014). Choosing DBSCAN Parameters Automatically using Differential Evolution. Int. J. Comput. Appl.. 91. 10.5120/15890-5059.

[20] B. Sabyasachi, M. Martin, “Automatic outlier detection for time series: An application to sensor data” Knowl. Inf. Syst.. 11. pp. 3, 10.1007/s10115-006-0026-6.

[21] Santos, M. F.; Amorim, M.; De Oliveira, Wilson; Stosic, T.. Análise Topológica De Dados Para Caracterização De Periodicidade Em Séries Temporais De Dados Pluviométricos. Revista Mundi Engenharia, Tecnologia e Gestão (ISSN: 2525-4782), v. 4, p. 150-1-150-12, 2019.

[22] Gidea, Marian & Katz, Yuri. (2017). Topological Data Analysis of Financial Time Series: Landscapes of Crashes. SSRN Electronic Journal. 10.2139/ssrn.2931836.

[23] Centro de Estudos Avançados em Economia Aplicada, Metodologia - PIB do Agronegócio, 2022, <https://www.cepea.esalq.usp.br/br/metodologia-pib-do-agronegocio-brasileiro.aspx>.

[24] International Society for Agricultural Meteorology, 11/10/2016, “Weather and Climate Forecasts for Agriculture” <http://www.agrometeorology.org/files-folder/ repository/ gamp_chapt5.pdf>.

[25] Open Weather, The influence of temperature on plant productivity in agriculture: Accumulated temperature, 28/09/2017, <https://openweather.co.uk/blog/post/influence-temperature-plant-productivity-agriculture-accumulated-temperature>.

[26] Jerry L. Hatfield, John H. Prueger, Temperature extremes: Effect on plant growth and development, Weather and Climate Extremes, Volume 10, Part A, 2015, Pages 4-10, ISSN 2212-0947, https://doi.org/10.1016/j.wace.2015.08.001.

[27] Ministry of Agriculture of British Columbia, Understanding Humidity Control in Greenhouses, 09/2015, <https://www2.gov.bc.ca/assets/gov/farming-natural-resources-and-industry/agriculture-and-seafood/animal-and-crops/crop-production/understanding_humidity_control.pdf>.

[28] Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), 08/2012, “Eventos Climáticos Extremos no Brasil - Impactos, Ciência e Políticas Públicas”, <https://fapesp.br/eventos/2012/08/IPCC/Rittl.pdf>.

[29] Climate Centre, "Abnormal is the new normal". May sets global climate records, 15/06/2016, <https://www.climatecentre.org/1601/abnormal-is-the-new-normal-may-sets-global-climate-records/>.

[30] Foorthuis, R. On the nature and types of anomalies: a review of deviations in data. *Int J Data Sci Anal* 12, 297–331 (2021). https://doi.org/10.1007/s41060-021-00265-1.

[31] Dasgupta, D., & Forrest, S. (1996). Novelty detection in time series data using ideas from immunology.

[32] Agarwal, Amulya & Gupta, Nitin “Comparison of Outlier Detection Techniques for Structured Data” *ArXiv* abs/2106.08779 (2021).

[33] IBM, “Supervised vs. Unsupervised Learning: What’s the Difference?”, 12/03/2021
<https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning>

[34] A. B. Nassif, M. A. Talib, Q. Nasir and F. M. Dakalbab, "Machine Learning for Anomaly Detection: A Systematic Review," in *IEEE Access*, vol. 9, pp. 78658-78700, 2021, doi: 10.1109/ACCESS.2021.3083060.

[35] Ester, M.; Kriegel, H.; Sander, J.; Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp. 226–231.

[36] Liu, Fei Tony & Ting, Kai & Zhou, Zhi-Hua. (2009). Isolation Forest. 413 - 422. 10.1109/ICDM.2008.17.

[37] Ministério da Agricultura, Pecuária e Abastecimento, 30/08/2021, “Projeções do Agronegócio 2020-2021 a 2030-2031”, <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>

[38] https://portal.inmet.gov.br/manual/manual-de-uso-da-api-esta%C3%A7%C3%B5es.

[39] Chergui, Nabila & Kechadi, M.-Tahar & McDonnell, Michael. (2020). The Impact of Data Analytics in Digital Agriculture: A Review. 1-13. 10.1109/OCTA49274.2020.9151851.

[40] Zemicheal, Tadesse & Dietterich, Thomas. (2019). Anomaly detection in the presence of missing values for weather data quality control. 65-73. 10.1145/3314344.3332490.

[41] Wibisono, S & Anwar, M & Supriyanto, Aji & Amin, I. (2021). Multivariate weather anomaly detection using DBSCAN clustering algorithm. Journal of Physics: Conference Series. 1869. 012077. 10.1088/1742-6596/1869/1/012077.