Predição de descargas atmosféricas utilizando Machine Learning para prevenção de acidentes

Marcos A. Alves [***,****] **Bruno A. S. Oliveira** [***,****]
Fernando P. Silvestrow [***] **Luiz F. M. Rodrigues** [***]
Eugenio L. Daher [***] **Willian Maia** [*,****] **Waterson S. Soares** [*,****]
Ana Paula P. Santos [**] **Douglas B. S. Ferreira** [**] **Osmar Pinto Júnior** [†]

[*] *Vale SA (e-mail: willian.maia, waterson.soares@vale.com)*
[**] *Instituto Tecnológico Vale - ITV, (e-mail: douglas.silva.ferreira, ana.santos@pq.itv.org)*
[***] *Fundação para Inovações Tecnológicas - FITec (e-mail: marcosaalves, brunooliveira, fsilvestrow, lfrodrigues, edaher@fitec.org.br)*
[****] *Programa de Pós-Graduação em Engenharia Elétrica - Universidade Federal de Minas Gerais - Belo Horizonte, MG, Brasil*
[†] *Grupo de Eletricidade Atmosférica (ELAT) do Instituto Nacional de Pesquisas Espaciais (INPE)(e-mail: osmar.pinto@inpe.br)*

Abstract:
The occurrence of atmospheric discharges poses risks to the the company operations and workers in open-air activities. Due to this, this paper aimed to cluster lightning data, simulating real-time monitoring of storms for three different target regions. In addition, storms information were used to predict, 15-minutes earlier, the probability of a lightning strikes these areas. Using a multi-source database from ELAT/INPE, different clusterization methods were evaluated in terms of the Calinski Harabasz, Davies Bouldin and Silhouette metrics. Overall, the best one was the MeanShift which cluster the data in 3-5 storms. Number of storms, density and distance were used into a classification machine learning model to generate warning alerts. The Extreme Gradient Boosting and Support Vector Machines achieved the best results in terms of precision and recall, important metrics to evaluate true and false alerts in this context. Both the false alerts, which implies in inactivity of operations and failure rate were equal to or lower than 40%.

Resumo:
A ocorrência de descargas atmosféricas gera riscos às operações e aos profissionais que atuam a céu aberto. Por isso, este trabalho simulou o monitoramento em tempo real de três regiões, utilizando informações das tempestades para prever, 15 minutos antes, a probabilidade de uma descarga atingir os alvos. Utilizando uma base de múltiplas fontes do ELAT/INPE, diferentes métodos de agrupamento foram avaliados observando as métricas Calinski Harabasz, Davies Bouldin e Silhueta. O melhor, no geral, foi o MeanShift que agrupou as descargas em $3-5$ tempestades. O número de tempestades, intensidade e distância foram utilizados para construir um modelo de classificação para gerar alertas de advertência. Extreme Gradient Boosting e Máquinas de Vetores de Suporte apresentaram melhores resultados em precisão e revocação. Tanto a taxa de falsos alertas que implica na inatividade de operações, quanto a de falhas que indica a ausência de um alertas preventivos foram iguais ou inferiores a 40%.

Keywords: Atmospheric discharges; Storms; Lightning warning; Machine learning; Mining; Clustering; Classification; Safety engineering.

Palavras-chaves: Descargas atmosféricas; Tempestades; Alertas de raios; Aprendizado de máquina; Mineração; Clusterização; Classificação; Engenharia de segurança.

1. INTRODUÇÃO

Descargas atmosféricas são belas, porém perigosas. Dado seu alto poder de destruição, elas podem trazer riscos operacionais como queima de equipamentos eletrônicos, avarias em torres de telecomunicações e em edifícios, desligamentos de linhas de transmissão e distribuição de energia, além de fatalidades com pessoas que atuam nestas regiões (Santos et al., 2018). Especialmente no Brasil, país com maior número de descargas atmosféricas no mundo (Pinto Jr e Pinto, 2021), o setor privado e a comunidade cientifica têm buscado soluções para mitigar potenciais riscos em locais como no caso da mineração, aviação, parques eólicos, agricultura e, também, em redes aéreas (Srivastava et al., 2015; Mostajabi et al., 2019).

Monitorar tempestades e gerar alertas de prevenção de descargas atmosféricas têm sido um grande desafio. As fontes de dados são, normalmente, provenientes de sensores de campo elétrico, redes de detecção com sensores em superfície, satélites e/ou previsão meteorológica (Geng et al., 2021). Já as soluções existentes, conforme reportadas na literatura, são variadas. Entre elas, destacam-se aquelas focadas em limiares de campo elétrico (Ferro et al., 2011; Igarashi et al., 2011; Costa et al., 2014; Srivastava et al., 2015), identificação e mapeamento espaço-temporal das tempestades (Juntian et al., 2011) e/ou aprendizado de máquina, ou *machine learning* (ML) (Mostajabi et al., 2019; Bala et al., 2017; Tervo et al., 2021; Geng et al., 2021; Santos et al., 2017). No campo da otimização, Soares e Fonseca (2019); Libório et al. (2020) propuseram uma abordagem para geração de alertas de predição de descargas atmosféricas utilizando medição de campo elétrico pontual, associado a dados de satélites, previsão meteorológica e redes de sensores de superfície aplicado ao ambiente aberto de mineração.

Um consenso existente é que a geração de alertas preventivos é uma tarefa complexa, nem sempre factível e depende de inúmeros fatores, da fonte de dados à metodologia utilizada. Por um lado, não se pode deixar de emitir um alerta de uma possível descarga e esta efetivamente acontecer, i.e., falha de alerta. Embora uma falha de alerta não significar que a descarga efetivamente vai atingir uma pessoa ou equipamento específico, a negligência deste critério é inadmissível do ponto de vista de segurança do trabalho. Por outro lado, também não é viável, em termos econômicos e de confiabilidade de sistema, parar as operações sempre que houver indícios de tempestade, pois os custos por inatividade de operações podem ser demasiadamente elevados, dado que um falso alerta significa, propriamente, uma parada nas atividades ou em parte delas. A tomada de decisão, neste caso, é encontrar aquela solução que minimize as falhas e os falsos alertas. A importância de cada critério passa a ser uma decisão de negócio.

Atualmente, os parques eólicos possuem normas técnicas internacionais como a IEC/TR 61400-24-2010 desenvolvida pela Comissão Eletrotécnica Internacional (*International Electrotechnical Commission* - IEC). Esta norma, detalhada em Lemos (2012) e Sorensen et al. (2008), traz vários aspectos de proteção contra descargas em turbinas eólicas, que ainda inclui procedimentos e recomendações para avaliação de risco, métodos apropriados para proteção e, ainda, orientação sobre segurança pessoal. Embora boa parte da proteção seja via sensores ou redes, por exemplo, abordagens que auxiliem na identificação e monitoramento em tempo real das tempestades também podem auxiliar a reduzir riscos de forma a atender critérios de segurança e confiabilidade. Outro ponto de atenção é que as áreas mais propícias para implantação de parques eólicos coincidem com os locais contendo frequentes registros de descargas atmosféricas (isso sem voltar ao fato levantado por Pinto Jr e Pinto (2021) de que o Brasil é o campeão em número de raios). Os aerogeradores se tornam, portanto, um ponto-alvo que deve ser monitorado com bastante cautela dada a sua maior probabilidade de ser atingido por uma descarga. Uma vez que é impossível impedir que uma descarga atmosférica ocorra, o que se deve fazer é melhorar as ferramentas existentes de proteção para que os riscos e prejuízos sejam reduzidos.

Desta forma, este artigo propõe uma abordagem capaz de fazer o monitoramento em tempo real de tempestades por meio de técnicas de aprendizado de máquina, não-supervisionadas e supervisionadas. Foi utilizado 1 ano de dados de descargas ocorridas ao redor de três pontos de interesse previamente definidos. Estes dados foram discretizados em intervalos de 5-5 minutos para simular a aplicação real. Para cada intervalo, técnicas de cluster foram utilizadas para extrair informações das possíveis tempestades naquele instante e, em seguida, estes dados servem como entrada para algoritmos supervisionados a fim de gerar alertas de advertência para regiões a serem protegidas. Estas regiões, na prática, podem ser mineradoras, ferrovias, redes aéreas ou parques eólicos que têm equipes que atuam a céu aberto e precisam garantir critérios de segurança e confiabilidade aos colaboradores e, ao mesmo tempo, buscam reduzir o tempo de inatividade das operações.

O restante do artigo é dividido da seguinte forma: a Seção 2 aborda os principais conceitos, fontes de dados e abordagens de inteligência computacional aplicadas ao tema; a Seção 3 descreve a base de dados e a metodologia desenvolvida; a Seção 4 apresenta os resultados e discussão, seguida da Seção 5 que endereça as considerações finais.

2. REFERENCIAL TEÓRICO

Esta seção apresenta os principais conceitos, trabalhos e métodos comumente utilizados para identificação e monitoramento de tempestades e, em especial, aqueles baseados em ML para geração de alertas de descargas atmosféricas.

As descargas podem ser (i) intra-nuvem, (ii) nuvem-nuvem, (iii) solo para ar e (iv) nuvem para solo (*cloud-to-ground*, ou CG), sendo esta última a mais perigosa e explorada. Elas podem ter cargas negativas (maioria) ou positivas (minoria) (Pinto Jr e Pinto, 2021). Já as tempestades podem ser categorizadas pelas suas características físicas e divididas em quatro tipos: (i) tempestades de célula única, (ii) aglomerados com múltiplas células, (iii) tempestades de linhas multicelulares e tempestades de super-células (Bala et al., 2017).

As soluções existentes para gerenciar tais eventos são baseadas, como antes mencionado, em sensores, redes de detecção, satélites meteorológicos com sensores imageadores

ou uma combinação de todos estes. Há ainda, modelos de aprendizado de máquina, supervisionados ou não, que usam estes dados em diferentes propostas, como: prever movimento de tempestades (Juntian et al., 2011), gerar alertas com base em diferentes valores do campo elétrico (Costa et al., 2014), gerar alertas com base no valor do campo elétrico, correlacionado com análise meteorológica e redes de satélites (Soares e Fonseca, 2019) ou variáveis meteorológicas (Mostajabi et al., 2019; Tervo et al., 2021).

Os sensores são equipamentos instalados no ponto a ser monitorado e medem o campo eletromagnético emitido pelas descargas. Embora sejam capazes de detectar a aproximação de tempestades severas e a distância da atividade elétrica, podem gerar muitos falsos alertas. A rede de detecção é formada por um conjunto de antenas que operam em uma frequência específica, VLF (frequência muito baixa), LF (baixa frequência) ou VHF (frequência muito alta), para detectar emissões eletromagnéticas dos raios. A frequência de operação e a proximidade delas determinam a qualidade do sinal e dos alertas emitidos. Nas redes, as antenas operam em conjunto e trazem maior informação e qualidade para determinação das tempestades. Satélites com sensores imageadores e sistemas com múltiplas fontes, ou *multisource*, auxiliam na detecção e rastreamento das tempestades e descargas, pois tendem a explorar o melhor dos dados de cada solução.

Aliados a todas estas fontes, há soluções baseadas em ML que trazem novas descobertas e melhorias às ferramentas existentes. Juntian et al. (2011) propuseram uma metodologia de identificação e agrupamento das descargas formando os clusters e rastreamento das tempestades no espaço-tempo. Com uma acurácia média de 75%, os resultados reportados indicaram uma capacidade de predizer a área de risco. Embora os autores não tenham focado na geração antecipada de alertas, entende-se que a metodologia pode ser estendida e aplicada. Neste caso, a velocidade da tempestade entre os intervalos de tempo poderia ser utilizada para calcular quando emitir o alerta.

Dada a imprevisibilidade no espaço e no tempo da dinâmica das tempestades e descargas, muitas pesquisas têm direcionado os seus esforços a soluções baseadas em inteligência computacional. Dentre as soluções baseadas em ML, Juntian et al. (2011) e Tervo et al. (2021) propuseram agrupar espacialmente as descargas e rastrear o movimento (direção) e quantidade de raios que pudessem ser emitidos nas tempestades. Estes últimos autores, indo além, também utilizaram diferentes modelos de ML para prever quanto de dano à rede elétrica as descargas poderiam causar. Mostajabi et al. (2019) utilizaram técnicas de aprendizado de máquina para prever, com sucesso, os riscos de raios próximos e distantes, tendo como base informações climáticas obtidas da rede de monitoramento SwissMetNet como: pressão do ar ao nível da estação observada, temperatura do ar, umidade relativa e velocidade do vento. Diferentes algoritmos foram treinados para diferentes estações na Europa, como La Dôle e Säntis, na Suíça, sendo que algumas delas eram situadas em elevadas altitudes, cuja probabilidade de serem atingidas por raios é maior. Santos et al. (2017) empregaram um modelo de Regressão Linear Múltipla e modelos globais HadGEM2-ES e CSIRO-Mk3.6 para a previsão de descargas do tipo nuvem-solo para o estado de São Paulo. Bala et al. (2017) revisaram diversas pesquisas que eram direcionadas à utilização de técnicas de inteligência computacional, otimização, mineração de dados, dentre outras sobre tempestades e/ou descargas. Alguns destaques do levantamento: a abordagem utilizada está diretamente relacionada à fonte e ao tipo de dado disponível; o desempenho dos modelos variam de $60-95\%$, embora a quantidade e qualidade dos dados disponíveis sejam muito distintas e, quase sempre, determinantes para a qualidade da predição; e, por fim, poucas pesquisas eram voltadas à geração de alertas para minimização de potenciais riscos, seja a pessoas, sistemas ou equipamentos.

Dentre as pesquisas mais similares à este presente estudo, podem ser citadas Tervo et al. (2021) e Mostajabi et al. (2019). Porém, há pelo menos três diferenças principais: (i) a base/fonte de dados, pois esta é proveniente do principal grupo de pesquisa nesta área no Brasil, (ii) o estudo é sobre três pontos situados no país, cuja dinâmica de tempestades é diferente dos demais – o Pará está localizado no Norte, onde os sistemas atmosféricos têm natureza convectiva, ou seja, calor e umidade atuam em conjunto para potencializar a energia para a formação das tempestades. As condições climáticas são moduladas por sistemas de diferentes escalas espaciais (local, meso e larga escala). Já em Minas Gerais, os sistemas atmosféricos são basicamente frontais, cujo pico da estação de raios coincide com os meses de verão do Hemisfério Sul e estão relacionados principalmente à ação da Zona de Convergência do Atlântico Sul (ZCAS) e (iii) nesta pesquisa os dados para a geração de alertas são extraídos das próprias descargas e utilizados como entrada para o método de ML. Além disso, argumenta-se que o foco desta pesquisa não é comparação de modelos, mas sim propor uma solução que possa dar apoio nas atividades das equipes de segurança e operações.

3. METODOLOGIA

3.1 Base de dados

A base de dados utilizada neste trabalho foi fornecida pelo Grupo de Eletricidade Atmosférica do Instituto Nacional de Pesquisas Espaciais (ELAT/INPE). Os dados são provenientes de três redes de superfície formadas, no total, por cerca de 110 sensores espalhados pelo Brasil que detectam radiação nas faixas de frequência de VLF e LF e, também, pelo sensor *Geostationary Lightning Mapper* (GLM) a bordo do satélite geoestacionário GOES-16, que detecta radiação na faixa do visível. As diferentes bases de dados são integradas de modo a permitir uma maior eficiência de detecção das descargas, visto que nenhum sistema individual detecta todas as descargas que ocorrem (Pinto Jr e Pinto, 2021).

O formato do dado utilizado neste trabalho foi o dia e hora das descargas com precisão em milissegundo, *timestamp*, em dd/mm/aaaa hh:mm:ss já com a correção do fuso horário, e as coordenadas onde as descargas ocorreram, Latitude e Longitude. Foi utilizado 1 ano de dado, de 04/04/2020 a 31/03/2021.

As regiões monitoradas correspondem a três áreas com raios de 50km de cobertura. Para facilitar a visualização, as descargas ocorridas nas três áreas monitoradas são apresentadas na Figura 1. É possível perceber um número maior de descargas nas duas primeiras figuras, que

correspondem a duas regiões do Brasil no estado de Minas Gerais (pontos P_1 e P_2), e menor na terceira situada no estado do Pará (ponto P_3).

3.2 Procedimentos Metodológicos

Nesta subseção são descritos os passos utilizados para fazer o monitoramento das tempestades e para geração de alertas preventivos.

O monitoramento das tempestades foi baseado em técnicas não-supervisionadas de aprendizado de máquina, também conhecidas como agrupamento ou clusters, tomando como inspiração os trabalhos de Juntian et al. (2011) e Tervo et al. (2021). Os passos desta etapa são descritos abaixo e eles foram feitos para cada ponto de interesse.

- Passo 1: Pré-processamento dos dados.
  Considerando 1 ano de dados, separou-se 2020 para treino e 2021 para teste. Acredita-se que a dinâmica das tempestades que ocorreram no primeiro ano também ocorreram no segundo sobre as regiões analisadas, visto que têm período de chuvas bem definido. Do início ao fim de cada período, estes dados foram separados em intervalos de 5 minutos para simular a aplicação real. Neste caso, cada subconjunto representava um intervalo de 5 minutos de dados de descargas.
- Passo 2: Determinar a área limite a ser monitorada.
  Filtrar as descargas que aconteceram dentro de um raio r do ponto de interesse P_i. Neste trabalho foi considerado $r = 30$km. Para calcular a distância das descargas de P foi utilizada a distância geodésica, disponível no *framework* GeoPy (Esmukov, 2022), mais fiel à curvatura da terra.
- Passo 3: Selecionar os métodos de agrupamento para determinar as tempestades.
  Diferentes técnicas estão disponíveis na literatura. Para esta aplicação foram comparados os métodos de aprendizado de máquina não-supervisionados MeanShift, DBScan e Optics implementados na Scikit-Learn (Pedregosa et al., 2011) e Clusteval (Taskesen, 2020). Estes métodos possuem fácil implementação, formam os agrupamentos com base na distância entre as descargas e possuem bom desempenho reportado na literatura, vide Wiwie et al. (2015).
- Passo 4: Determinar as métricas de desempenho.
  Métodos não-supervisionados são conhecidos por não possuírem uma saída/resposta conhecida para cada amostra. O objetivo neste caso foi agrupar as descargas e definir as tempestades. Neste trabalho foram utilizadas as métricas: coeficiente de silhueta e os índices Davies Bouldin e Calinski-Harabaz, conhecidas para este tipo de análise (Krasnov e Sen, 2019).
- Passo 5: Seleção do melhor método de cluster.
  Para cada intervalo, verificava se havia descargas. Quando eram detectadas, as posições de Latitude e Longitude eram submetidas aos métodos e calculadas as métricas antes mencionados. O melhor método de clusterização foi aquele que apresentou os melhores resultados nas métricas.
- Passo 6: Extrair as informações de cluster.
  Cada cluster formado correspondia a uma tempestade. Além de agrupar as descargas, também foram extraídas outras informações, quais sejam: (i) o número de tempestades, (ii) número de descargas na tempestade mais intensa, (iii) distância da tempestade mais próxima de P, (iv) distância da tempestade mais densa, entre outras.

As informações obtidas nesse processo são valiosas, pois elas indicam a situação meteorológica nas proximidades de cada ponto monitorado. Além disso, as informações extraídas foram utilizadas para gerar alertas que antecipem a possibilidade de uma descarga atingir o ponto de interesse. Este local pode ser, por exemplo, uma mina de céu aberto, linha férrea ou parque eólico onde esteja ocorrendo atividades de operação, montagens ou manutenção.

Os passos para geração antecipada de alertas são descritos a seguir.

- Passo 1: Geração da base de dados.
  O melhor algoritmo de cluster é utilizado para cálculo das variáveis de entrada (*features*) para o modelo supervisionado de classificação. Cada amostra contém as 4 variáveis descritas no Passo 6 anterior. Para estimar a saída, ou resposta do modelo, durante o cálculo dos clusters, verificava se 15 minutos à frente alguma descarga atingiu, ou não, a área alvo. Para exemplificar: suponha que para um dia qualquer, no período de 10:55 as 11:00 foram detectadas 15 descargas. Sobre elas eram obtidas as informações de cluster. Verificava-se, também, se no horário de 11:15 às 11:30 teve alguma descarga em um raio de 2km do ponto monitorado – área a ser protegida. Se sim, a saída esperada era 1, senão 0.
  Em termos práticos o que se buscava era um possível padrão nas tempestades que, caso ele ocorresse, poderia cair uma descarga sobre o ponto monitorado durante um intervalo de 15 minutos de operação, 15 minutos à frente. Se sim, um alerta de 30 minutos era gerado para que potenciais riscos, seja de pessoas ou equipamentos, fossem mitigados.
- Passo 2: Dividir a base entre treino e teste.
  Dados de 2020 foram utilizados para treino e 2021 para teste. Acredita-se que a dinâmica das tempestades ocorridas em 2020 sejam suficientes para representar o restante do período. Além disso, as regiões monitoradas possuem o período chuvoso bem definido, de setembro a março, tendo, portanto, amostras similares no treino e no teste.
- Passo 3: Pré-processamento.
  Balanceamento: Os dados gerados eram altamente desbalanceados, com apenas 5% de amostras com a classe 1. Para resolver este problema, duas técnicas foram aplicadas para remover amostras ruidosas na classe majoritária e depois balanceá-las, sendo elas: *TomekLinks* (Tomek, 1976) e *Condensed Nearest Neighbour* (Hart, 1968), respectivamente.
  Normalização/Padronização: Em relação a transformação nos dados, foram testados tanto dados com a normalização z-score, onde $z = (\mathbf{x} - \mu)/\sigma$, quanto com a normalização *Min-Max*, ou $\mathbf{x}_{scaled} = (\mathbf{x} - \mathbf{x}_{min})/(\mathbf{x}_{max} - \mathbf{x}_{min})$.
- Passo 4: Seleção de métodos e hiperparâmetros.
  Os seguintes métodos foram avaliados: (i) Naive Bayes, (ii) Florestas Aleatórias, conhecidas como *Random Forest* (RF), (iii) Máquina de Vetores de Suporte (SVM) e *Extreme Gradient Boosting* (XGBoost). Os hiperparâmetros para cada modelo foram

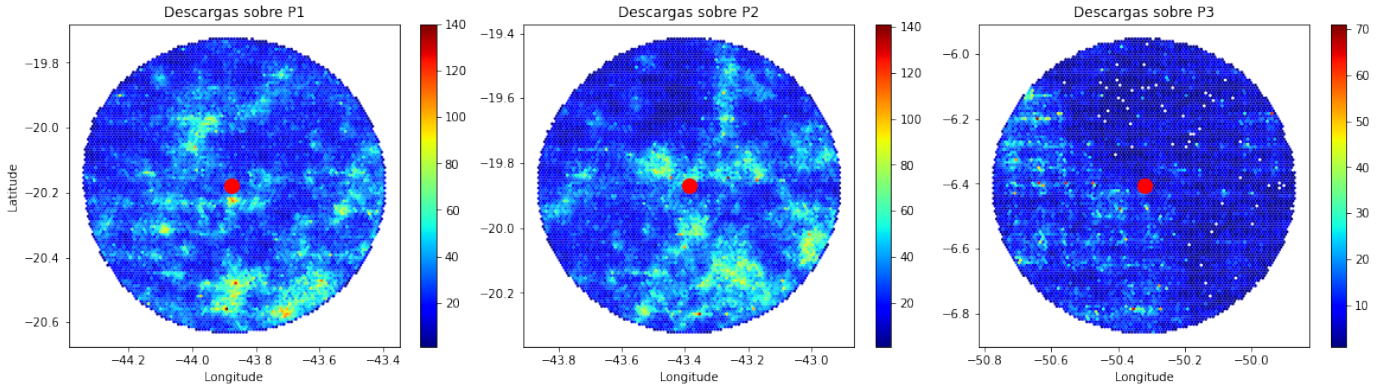

Figura 1. Ocorrência de descargas sobre P_1, P_2 e P_3 considerando um raio de observação de 50km.

estimados utilizando GridSearch. Todos implementados na SkLearn (Pedregosa et al., 2011)

- Passo 5: Desempenho dos modelos.
  Os modelos foram treinados para cada ponto monitorado (P_1, P_2, P_3), cada algoritmo e para cada tipo de pré-processamento. As métricas, explicadas em Tervo et al. (2021), utilizadas para calcular o desempenho dos modelos com base no conjunto de testes foram: acurácia, precisão, revocação (ou *recall*), F1-score e suporte.

Todos os experimentos foram implementados em Python e executados em um computador Intel(R) Core i7 com 8GB de RAM.

4. RESULTADOS E DISCUSSÃO

A média do número de clusters (N clusters) gerado para todos os intervalos de tempo, bem como os valores das métricas Calinski Harabasz (CH), Davies Bouldin (DB) e Silhueta são apresentados na Tabela 1 para o ponto P_1, P_2 e P_3.

Tabela 1. Resultado dos métodos de agrupamento para os pontos monitorados P_1, P_2 e P_3 considerando o número médio de clusters formados em cada janela de tempo e as métricas. Os melhores resultados estão destacados.

Ponto	**Método**	**N clusters**	**CH**	**DB**	**Silhueta**
P_1	MeanShift	3.32 ± 1.81	**603.35**	**0.28**	0.63
	Clusteval	2.80 ± 1.94	349.27	0.32	**0.83**
	DBScan	1.95 ± 2.29	17.33	1.08	0.09
	Optics	4.83 ± 3.68	69.49	1.32	0.25
P_2	MeanShift	3.35 ± 1.84	326.61	**0.28**	0.61
	Clusteval	2.71 ± 1.78	**332.97**	0.33	**0.82**
	DBScan	1.70 ± 3.15	5.09	0.75	-0.04
	Optics	4.26 ± 3.44	56.44	1.31	0.22
P_3	MeanShift	2.96 ± 2.03	**722.67**	**0.24**	0.50
	Clusteval	2.69 ± 1.85	194.19	0.33	**0.79**
	DBScan	0.70 ± 1.37	28.76	0.54	0.01
	Optics	2.89 ± 3.05	50.42	0.99	0.25

Utilizando o MeanShift como método de agrupamento, dado seu melhor desempenho no geral, a Figura 2 ilustra algumas tempestades agrupadas para o ponto P_2, considerando intervalos de $5-5$ minutos, durante 40 minutos. Observa-se que há 3 tempestades a Nordeste de P_2, sendo a de cor azul a mais densa. Com o tempo, as tempestades se intensificam (em número de descargas) e, aos poucos, vão se afastando para Leste. A atenção, neste caso, seria sobre uma possível aproximação desta ou de outra tempestade.

Com as informações de agrupamento das tempestades, a próxima etapa foi direcionada à geração de alertas. Por questões de simplicidade, apenas os resultados dos melhores métodos são apresentados a seguir.

Para o ponto P_1, o melhor modelo treinado foi um XGBoost utilizando normalização z-score nos dados. As respostas para cada classe, ou matriz de confusão, são apresentadas na Tabela 2 e os resultados das métricas sumarizados na Tabela 3.

Tabela 2. Matriz de confusão para o ponto P_1

		Preditos pelo ML	
		Sem descarga	Com descarga
Real	Sem descarga	0.68	0.32
	Com descarga	0.33	0.67

Tabela 3. Desempenho do modelo XGBoost com normalização z-score para o ponto P_1

Classe	**Acurácia**	**Precisão**	**Recall**	**F1-score**	**Suporte**
0	0.677404	0.953589	0.678137	0.792614	7817
1	0.677404	0.172368	0.670077	0.274202	782

Como apresentado nas Tabelas 2 e 3 em relação ao ponto P_1, o modelo treinado consegue acertar a classificação em quase 70% dos casos, independentemente da classe em questão. Considerando o cenário do problema, onde não se tem uma certeza física e exata de como se origina tempestades, pode-se considerar que os resultados alcançados pelo modelo foram bons.

Para o ponto monitorado P_2 o melhor modelo foi baseado no método SVM também com a normalização z-score. As respostas para cada classe são apresentadas na Tabela 4 e as métricas na Tabela 5.

Tabela 4. Matriz de confusão para o ponto P_2

		Preditos pelo ML	
		Sem descarga	Com descarga
Real	Sem descarga	0.75	0.25
	Com descarga	0.40	0.60

Como apresentado nas Tabelas 4 e 5 em relação ao ponto P_2, o modelo treinado obteve resultados não tão balanceados quanto o melhor modelo do ponto P_1. O método

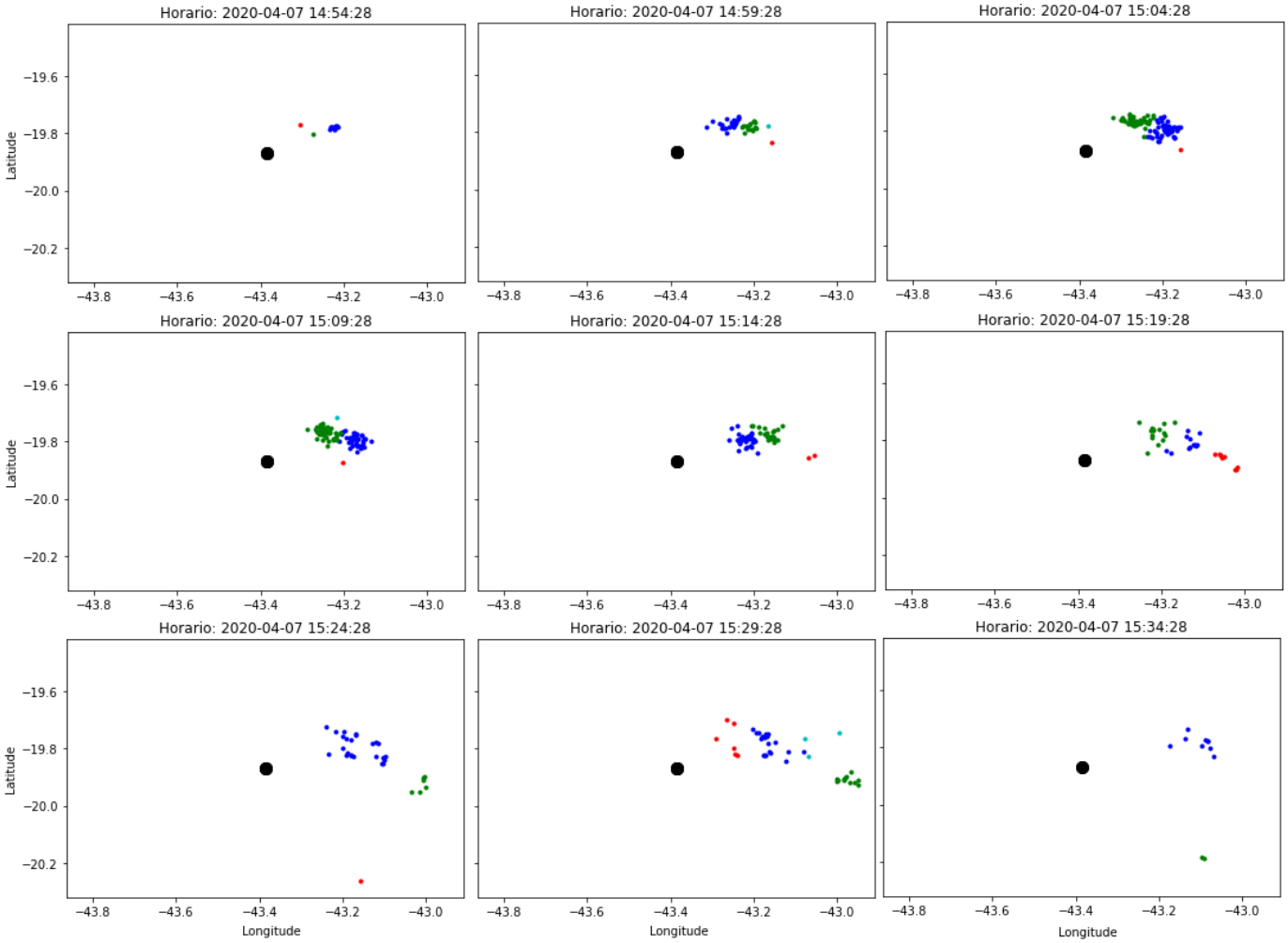

Figura 2. Agrupamento de tempestades próximas a P_2 durante 40 minutos.

Tabela 5. Desempenho do modelo SVM com normalização z-score para o ponto P_2

Classe	Acurácia	Precisão	Recall	F1-score	Suporte
0	0.737795	0.946457	0.752753	0.838565	5448
1	0.737795	0.202487	0.595819	0.302254	574

SVM com a normalização z-score classifica corretamente 75% das vezes para a classe 0 e 60% quando a classe é 1.

Para o ponto P_3 o melhor modelo foi gerado com base no método SVM com normalização $min-max$, ou *StandardScaler*. A matriz de confusão para este ponto é apresentada na Tabela 6 e as métricas na Tabela 7.

Tabela 6. Matriz de confusão para o ponto P_3

		Preditos pelo ML	
		Sem descarga	Com descarga
Real	Sem descarga	0.74	0.26
	Com descarga	0.40	0.60

Tabela 7. Desempenho do modelo SVM com normalização $min-max$ para o ponto P_3

Classe	Acurácia	Precisão	Recall	F1-score	Suporte
0	0.727992	0.961130	0.737504	0.834597	8583
1	0.727992	0.145944	0.600624	0.234828	641

Por fim, Como mostrado nas Tabelas 6 e 7 em relação ao ponto P_3, os resultados se aproximam significativamente com os obtidos no P_2. Outra coincidência é que em ambas as regiões monitoradas o melhor modelo foi o SVM, tendo como diferença apenas o método de normalização.

Como o conjunto de testes, que reflete os dados reais, são altamente desbalanceados (vide o valor do suporte nas Tabelas 3, 5 e 7) e o foco está em reduzir tanto falhas quanto falsos alertas, as métricas mais importantes a serem independentemente observadas são revocação e precisão, respectivamente, ou F1-score no caso de analisar harmonicamente as duas anteriores. A revocação é dada por $VP/(VP+FN)$ e a precisão por $VP/(VP+FP)$, sendo VP os verdadeiros positivos, FN os falsos negativos e FP os falsos positivos. A saber, F1-score é calculado por $2*((Pr*Re)/(Pe+Rr))$, sendo Pr a previsão e Re a revocação.

O foco principal da metodologia, neste caso, foi o de gerar alertas que antecipem a ocorrência de descargas a fim de minimizar potenciais riscos sobre pessoas. Em uma aplicação prática, os dados são recebidos em tempo real e consumidos de 5 em 5 minutos, e a predição é realizada para aquela janela de dados, no intervalo de tempo $t_0...t_5$. Caso o resultado do modelo de ML seja positivo, um alerta é emitido indicando que de 15 a 30 minutos à frente,

$t_{20}...t_{35}$, há a probabilidade de uma descarga atingir a área em um raio de 2km do ponto de interesse. Em tempestades intensas, mais descargas estarão acontecendo e o modelo deverá gerar alertas em intervalos seguidos de tempo. Na prática, há pelo menos duas formas de gerenciar os alertas: (i) renovação: mantém o mesmo alerta e apenas atualiza o horário do final do alerta; (ii) dispara um novo alerta a cada predição. Uma decisão de negócio deve ser tomada e a solução mais aceitável é implementada.

Como novas perspectivas, acredita-se que a aplicação desta abordagem de ML com dados meteorológicos como aqueles utilizados por Mostajabi et al. (2019) e/ou informações de campo elétrico, vide Igarashi et al. (2011); Ferro et al. (2011); Srivastava et al. (2015), quando disponíveis, são fortes candidatas a gerar bons resultados para uma aplicação. Além disso, também podem ser exploradas a aplicação para eventos extremos que, a cada dia, têm acontecido com mais frequência (Medeiros et al., 2019). A inclusão de dados meteorológicos também é viável, vide os bons resultados reportados por Mostajabi et al. (2019) e permitiria novas abordagens, como séries temporais (Oliveira e Lucas et al., 2020; Silva et al., 2019).

5. CONSIDERAÇÕES FINAIS

Este artigo propôs o desenvolvimento de uma aplicação para agrupar os dados de descargas atmosféricas e monitorar as tempestades em tempo real, com base em uma fonte de dados altamente confiável do grupo ELAT/INPE. Informações como a quantidade de tempestades, número de raios e distância das tempestades foram utilizadas em um modelo de ML capaz de gerar alertas antecipados de descargas. Com este tipo informação, setores como mineradoras, concessionárias de energia, empresas de aviação e outras podem desenvolver ações de prevenção e mitigação de riscos.

Em uma aplicação prática, a visualização em tempo real da formação das tempestades e o movimento das mesmas servem de apoio às equipes de segurança do trabalho, engenharia elétrica e operacional em relação ao entendimento dos alertas gerados, e vice-versa. Tempestades muito intensas, em número de descargas, mesmo que formem poucos clusters/grupos, indica uma desordem na atividade eletromagnética no local e devem ser vistas com atenção. Ao mesmo tempo, tempestades menos intensas e distantes geograficamente dos pontos monitorados devem ser observadas, especialmente em relação a direção. Dados de outras fontes, quando disponíveis, podem ser utilizados em conjunto com esta abordagem, por exemplo, sensores de campo elétrico. Uma vez que eles são instalados localmente e no ponto-alvo, alterações no campo elétrico podem indicar e confirmar a necessidade de emissão de alertas. Acredita-se que esta agregação possa minimizar as paradas indevidas e possibilite a emissão de alertas com mais eficácia.

As soluções existentes variam da abordagem à fonte e tipo de dado utilizado. O que se percebe na literatura é que múltiplas fontes de dados, uso de informação climática e de sensores poderiam melhorar ainda mais a precisão dos resultados, porém nem sempre elas estão disponíveis e/ou são de domínio público. Além disso, cada região possui dinâmicas diferentes e isso requer o desenvolvimento de soluções especialistas. Embora o período chuvoso seja bem definido em algumas regiões do Brasil, por exemplo, ele ainda possui outras especificidades que podem ser exploradas na modelagem de predição como horários mais comuns de tempestade, sentido do vento, campo elétrico e entre outros.

AGRADECIMENTOS

Os autores agradecem o financiamento da pesquisa proporcionada pela Vale S.A. e a oportunidade do desenvolvimento do trabalho nas dependências da FITec (`www.fitec.org.br/`).

REFERÊNCIAS

Bala, K., Choubey, D.K., e Paul, S. (2017). Soft computing and data mining techniques for thunderstorms and lightning prediction: a survey. In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, volume 1, 42–46. IEEE. doi: 10.1109/ICECA.2017.8203729.

Costa, P.F., Ferreira, M.A., e Salame, Y.C. (2014). Preventive lightning protection using local static electric field measurements without mobile elements: first brazilian experience. In *International Conference on Grounding and Earthing & 6th International Conference on Lightning Physics and Effects*, 6.

Esmukov, K. (2022). Geopy. URL `github.com/geopy/geopy`. [Online; acessado em 02-02-2022].

Ferro, M.A.D.S., Yamasaki, J., Pimentel, D.R.M., Naccarato, K.P., e Saba, M.M.F. (2011). Lightning risk warnings based on atmospheric electric field measurements in brazil. *Journal of Aerospace Technology and Management*, 3, 301–310. doi:10.5028/jatm.2011.03032511.

Geng, Y.a., Li, Q., Lin, T., Yao, W., Xu, L., Zheng, D., Zhou, X., Zheng, L., Lyu, W., e Zhang, Y. (2021). A deep learning framework for lightning forecasting with multi-source spatiotemporal data. *Quarterly Journal of the Royal Meteorological Society*, 147(741), 4048–4062. doi:10.1002/qj.4167.

Hart, P. (1968). The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 14(3), 515–516. doi:10.1109/TIT.1968.1054155.

Igarashi, A.Y.S., Leando, G., e Leite, E.A. (2011). Alerta de incidências de descargas atmosféricas utilizando lógica fuzzy. In *10th Brazilian congress on computational intelligence*.

Juntian, G., ShanQiang, G., e Wanxing, F. (2011). A lightning motion prediction technology based on spatial clustering method. In *2011 7th Asia-Pacific International Conference on Lightning*, 788–793. IEEE. doi: 10.1109/APL.2011.6110234.

Krasnov, F. e Sen, A. (2019). The number of topics optimization: Clustering approach. *Machine Learning and Knowledge Extraction*, 1(1), 416–426. doi:10.3390/make1010025.

Lemos, D.F.A. (2012). Normalização e Desempenho de Aerogeradores. Technical report, Centro de Tecnologias do Gás e Energias Renováveis. [Online; acessado em 02-02-2022].

Libório, M., Maia, W., Martins, C., Ekel, P., Laudares, S., e Bernardes, P. (2020). Reducing costs of preventive lightning systems by locational optimization. *GOT:*

Revista de Geografia e Ordenamento do Território, (20), 149. doi:10.17127/got/2020.20.007.

Medeiros, E.S.d., Alves, M.A., e Souza, S.A. (2019). Estimação de nível de retorno da precipitação máxima diária na cidade de jataí-go. *Ciência e Natura*, 41, 36. doi:10.5902/2179460X35639.

Mostajabi, A., Finney, D.L., Rubinstein, M., e Rachidi, F. (2019). Nowcasting lightning occurrence from commonly available meteorological parameters using machine learning techniques. *Npj Climate and Atmospheric Science*, 2(1), 1–15. doi:10.1038/s41612-019-0098-0.

Oliveira e Lucas, P., Alves, M.A., e Silva, P.C.d.L., e Guimarães, F.G. (2020). Reference evapotranspiration time series forecasting with ensemble of convolutional neural networks. *Computers and Electronics in Agriculture*, 177, 105700. doi:10.1016/j.compag.2020.105700.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12, 2825–2830.

Pinto Jr, O. e Pinto, I.R.C.A. (2021). *Brasil campeão mundial de raios*. Artliber.

Santos, A.P.P., Coelho, C.A., Pinto Junior, O., dos Santos, S.R.Q., de Lima, F.J.L., e de Souza, E.B. (2018). Climatic diagnostics associated with anomalous lightning incidence during the summer 2012/2013 in southeast brazil. *International Journal of Climatology*, 38(2), 996–1009. doi:doi.org/10.1002/joc.5227.

Santos, A.P.P.d., Júnior, O.P., dos Santos, S.R.Q., de Lima, F.J.L., de Souza, E.B., de Morais, A.A.R., Ávila, E.E., Pedernera, A., et al. (2017). Climatic projections of lightning in southeastern brazil using cmip5 models in rcp's scenarios 4.5 and 8.5. *American Journal of Climate Change*, 6(03), 539. doi:10.4236/ajcc.2017.63027.

Silva, P.C.L., Sadaei, H.J., Ballini, R., e Guimarães, F.G. (2019). Probabilistic forecasting with fuzzy time series. *IEEE Transactions on Fuzzy Systems*, 28(8), 1771–1784. doi:10.1109/TFUZZ.2019.2922152.

Soares, W.S. e Fonseca, F.O.G. (2019). Lightning monitoring systems a case study applied to the iron ore minning. an approach to meeting standards iec 62793-5, iec 627139 iec 62305/nbr-5419. In *2019 International Symposium on Lightning Protection (XV SIPDA)*, 1–6. doi:10.1109/SIPDA47030.2019.8951594.

Sorensen, T.S., Plumer, J., Montanyà, J., Krogh, T.H., Hermoso, B., Birkl, J., Gehlhaar, T., McNiff, B., Bertelsen, K., Peesapati, V., et al. (2008). The update of iec 61400-24 lightning protection of wind turbines. In *29th International Conference on lightning protection*, 5.

Srivastava, A., Mishra, M., e Kumar, M. (2015). Lightning alarm system using stochastic modelling. *Natural Hazards*, 75(1), 1–11. doi:10.1007/s11069-014-1247-8.

Taskesen, E. (2020). Clusteval: a python package for unsupervised cluster validation. doi:10.5281/zenodo.5745348. URL `github.com/erdogant/clusteval`. [Online; acessado em 02-02-2022].

Tervo, R., Láng, I., Jung, A., e Mäkelä, A. (2021). Predicting power outages caused by extratropical storms. *Natural Hazards and Earth System Sciences*, 21(2), 607–627. doi:10.5194/nhess-21-607-2021.

Tomek, I. (1976). Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics*, 6, 769–772. doi:10.1109/TSMC.1976.4309452.

Wiwie, C., Baumbach, J., e Röttger, R. (2015). Comparing the performance of biomedical clustering methods. *Nature methods*, 12(11), 1033–1038. doi:10.1038/nmeth.3583.