UNIVERSIDADE FEDERAL DE ITAJUBÁ

INSTITUTO DE RECURSOS NATURAIS
PROGRAMA DE GRADUAÇÃO EM CIÊNCIAS ATMOSFÉRICAS

RELAÇÃO ENTRE RELÂMPAGOS E AS PROPRIEDADES FÍSICAS DAS TEMPESTADES

MONOGRAFIA DE GRADUAÇÃO

Juliano dos Reis Monteiro

Itajubá, MG, Brasil

2022

RELAÇÃO ENTRE RELÂMPAGOS E AS PROPRIEDADES FÍSICAS DAS TEMPESTADES

---

por

Juliano dos Reis Monteiro

Monografia apresentada à comissão examinadora Programa de Graduação em Ciências Atmosféricas da Universidade Federal Itajubá (UNIFEI, MG), como requisito parcial para obtenção do grau de
Bacharel em Ciências Atmosféricas.

Orientador: Dr. Prof. Enrique Vieira Mattos

Itajubá, MG, Brasil
2022

Universidade Federal de Itajubá
Instituto de Recursos Naturais
Programa de Graduação em Ciências Atmosféricas

A Comissão Examinadora, abaixo assinada, aprova a Monografia

RELAÇÃO ENTRE RELÂMPAGOS E AS PROPRIEDADES FÍSICAS DAS TEMPESTADES

elaborada por

Juliano dos Reis Monteiro

Como requisito parcial para a obtenção do grau de

Bacharel em Ciências Atmosféricas

Comissão Examinadora:

Enrique Vieira Mattos

Enrique Vieira Mattos, Dr. (UNIFEI)

(Presidente/Orientador)

Michelle S. Reboita

Michele Simões Reboita, Dra.
(UNIFEI)

Ricardo Batista Vilela

Ricardo Batista Vilela, Msc. (CLIMATEMPO)

Itajubá, 12 de Julho de 2022.

AGRADECIMENTOS

Primeiramente, aos meus queridos pais Nivaldo e Maria Zely por todo o apoio incondicional concedido durante toda a minha vida, zelando por me ensinar características éticas fundamentais como: caráter, persistência, comprometimento, valores morais e, acima de tudo, respeito. Agradeço e reconheço imensamente toda a dedicação deles durante todos esses anos e obrigado por tornarem possível minha qualificação em Ensino Superior. Essa formação representa a vitória da batalha de vocês em, não só me proporcionar esta qualificação, mas também na minha formação pessoal. Obrigado por tudo!

À minha querida irmã Bruna que sempre foi minha fonte de inspiração por tudo que ela é, já conquistou e, acima de tudo, um exemplo de pessoa. Obrigado por sempre estar disposta em me ajudar, seja no conhecimento técnico como através de conselhos.

Aos meus amigos e irmãos cunhenses Lucas e João por todo apoio, conselhos e momentos durante todos esses anos, os quais foram muito valiosos para a construção dessa importante etapa da minha vida.

Ao meu orientador e amigo Prof. Dr. Enrique Vieira Mattos por toda a atenção, didática e confiança depositada a mim para o desenvolvimento deste trabalho. Obrigado pelo exemplo de pessoa, profissionalismo, competência, seriedade e organização. Agradeço também pela primeira oportunidade no ano de 2018 através do projeto de iniciação científica, a qual foi responsável por abrir as portas da universidade para mim fazendo com que eu me identifica-se pela área da pesquisa.

A todos meus professores da UNIFEI, em especial aos do Instituto de Recursos Naturais (IRN), por passarem seus conhecimentos a diante e por toda a amizade desenvolvida nesses anos.

Aos meus queridos amigos e colegas de turma do Curso de Ciências Atmosféricas da UNIFEI. Em especial aos amigos: Aline Araújo, Thales Victor, Alysson Fernando, Thaís Aparecida, Raquel Gonçalves, Denis William, Geovane Carlos, Thales Heitor e Fabiana Teixeira. Obrigado pelo ótimo convívio, companhia e amizade.

Ao Centro de Previsão do Tempo e Estudos Climáticos (CPTEC) do Instituto Nacional de Pesquisas Espaciais (INPE) pelo fornecimento das imagens de satélite do GOES-13 e do algoritmo ForTracC.

À empresa Climatempo que gentilmente forneceu os dados de relâmpagos para a realização dessa pesquisa, em especial ao Wagner Flauber Araújo de Lima e Ricardo Batista Vilela.

A meus pais NIVALDO e MARIA ZELY, minha Irmã BRUNA e meus QUERIDOS AMIGOS.

"Característica marcante das tempestades, os relâmpagos causam espanto tanto pela beleza quanto pelo poder de destruição. Eles podem ser definidos, de modo simplificado, como transferências de cargas elétricas entre as nuvens e entre estas e o solo, mas a origem dessas cargas e muitos fatores envolvidos na liberação das faíscas são pouco conhecidos".

Pinto Jr et al, 1997, p. 26.

RESUMO

Monografia de Graduação
Programa de Graduação em Ciências Atmosféricas
Universidade Federal de Itajubá, MG, Brasil

RELAÇÃO ENTRE RELÂMPAGOS E AS PROPRIEDADES FÍSICAS DAS TEMPESTADES.

AUTOR(A): JULIANO DOS REIS MONTEIRO
ORIENTADOR: ENRIQUE VIEIRA MATTOS
Local e Data da Defesa: Itajubá, 12 de julho de 2022

O estado de São Paulo é anualmente afetado por alta incidência de relâmpagos produzidas por nuvens de tempestades que podem causar diversos prejuízos socioeconômicos aos diversos setores da economia. No entanto, ainda inexistem ferramentas que possibilitem prever as características elétricas das tempestades. Nesse sentido, o objetivo desse trabalho é analisar as relações entre as propriedades físicas de sistemas covnectivos com e sem relâmpagos e, por meio da aplicação de técnicas de inteligência artificial (IA) avaliar ferramentas que possibilitem prever as características elétricas das tempestades. Nesse contexto foram utilizadas imagens do canal infravermelho (10,7 μm) do satélite geoestacionário *Geostationary Operational Environmental Satellite*-13 (GOES-13) e dados de relâmpagos da rede *Earth Networks Total Lightning Network* (ENTLN) compreendendo os anos de 2013 a 2017. Os sistemas convectivos foram identificados e rastreados através do algoritmo *Forecast and Tracking the Evolution of Cloud Clusters* (ForTraCC). Ao todo, foram identificadas 57446 tempestades, as quais 70% foram divididas para treinamento dos modelos de aprendizado de máquina (regressão linear, regressão logística, árvore de decisão e florestas aleatórias) e 30% para validação. Foram observadas diferentes correlações de Pearson entre o tamanho (0,49), taxa de expansão (0,03) e temperatura mínima (-0,42) dos sistemas convectivos e a ocorrência de relâmpagos. No contexto da estimativa de relâmpagos através de modelos de IA, o modelo de regressão linear não apresentou uma boa (coeficiente R^2 de 0,3) performance, enquanto os modelos de classificação mostraram uma eficiência superior a 70% em predizer corretamente o tipo de tempestade, sendo que o modelo de regressão logística obteve o melhor desempenho (acurácia de 80%). Tais resultados indicam um importante passo para auxiliar no desenvolvimento de uma ferramenta que auxilie a previsão de curtíssimo prazo de tempo *(nowcasting*).

Palavras-chave: Tempestades; Relâmpagos; Inteligência artificial

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS, SIGLAS E SÍMBOLOS

A	-	Ampere
CCM	-	Complexos Convectivos de Mesoescala
ENTLN	-	Earth Networks Total Lightning Network
ER	-	Erro Residual
FN	-	Falso Negativo
FP	-	Falso Positivo
GLM	-	Geostationary Lightning Mapper
IA	-	Inteligência Artificial
IN	-	Intra-Nuvem
IQR	-	Intervalo Interquartílico
kHz	-	QuiloHertz
MAE	-	Erro Absoluto Médio
MSE	-	Erro Quadrático Médio
NS	-	Nuvem-Solo
RMSE	-	Raiz do Erro Quadrático Médio
SC	-	Sistemas Convectivos
SCM	-	Sistemas Convectivos de Mesoescala
SCAs	-	Sistemas Convectivos Alongados
TB	-	Temperatura de Brilho
TI	-	Temperatura de Inversão
TC	-	Temperatura de Colisão
TMED	-	Temperatura Média
TMIN	-	Temperatura Mínima
TMIN9	-	Temperatura Mínima do *Kernel* de 9 *Pixels*
VN	-	Verdadeiro Negativo
VP	-	Verdadeiro Positivo
ZH	-	Refletividade Horizontal

SUMÁRIO

1. Introdução

Os Sistemas Convectivos de Mesoescala (SCM) representam um dos mais importantes sistemas atmosféricos. O entendimento sobre os SCM é fundamental para compreender o tempo e o clima de uma determinada região. Os SCM podem gerar danos decorrentes de grandes volumes de precipitação, ventos fortes, queda de granizo e ampla atividade elétrica (HOUZE JR, 1993; MADDOX, 1980; PINTO JR; PINTO, 2000; WALLACE; HOBBS, 2006). Geralmente, esses sistemas são compostos por aglomerados de nuvens profundas e organizadas caracterizadas por uma extensa parte estratiforme no limite da troposfera assumindo a forma de bigorna. São formados por um conjunto de nuvens Cumulonimbus (Cb) que duram de 6 a 12 horas associados a intensas correntes ascendentes em seu interior e com área de 100 km ou mais em escala horizontal (HOUZE JR, 1993). Os SCM podem ser divididos em: Linhas de Instabilidade, Complexos Convectivos de Mesoescala (CCM) e Sistemas Convectivos Alongados (SCAs). Linhas de instabilidade são caracterizadas pela organização linear de tempestades severas composta por diversas nuvens Cb na vanguarda de sistemas convectivos formados geralmente em decorrência da passagem de um sistema frontal frio (NEWTON, 1950; RIBEIRO, 2018) Os CCM foram definidos por Maddox (1980) como tempestades que possuem excentricidade aproximadamente circular ($\geq 0,7$), área total maior ou igual a 100000 km² e temperatura de brilho do topo menor ou igual a -32 ºC. Ao mesmo tempo, o núcleo convectivo da tempestade deve apresentar área maior ou igual a 50000 km² e temperatura de brilho menor ou igual a -52 ºC, ao passo que todas essas condições devem perdurar por no mínimo 6 horas. Já os SCAs foram analisados por Anderson e Arritt (1998) que perceberam a formação de sistemas convectivos que se assemelhavam aos critérios de identificação de CCM, no entanto, possuíam uma forma mais alongada (excentricidade baixa).

A região Sul e Sudeste do Brasil são um dos locais mais favoráveis à ocorrência de SCM no mundo (ZIPSER *et al*., 2006). Esses sistemas são usualmente formados a leste da Cordilheira dos Andes, próximo a Bacia do Prata (onde se concentra 80% da produção hidroelétrica do continente). Possuem como principal característica de formação o transporte de umidade proveniente da Amazônia devido ao escoamento meridional ou atuação dos Jatos de Baixos Níveis (JBN). Durkee e Mote (2009) e Velasco e Fritsch (1987) analisaram as ocorrências de CCM em latitudes médias na América do Sul e notaram que os CCM sul-americanos são, em média maiores espacialmente e mais duradouros em relação àqueles encontrados na América do Norte.

Os SCM são caracterizados por possuírem elevada quantidade de gelo, água líquida super-resfriada e intensas correntes ascendentes e descendentes, sendo essas condições essenciais para a formação dos relâmpagos. Nuvens com grande desenvolvimento vertical possuem regiões com particularidades diferentes, seja em razão da presença de diferentes tipos e formatos de hidrometeoros ou em questão das propriedades termodinâmicas. Pinto Jr e Pinto (2000) relatam que a existência de centros de cargas no interior das nuvens acima da isoterma de 0 ºC indica que a presença de hidrometeoros congelados seja um importante fator para o processo de eletrização das nuvens. Em consistência MacGorman e Rust (1998) e Saunders (1993) observaram a existência de uma região das nuvens localizada entre a isoterma de 0 ºC e -20 ºC, a qual dispõe da presença de granizo e cristais de gelo, denominada "fase mista". Nessa região, quando sujeita a fortes correntes ascendentes e descendentes, ocorre a colisão entre tais hidrometeoros que por meio de trocas de cargas elétricas podem possibilitar a eletrificação das nuvens.

Devido à complexidade dos processos microfísicos e macrofísicos envolvidos na eletrificação das nuvens, os processos que explicam a formação das cargas elétricas no seu interior ainda requerem grande avanço científico. Por outro lado, as teorias sobre os mecanismos de separação de cargas dentro das nuvens são bem consolidadas. Partindo da concepção que a colisão entre cristais de gelo e granizo de tamanhos diferentes seja um dos princípios dos processos de formação de relâmpagos (REYNOLDS *et al.*, 1957), existem três principais teorias que explicam a separação dos centros de cargas elétricas dentro de uma nuvem de tempestade. O primeiro mecanismo é conhecido como processo colisional indutivo (Figura 1). Este mecanismo considera a pré-existência de um campo elétrico (orientado para baixo) que possibilita a indução de cargas positivas (negativas) na parte inferior (superior) do granizo e/ou *graupel* durante o movimento de queda dentro da nuvem. Dessa forma, ao colidir com cristais de gelo ocorre uma transferência de elétrons de modo que o granizo (cristal de gelo) se torna carregado negativamente (positivamente). Sendo assim, com a atuação da gravidade, o granizo (cargas negativas) por ser mais massivo se concentra na base da nuvem, enquanto os cristais de gelo (cargas positivas) ficam suspensos na nuvem, configurando assim dois centros de cargas bem definidos denotando uma estrutura elétrica dipolar (WALLACE; HOBBS, 1977). Entretanto, esse processo tem sido alvo de críticas, ao passo que, experimentos laboratoriais mostraram que a presença do campo elétrico de tempo bom não atuaria na indução de cargas no granizo, mas sim na intensificação dos centros de carga já existentes devido a processos distintos (MATTOS, 2009; PINTO JR; PINTO, 2000).

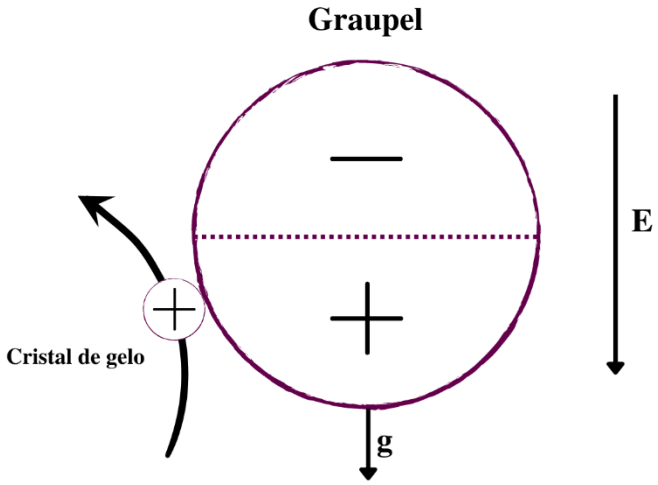

Figura 1 - Mecanismo de carregamento indutivo. A presença de um campo elétrico orientado na vertical induz a separação de cargas positivas e negativas no *graupel* durante sua queda em razão da gravidade. Ao colidir com cristais gelo, há uma remoção de cargas positivas do *graupel* tornando-o negativamente carregado, enquanto os cristais de gelo se tornam positivamente carregados. Fonte: Adaptada de Saunders (2008).

O segundo mecanismo de separação de cargas elétricas é o processo colisional termoelétrico (Figura 2. Esse processo propõe que a transferência de carga não depende do campo elétrico local, mas sim da temperatura de inversão (TI) e da temperatura do local de colisão (TC). A TI é estimada em aproximadamente -15 ºC e está localizada aproximadamente em 6 km de altura (próximo do centro de cargas negativas, WILLIAMS, 1989). Dessa forma, nesse mecanismo de eletrificação, se TC for menor (maior) que -15 ºC o granizo tende a ficar negativamente (positivamente) carregado. Em ambos os casos, devido à diferença de massa entre os hidrometeoros, ocorre a separação gravitacional, conduzindo a formação de uma estrutura de cargas tripolar dentro da nuvem (WALLACE; HOBBS 2006; WILLIAMS, 1989). Evidências que a temperatura do local de colisão e o conteúdo de água congelada no interior das nuvens eram importantes parâmetros acerca dos processos de eletrificação foram verificadas por Takahashi (1978). Neste estudo, através de um aparato experimental que simulava o processo de eletrificação durante a ocorrência de *riming*, isto é, congelamento de água líquida super-resfriada em contato com uma superfície congelada, notou-se que a polaridade dos hidrometeoros envolvidos nos processos de colisão (granizo ou *graupel* e cristais de gelo) se diferenciava dependendo do local de ocorrência da colisão, uma vez que, o *graupel* assumiria carga positiva (negativa) abaixo (acima) da temperatura de -10 ºC.

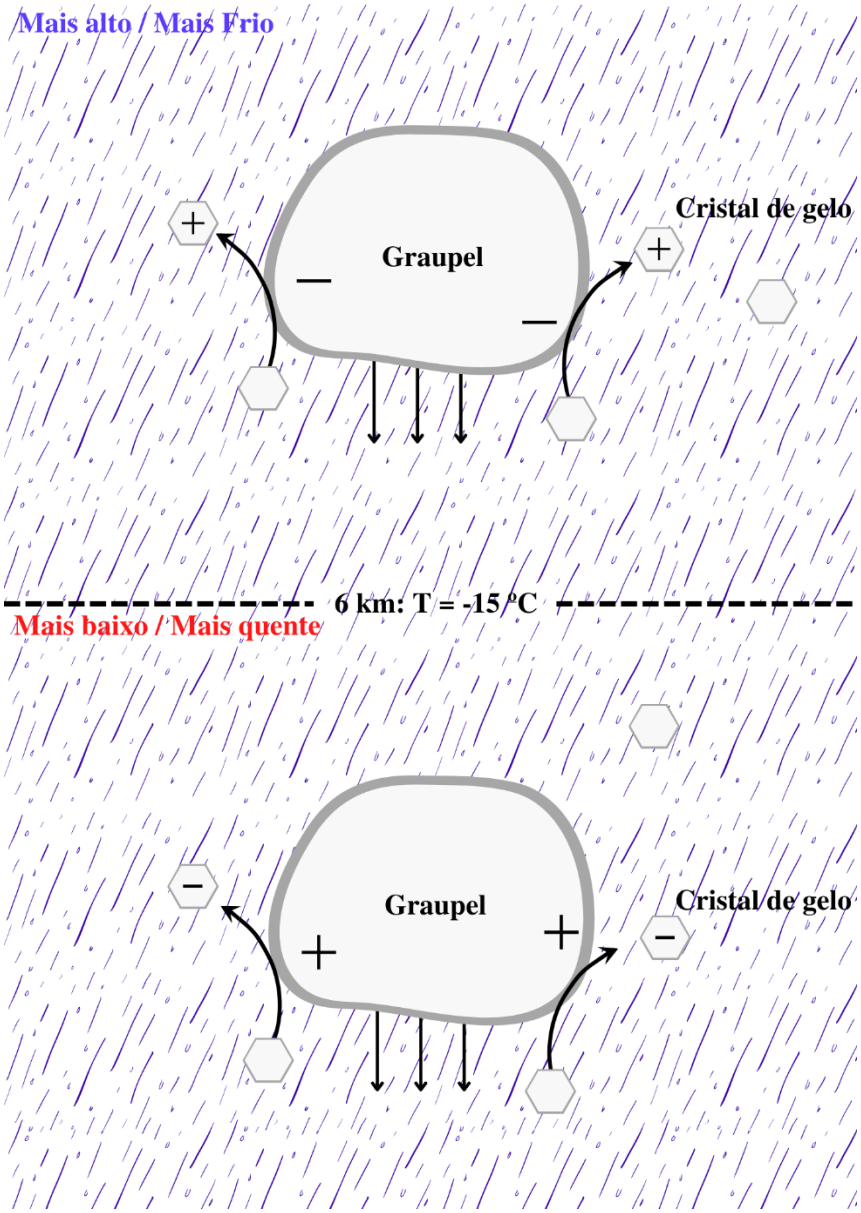

Figura 2 – Representação do mecanismo de eletrificação termoelétrico. A colisão entre *graupel* e cristais de gelo acima (abaixo) da isoterma de -15 ºC torna o *graupel* carregado negativamente (positivamente). Fonte: Adaptada de Williams (1988).

O terceiro mecanismo de eletrificação é conhecido como processo convectivo (Figura 3). Este mecanismo é fundamentado por dois processos principais: ionização de moléculas próximas ao topo da nuvem por raios cósmicos e pela existência de um intenso campo elétrico produzido por estruturas pontiagudas presentes na superfície terrestre (conhecido como efeito corona). Dessa maneira, nos estágios iniciais de uma nuvem *Cumulus* as correntes ascendentes elevam esses íons positivos presentes na superfície até o interior da nuvem, os quais entrando em contato com gotículas de água são aprisionados e conduzidos até o topo da nuvem, atraindo cargas negativas para essa região. Com isso tais cargas negativas são retidas pelas partículas

presentes na fronteira da nuvem, formando uma espécie de blindagem eletrostática nesse local. Além disso, devido ao transporte por correntes descendentes presentes nas regiões de fronteira as gotículas de água e cristais de gelo com íons negativos são transportadas até a base da nuvem, atraindo cargas positivas para essa região, e assim formando um dipolo positivo (WILLIAMS, 1989). Todavia, embora existam evidências que esse mecanismo ocorra, há algumas contradições ainda não resolvidas (VONNEGUT, 1991): i) não há evidências que íons positivos originados por raios cósmicos sejam suficientemente capazes de eletrificar as nuvens e ii) como um campo elétrico tão intenso não é capaz de romper a rigidez dielétrica no interior da nuvem. Dessa forma, essa teoria ainda necessita de revisões conceituais que fundamentem com mais precisão a ocorrência desse processo (MATTOS, 2009).

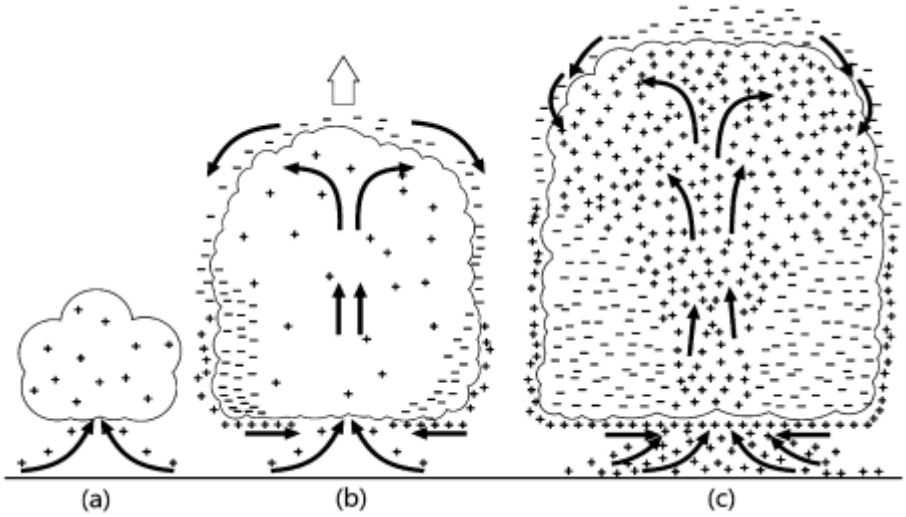

Figura 3 - Mecanismo de carregamento convectivo onde em (a) cargas positivas são transportadas até o interior da nuvem, (b) cargas positivas adicionais são introduzidas através da base da nuvem e uma camada de blindagem é formada por cargas negativas nas regiões fronteiriças que se estende desde o topo até a base da nuvem e (c) o menor acúmulo de cargas negativas aumenta a intensidade do campo elétrico para uma grandeza suficiente para gerar descargas coronas positivos de objetos em solo. Fonte: Saunders (2008).

A partir dos processos de eletrificação da nuvem, o campo elétrico gerado quando excedente pode romper a rigidez dielétrica do ar e iniciar a formação dos relâmpagos no interior das nuvens. Os relâmpagos assim formados são definidos como descargas atmosféricas de grande corrente elétrica (em média 30000 A) e comprimento de quilômetros (UMAN; KRIDER, 1989). O Brasil, por exemplo é um dos países com as maiores incidências de relâmpagos no mundo. Estima-se que ocorram aproximadamente 90 milhões de relâmpagos anualmente (ODA *et al.*, 2022), principalmente devido à sua localização tropical e vasta extensão territorial (PINTO JR, 2005). De acordo com Cardoso *et al.* (2014), somente os relâmpagos no Brasil são responsáveis por causar prejuízos de aproximadamente

R$500.000.000,00 aos setores de energia, telecomunicação e industrial, e estima-se que aproximadamente 120 pessoas morrem anualmente atingidas pelos relâmpagos.

Desde o lançamento do primeiro satélite meteorológico TIROS-1 (do inglês, *Television Infrared Observation Satellite*) no início da década de 60 houve grande avanço na compreensão sobre as tempestades. A partir dos anos 60 diversos estudos sobre a relação entre as propriedades do topo das nuvens e a ocorrência de relâmpagos foram possíveis devido ao surgimento dessas tecnologias (PURDOM *et al.*, 1996). Atualmente, devido à alta frequência temporal (~ 10-30 min) dos satélites geoestacionários, esses são mais utilizados do que satélites de orbita polar. Através da energia emitida pelo topo das nuvens, que compreende o espectro do infravermelho (comprimento de onda em torno de 10,2 μm) é possível determinar a temperatura da sua parte mais elevada, estimar o seu tamanho e a taxa de expansão da área da tempestade. Os satélites geoestacionários mais utilizados sobre a América do Sul são os lançados *pela National Aeronauticsand Space Administration* (NASA) denominados *Geoestationary Operational Envirionmental Satellite* (GOES). Entre 2006 e 2019 esteve operacional sobre as Américas o satélite GOES-13, o qual será utilizado no presente estudo.

Além dos sensores a bordo de satélites orbitando o planeta, outras ferramentas de sensoriamento remoto, como os sensores em superfície são capazes de monitorar com alta eficiência os fenômenos meteorológicos, como os relâmpagos. Assim é possível associar os dados provenientes de satélites com as informações das redes de relâmpagos em solo. A combinação das propriedades físicas das nuvens inferidas por satélites com informações de relâmpagos pode aprofundar a compreensão da formação e propagação dos relâmpagos e associá-los à física das nuvens (MATTOS, 2009). Dessa forma, as redes de monitoramento e detecção de relâmpagos em superfície estimam através das ondas eletromagnéticas emitidas, a localização, data, polaridade e o pico de corrente dos relâmpagos. Atualmente o Brasil abrange seis redes de relâmpagos: i) Rede Integrada Nacional de Detecção de Descargas Atmosféricas (RINDAT) possui cerca de 35 sensores capazes de monitorar relâmpagos nuvem-solo (NS) espalhados pelas regiões Centro-Oeste, Sudeste e Sul do país; ii) Sistema Brasileiro de Detecção de Descargas Atmosféricas (BrasilDAT) que integra nacionalmente cerca de 75 sensores da rede *Earth Networks Total Lightning Network* (ENTLN) espalhados pelo Brasil capazes de detectar relâmpagos do tipo intra-nuvem (IN) e NS; iii) *Sferics Timing and Ranging Network* (STARNET) a qual é caracterizada pelo método de detecção que abrange longas distâncias por meio de *Very Low Frequency* (VLF), entre 7 e 15 kHz e possui um total de 9 sensores instalados no Brasil; iv) *Lightning Network* (LINET) instalada no estado de São Paulo; v) *Global*

Lightning Dataset (GLD360) e vi) *Worldwide Lightning Location Network* (WWLLN), ambas operando com sensores de longo alcance (ODA *et al*., 2022).

Ao longo das últimas décadas diversos estudos sobre o comportamento espacial e temporal dos SCM na América do Sul (DURKEE e MOTE 2009; MACHADO et al. 1998; MARTINS *et al*. 2017; SALIO; NICOLINI; ZIPSER, 2007; VELASCO; FRITSCH, 1987; ZIPSER *et al*. 2006) e sua relação com relâmpagos (HAHN, 2021; LIMA, 2005; MACEDO *et al*., 2005; MONTEIRO *et al*., 2021; PETERSON *et al*., 2020; SPERLING, 2018) foram publicados. Esses estudos, de forma geral apresentam resultados que indicam uma boa relação entre o tamanho, temperatura e taxa de crescimento das tempestades com a produção de relâmpagos. Por exemplo, os resultados de Mattos e Machado (2011) mostraram que sistemas convectivos grandes associados a baixas temperaturas do topo da nuvem apresentavam maiores ocorrências de relâmpagos. Semelhantemente, Mecikalski *et al*. (2013) observaram uma tendência no aumento da área das tempestades a partir do estágio do ciclo de vida que possui propriedades capazes de fomentar o carregamento elétrico da nuvem e assim propiciar o início dos primeiros relâmpagos.

Com o avanço computacional, e com a utilização da ampla gama de dados meteorológicos disponíveis tornou-se possível uma maior compreensão sobre a complexidade envolvida na ocorrência de fenômenos naturais, como os relâmpagos. Nesse âmbito destaca-se a inteligência artificial (IA). Winston (1992, p. 5) define o termo "inteligência artificial" como: "o estudo dos cálculos que permitem perceber, raciocinar e agir". Ou seja, a partir de formulações matemáticas é possível entender o comportamento, os padrões e as tendências de uma série de informações que, posteriormente permitem tomar decisões. Dentre as subdivisões da IA, destaca-se o "aprendizado de máquina" (do inglês, *machine learning-ML*). O *ML* foi proposto primeiramente por Samuel (1959) que mostrou a capacidade dos computadores em aprender sem serem explicitamente programados. Com o avanço técnico-científico ao longo dos anos os modelos e algoritmos de aprendizagem foram sendo aprimorados e desenvolvidos de formas e para objetivos diferentes, e geralmente, se enquadram em duas categorias (Figura 4): aprendizado supervisionado e não supervisionado. O primeiro é caracterizado por possuir para cada valor previsto, um valor observado associado, ou seja, o objetivo do algoritmo é ajustar uma relação estatística entre os preditores (variáveis associadas ao que deseja prever) e a resposta (aquilo que irá ser previsto), a partir de observações já conhecidas. Já o aprendizado não supervisionado descreve uma situação mais desafiadora. Para cada valor (numérico ou classificador) das variáveis preditoras não há nenhuma observação ou resposta associada, isto é, o objetivo do algoritmo é prever uma nova categoria a partir das variáveis preditoras. Nesse

caso, refere-se como uma situação não supervisionada pois não há variáveis de resposta que possam supervisionar a análise (JAMES *et al*., 2013). Dessa maneira, devido a sua versatilidade de aplicação e capacidade de tomada de decisão o aprendizado de máquina vem sendo empregado em diversas áreas nas últimas décadas, como por exemplo na medicina (DARCY *et al*., 2016; DEO, 2015; HANDELMAN *et al*., 2018; RAJKOMAR *et al*., 2019; SIDEY-GIBBONS; SIDEY-GIBBONS, 2019), na educação (ALENEZI *et al*., 2020; KUCAK et al., 2018; SHAH *et al*., 2021), no esporte (HORVAT; JOB, 2020; RICHTER *et al*., 2021; ROSSI *et al*., 2021), em finanças (CULKIN; DAS, 2017; DIXON *et al*., 2020; GOODELL *et al*., 2021), entre inúmeras outras áreas.

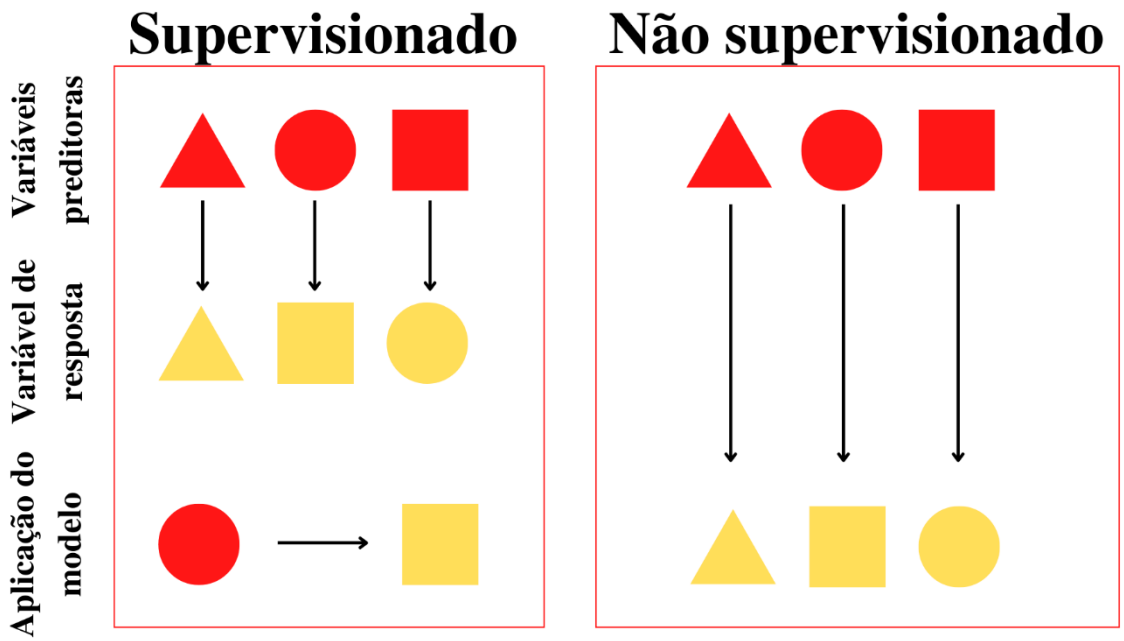

Figura 4 - Representação de modelo supervisionado (à esquerda) e não supervisionado (à direita). Modelos supervisionados ou de classificação são caracterizados por possuírem variáveis de resposta associadas às variáveis preditoras, enquanto modelos não supervisionados caracterizam-se pela ausência de variáveis de resposta necessitando da aplicação de mecanismos de associação e agrupamento para realizar a previsão.

Outra área que vem sendo bastante explorada por técnicas de aprendizagem de máquina, é a meteorologia e em suas diversas subaéreas como: agrometeorologia (LIAKOS *et al*., 2018; KELLEY *et al*., 2020; YING *et al*., 2020), qualidade do ar (NAIR *et al*., 2021; STIRNBERG, 2021; LIU *et al*., 2022), radiação solar (LI *et al*., 2016; VOYANT, 2017; ZHOU *et al*., 2021), previsão do tempo (HOLMSTROM *et al*., 2016; SCHER; MESSORI, 2018; BOCHENEK, 2022). Além disso, outro setor que se destaca é o de aplicação de modelos de aprendizagem de máquina para analisar sistemas convectivos. Por exemplo, Jergensen (2020) utilizou como variáveis preditoras dados de radar e dados de radiossondagem para descrever a atmosfera

próxima a uma célula convectiva. Como variáveis de resposta foram utilizadas a classificação das células convectivas identificadas como tempestades severas e não severas. A partir dessas informações e aplicação de modelos supervisionados (regressão logística, florestas com aumento de gradiente, florestas aleatórias e máquina de vetores de suporte, sendo respectivamente os dois primeiros explicados na seção 2 do presente estudo) foi observado em seus resultados diferentes precisões de acerto entre os modelos. O modelo que opera por florestas com aumento de gradiente apresentou os melhores resultados, evidenciando que a performance de sistemas de previsão varia de acordo com as habilidades de compreensão e fragilidades de cada algoritmo. Em contrapartida, os resultados de Zhou *et al*. (2020) mostraram que a partir da união de: i) dados de temperatura de brilho das nuvens do canal infravermelho do satélite Himawari-8, ii) dados de refletividade de radares meteorológicos e iii) dados de relâmpagos provenientes de redes em solo é possível construir um sistema eficiente de previsão de relâmpagos visando auxiliar a previsão de curtíssimo prazo (no inglês, *nowcasting*). Em consistência com esses resultados, Monteiro *et al*. (2021) mostraram que o início do estágio de maturação das tempestades está associado aos momentos de máxima atividade de relâmpagos. Estudos recentes como Drugan e Preston (2022) utilizaram dados de radares polarimétricos e dados de relâmpagos para desenvolver um modelo capaz de identificar os tipos de hidrometeoros dentro de um sistema convectivo e inferir a ocorrência de relâmpagos. Seus resultados mostraram que o algoritmo de melhor desempenho foi aquele que utilizou valores de refletividade horizontal abaixo do limiar de 40 dBZ na isoterma de -15 ºC após a fase de dissipação do *graupel*.

Em contrapartida, Cintineo *et al*. (2022) utilizaram uma arquitetura de rede neural convolucional com dados de temperatura de brilho do canal infravermelho e relâmpagos do sensor *Geostationary Lightning Mapper* (GLM) a bordo do satélite GOES-16 para fins de treinamento e validação do algoritmo. Os autores desenvolveram um produto denominado *"LightningCast"* capaz de fornecer previsões para o desenvolvimento e deslocamento de tempestades atribuindo sua atividade elétrica. Sendo assim, é possível notar que a construção de sistemas que visam predizer a severidade das tempestades têm sido enfoque de diversos estudos nos últimos anos.

Tendo em vista o que foi apresentado, é notável que a associação entre a física das nuvens convectivas e suas características elétricas juntamente com a aplicação de modelos de aprendizagem de máquina podem conceber uma maior compreensão dos padrões naturais desses fenômenos. Entretanto, no Brasil ainda inexistem ferramentas que possibilitem prever as características elétricas das tempestades. Dessa forma, este presente estudo tem como

finalidade analisar e avaliar, a partir de dados de satélite e de redes de monitoramento de relâmpagos em solo, a relação entre as propriedades físicas e elétricas dos sistemas convectivos ocorridos próximos ao estado de São Paulo entre 2013 e 2017. Com base nesta investigação tem-se como objetivo adicional aplicar ferramentas de aprendizagem de máquina com o intuito de criar um sistema inteligente que possa prever, com eficiência, a ocorrência de relâmpagos a partir das propriedades físicas (tamanho, taxa de expansão e temperatura) das tempestades.

2. Dados e metodologia

Esta seção apresentará a descrição dos dados de satélite e relâmpagos utilizados, o processamento envolvido na identificação e rastreamento das tempestades e o acoplamento com as informações de relâmpagos. Por fim será apresentada a utilização das técnicas de *ML* empregadas neste trabalho.

2.1 Imagens de satélite

No presente estudo o período de análise compreende janeiro de 2013 a dezembro de 2017. Foram utilizados dados no formato binário do canal 4 do infravermelho (10,7 μm) do satélite GOES-13 em formato retangular abrangendo a América do Sul. Esse canal pertence ao sensor GOES IMAGER presente no satélite que contém cinco canais espectrais, sendo: visível (0,65 μm) e infravermelho (3,9 μm, 6,55 μm – conhecido como canal do vapor d'água, 10,7 μm e 13,35 μm). No caso do visível, a resolução espacial utilizada é de 1 km, enquanto para o infravermelho é de 4 km, enquanto a resolução temporal é de 30 minutos (OSCAR, 2022). Os dados foram adquiridos pelo banco de dados do Centro de Previsão do Tempo e Estudos Climáticos (CPTEC) do Instituto Nacional de Pesquisas Espaciais (INPE) presente no website <http://ftp.cptec.inpe.br/goes/goes13/retangular_4km/>.

2.2 Relâmpagos

A rede *Earth Networks Total Lightning Network* (ENTLN) é uma rede de detecção de relâmpagos IN e NS que possui mais de 1800 sensores no mundo todo. Atualmente, no Brasil a rede ENTLN conta com dezenas de sensores distribuídos nas Regiões Sul, Sudeste (10 estações somente no estado de São Paulo) e parte das Regiões Centro-Oeste e Nordeste do Brasil, com somente uma estação na Região Norte como indicado na Figura 5. A rede ENTLN detecta os relâmpagos através do monitoramento de pulsos eletromagnéticos gerados quando há a ocorrência de relâmpagos. Dessa forma, por meio de sensores instalados na superfície

terrestre, é possível identificar relâmpagos por meio de ondas na frequência de radiação de *Low Frequency* (LF), *Mid Frequency* (MF, a qual é a faixa utilizada pela operação da ENTLN) e *Very High Frequency* (VHF), com eficiência de 90% e precisão menor que 1 km, permitindo um maior detalhamento de momentos específicos desse tipo de relâmpago (EARTH NETWORKS, 2022). Os dados utilizados correspondem à ocorrência de descarga de retorno (do inglês, *strokes*). Quando mencionada ao longo do texto a palavra relâmpago, implicitamente estará referindo-se a descarga de retorno. Os dados de relâmpagos para fins de pesquisa foram fornecidos pela empresa CLIMATEMPO.

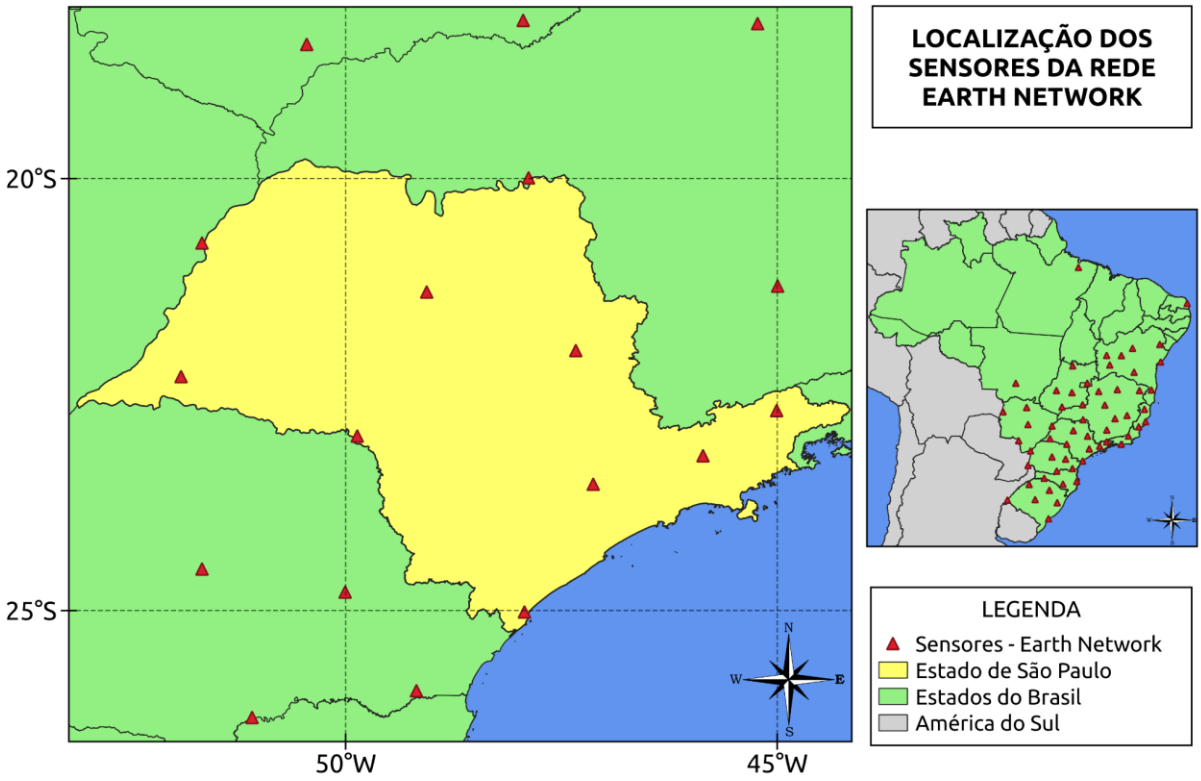

Figura 5 - Localização dos sensores da rede *Earth Networks Total Lightning Network* no estado de São Paulo e no Brasil em 2017.

2.3 Identificação e rastreamento das tempestades

Os SCM foram identificados e rastreados empregando-se o algoritmo *Forecast and Tracking of Active Convective Cells* (ForTraCC, VILA *et al*. 2008). O ForTraCC é um algoritmo que permite estimar as propriedades físicas e radiativas dos sistemas convectivos e prever sua evolução ao longo do tempo, aplicando-se limiares de temperatura de brilho (TB) e tamanho dos SCM (VILA *et al*. 2008). Dessa forma, os dados binários referentes ao canal

infravermelho (10,7 µm) do satélite GOES-13 foram utilizados para o processamento do ForTraCC e para o presente estudo foram aplicados os seguintes critérios:

i) TB menor que 235 K e 210 K, para identificação dos SCM e frações convectivas, respectivamente.
ii) Área do sistema convectivo maior que 75 pixels, isto é, (4 km x 4 km) x 75 pixels = 1200 km².

Após finalizado o processamento, o algoritmo forneceu uma série temporal de dados com as principais características radiativas e morfológicas das tempestades identificadas a partir da metodologia empregada, tais como: localização (latitude e longitude do centro geométrico), tamanho (em *pixels*), taxa de expansão normalizada (A_e, em 10^{-6} segundos^{-1}), temperatura média (T_{med}, em K) e mínima (T_{min}, em K), temperatura mínima média de brilho do kernel de 9 pixels (T_{min9}, em K), fração convectiva, excentricidade, ângulo de inclinação, entre outros parâmetros.

A partir da área do SCM fornecida pelo ForTraCC calculou-se o Raio Efetivo (R_e), através da seguinte expressão:

$$R_e = \sqrt{\frac{Apxl \,.Asc}{\pi}} \quad (1)$$

O R_e corresponde ao raio de um círculo cuja área seja equivalente à área do SCM, expresso em quilômetros. Onde, A_{pxl} é a área de um *pixel* do satélite GOES-13 (16 km^2) no canal infravermelho e A_{sc} é a área total do SCM (MACHADO *et al.*, 1998; MATTOS; MACHADO, 2011).

A propriedade A_e (Equação 2) é expressa em 10^{-6} segundos^{-1} e representa a taxa de expansão dos SCM, sendo crescimento ($A_e > 0$), decaimento ($A_e < 0$) e maturação ($A_e \sim 0$), sendo determinada pela seguinte expressão:

$$Ae = \frac{1}{\bar{A}}\left(\frac{\partial A}{\partial t}\right) \quad (2)$$

Onde, o parâmetro $\bar{A}$ corresponde à área média do SCM entre duas imagens consecutivas, ∂A é a variação da área entre essas imagens e ∂t o intervalo de tempo (em segundos) entre as duas imagens. A propriedade T_{min9} representa a média da temperatura entre os nove pixels que

possuem as menores temperaturas do topo do SCM. Os pixels mais frios podem estar em localidades diferentes do topo da nuvem.

A partir do banco de dados de tempestades rastreadas pelo algoritmo ForTraCC foram necessários aplicar alguns filtros para a escolha das tempestades a serem avaliadas. Por exemplo, foram analisadas apenas as tempestades que:

i) Iniciaram-se e dissiparam-se dentro da região de estudo (estado de São Paulo destacada em amarelo na Figura 6).
ii) Apresentaram pouca falta de dados a respeito do seu ciclo de vida (menos que 51 % de faltas de dados válidos ao longo do ciclo de vida, caso contrário, o sistema convectivo é eliminado).
iii) Iniciaram-se (não iniciaram como resultado de uma divisão de uma tempestade) e dissiparam-se (não dissiparam devido a união de tempestades) espontaneamente.
iv) Não apresentaram divisão ou união entre tempestades ao longo do ciclo de vida.

Como discutido por Machado e Laurent (2004), estas limitações são essenciais para assegurar que o crescimento inicial das tempestades está associado à própria dinâmica interna desses sistemas. O processamento anteriormente discutido foi aplicado para os 5 anos de dados de satélite (2013-2017) para o estado de São Paulo, e com isso foram identificados 2304 sistemas convectivos (Figura 6) que totalizaram em 57446 etapas do ciclo de vida analisados. Importante ressaltar, que os termos “tempestade” ou “sistema convectivo” a partir desse momento do presente texto refere-se a uma etapa do ciclo de vida de um sistema convectivo e não ao seu ciclo de vida total.

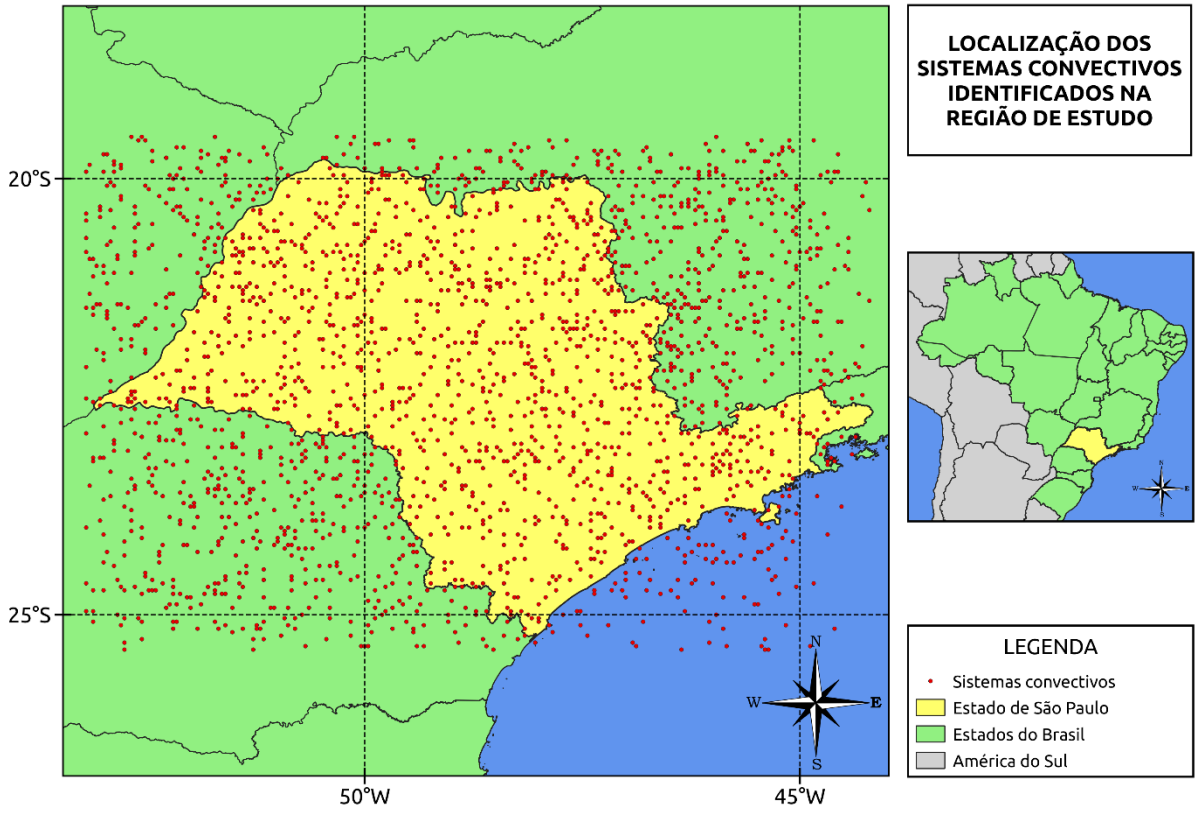

Figura 6 - Localização dos 2304 sistemas convectivos identificados pelo algoritmo ForTraCC na região de estudo.

A combinação das propriedades físicas e morfológicas dos SCM identificados pelo ForTraCC com as informações de relâmpagos da ENTLN foi realizada considerando um intervalo de tempo de 15 minutos antes e 15 minutos depois de cada imagem de satélite. Para cada pixel de 4 km por 4 km pertencente à tempestade foi contabilizado o número de relâmpagos IN, -NS e +NS. A partir da contabilização dos relâmpagos para cada pixel contabilizou-se o número total de relâmpagos IN, -NS e +NS pertencente a cada tempestade. Neste trabalho as tempestades foram divididas em dois grupos principais: i) aquelas com relâmpagos (IN+NS) e ii) aquelas sem relâmpagos (número total de relâmpagos igual a zero). Ao todo foram identificadas 24384 e 33062 tempestades com e sem relâmpagos, respectivamente.

2.4 Análise da relação entre as propriedades das tempestades e relâmpagos

Com o intuito de identificar as principais características físicas dos sistemas convectivos com e sem relâmpagos foram realizadas análises exploratórias dos dados. Para isto foram empregados histogramas, gráficos de dispersão (*boxplot* e *violinoplot*) e mapas de correlação

de Pearson evidenciando quais propriedades físicas são mais fortemente correlacionadas à ocorrência de relâmpagos.

O uso de diferentes tipos de gráficos auxilia tanto na visualização do comportamento dos dados quanto na interpretação dos resultados. Parte das análises realizadas neste trabalho foram através da junção de dois gráficos: gráfico *boxplot* ou diagrama de caixa e *violinoplot*. Os diagramas de caixa são uma excelente ferramenta para comparar a frequência de amostras fornecendo de maneira intuitiva informações acerca da centralização e distribuição dos dados ao longo de um intervalo. Por exemplo, é possível visualizar a concentração e dispersão em quartis, assim como os limites máximo e mínimo dentro de um intervalo de valores e a presença de valores que anômalos (*outliers*). Já o *violinoplot* permite a visualização rápida da distribuição de frequências de um conjunto de dados evidenciando a densidade dos dados, isto é, a concentração das maiores e menores frequências ao longo de determinados valores. A Figura 7 mostra a ilustração desse tipo de gráfico e seus recursos.

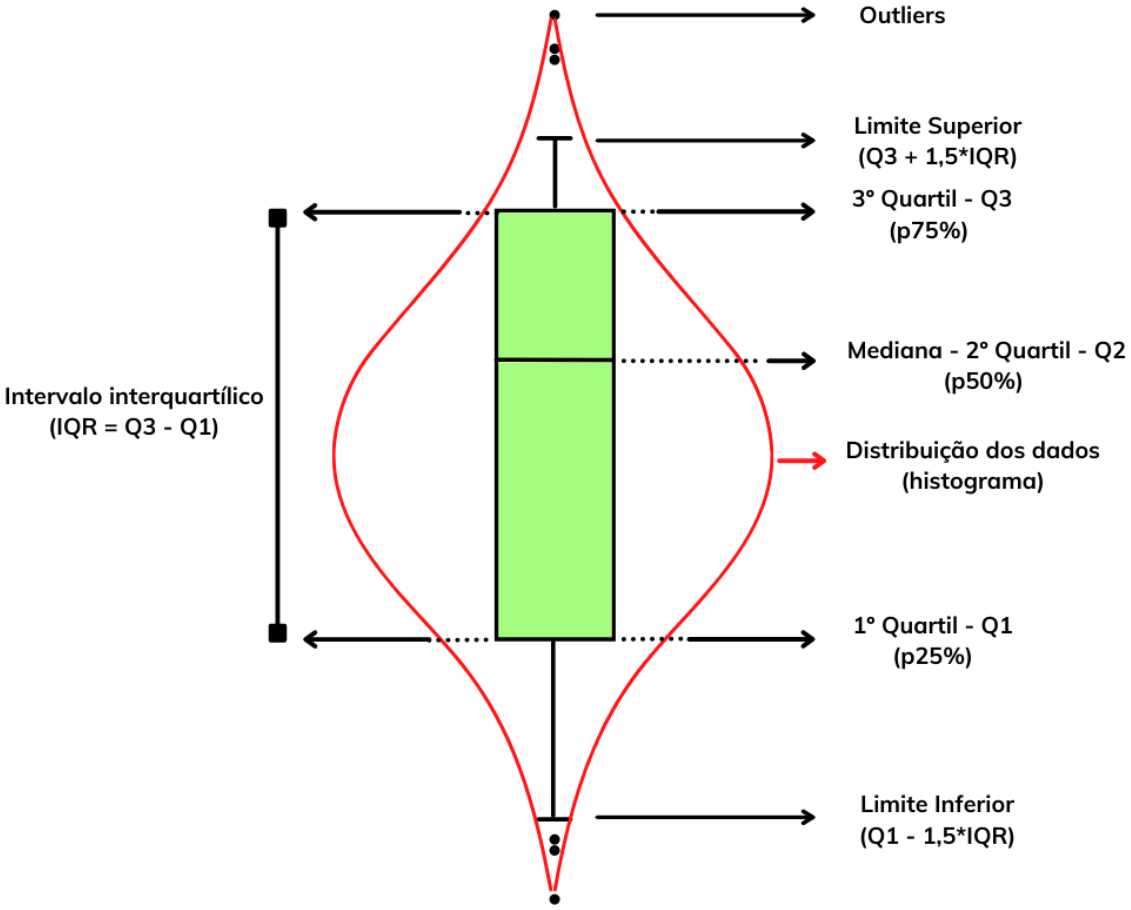

Figura 7 - Representação da união de gráficos *boxplot* (destacado na cor verde) e *violinoplot* (destacado na cor vermelha).

Os *outliers* são os dados que se diferenciam drasticamente de todos os outros e estão acima (abaixo) do limite superior (inferior), os quais são definidos como o máximo (mínimo) valor da amostra dentro do critério de seleção de *outliers*. Além disso, traz três valores que dividem o conjunto de dados em quartis, sendo:

- ➔ Primeiro Quartil (Q1): separa os 25% dos valores inferiores dos 75% dos valores superiores.
- ➔ Segundo Quartil (Q2): separa os 50% dos valores inferiores dos 50% dos valores superiores (mediana).
- ➔ Terceiro Quartil (Q3): separa os 75% dos valores inferiores dos 25% dos valores superiores.

Já o Intervalo Interquartílico (IQR) é definido como a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1), sugerindo, dessa maneira, a variabilidade da amostra, uma vez que quanto maior o IQR maior a variabilidade.

$$IQR = Q3 - Q1 \tag{3}$$

O método de correlação de Pearson (r) mensura a direção e a relação linear entre duas variáveis quantitativas a partir do compartilhamento de variância (MOORE, 2007).

$$r = \frac{1}{n-1} \Sigma \left(\frac{xi - \underline{X}}{Sx}\right)\left(\frac{yi - \underline{Y}}{Sy}\right) \tag{4}$$

Onde n é o número de total de dados, xi é um dado qualquer da variável x, $\underline{X}$ a média da variável x e Sx o desvio padrão da variável x. A mesma caracterização adequa-se para yi, $\underline{Y}$ e Sy. Desse modo, o coeficiente de Pearson se ajusta entre -1 e 1, uma vez que o sinal sugere a tendência de relacionamento (diretamente ou inversamente proporcionais), a proximidade do valor zero revela fraca correlação e os extremos indicam forte correlação entre as variáveis analisadas (FIGUEIREDO FILHO; SILVA JÚNIOR, 2009). No entanto, geralmente valores exatos (-1, 0 ou 1) são difíceis de serem obtidos na prática. Dessa forma, Dancey e Reidy (2005) propõem que valores de *r* entre 0,1 à 0,3 seja considerado fraco, entre 0,4 e 0,6 moderado e acima de 0,7 seja considerado forte. Já Cohen (1988), sugere que valores de *r* entre 0,1 e 0,29 é considerado fraco, entre 0,3 e 0,49 é moderado e entre 0,5 e 1 é considerado forte.

Como os resultados de Mattos e Machado (2011) sugerem que o tamanho, a taxa de expansão e a temperatura mínima dos SCM são os principais parâmetros físicos associados à ocorrência de relâmpagos, no presente estudo as variáveis utilizadas nas análises foram: Raio

Efetivo (km), Taxa de Expansão Normalizada ($10^{-6}s^{-1}$), Temperatura Média (ºC), Variação da Temperatura Média (ºC), Temperatura Mínima (ºC), Variação da Temperatura Mínima (ºC), Temperatura Mínima do *Kernel* de 9 *pixels* (ºC) e Variação da Temperatura Mínima do *Kernel* de 9 *pixels* (ºC).

2.5 Aplicação dos modelos de *Machine Learning*

Os modelos de *ML* foram aplicados com o objetivo de analisar a capacidade de estimar a ocorrência de relâmpagos. Para isto a capacidade de estimativa de relâmpagos com modelos de *ML* foi aplicada em duas vertentes: i) estimar a quantidade específica do total de relâmpagos de um SCM e ii) ocorrência e não ocorrência de qualquer quantidade de relâmpago, ou seja, classificação binária (SCM com relâmpago = 1 e SCM sem relâmpago = 0).

Neste trabalho, partindo de uma metodologia de aprendizado supervisionado, foram utilizados os seguintes modelos de *ML*: a) Regressão Linear Múltipla, b) Regressão Logística, c) Árvore de Decisão e d) Floresta Aleatória. Os modelos serão descritos em detalhes na seção seguinte. De modo geral, os dados de entrada não se diferem para cada tipo de modelo, mas sim, a forma de aplicação (dados discretos ou binarizados) a partir do processamento e metodologia adotada por cada um deles. Por exemplo, para o modelo de regressão linear, 70% (17069 eventos) dos dados de tempestades com relâmpagos foram utilizados para aprendizagem do modelo e 30% (7315 eventos) usados para fins de validação, isto é, se a previsibilidade de relâmpagos pelo modelo está coerente ou não. Em contrapartida, os modelos de classificação (regressão logística, árvore de decisão e floresta aleatória) foram utilizados a união dos dados de tempestades com e sem relâmpagos (dados binarizados, sendo o valor 0 para tempestades que não tiveram relâmpagos e o valor 1 para tempestades que tiveram ao menos um relâmpago), uma vez que 70% (40212 eventos) são utilizados para a aprendizagem do modelo e 30% (17234 eventos) para validá-lo. Em seguida, foram empregadas métricas para avaliar a qualidade das estimativas de cada modelo. Dessa maneira foram utilizados dois tipos de métricas para avaliar a qualidade das estimativas: a) Relatório de Classificação e b) Matriz de Confusão. A Figura 8 mostra um fluxograma representando as etapas de manipulação e aplicação dos modelos de *ML*.

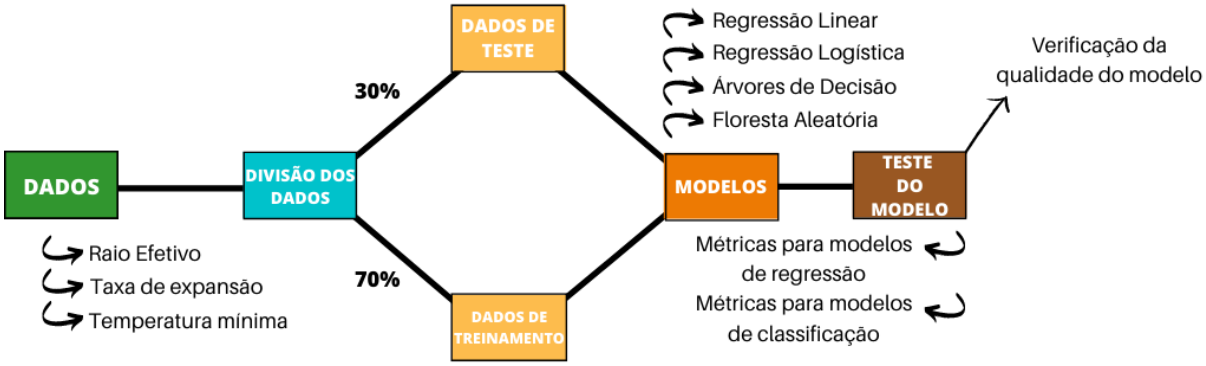

Figura 8 - Fluxograma dos processos de *Machine Learning*.

2.5.1 Modelo de Regressão Linear

Os modelos de regressão linear consistem em métodos que permitem ajustar a relação entre duas ou mais variáveis em uma reta a fim de encontrar níveis de correlação que possibilitem estimar variáveis através de variáveis preditoras (GROSS; GROß, 2003). Dessa maneira pode-se denotar a variável estimada como Y e as variáveis preditoras como X:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i + \varepsilon \qquad (5)$$

onde, α e β são coeficientes de regressão e ε é o erro atrelado a variabilidade que não pode ser expressa exatamente como uma combinação linear pelas variáveis preditoras (GUNST; MASON, 2018).

No presente trabalho, o modelo de regressão linear foi aplicado para as propriedades físicas das tempestades (variáveis preditoras) para predizer a quantidade aproximada de relâmpagos (variável estimada) que um sistema convectivo poderá apresentar.

2.5.2 Modelo de Regressão Logística

A regressão logística é um modelo estatístico utilizado para estimar a probabilidade de determinado evento ocorrer. De acordo com Hosmer e Lemeshow (1989) esse modelo permite relacionar η variáveis independentes ($X_1, X_2, \ldots, X\eta$) a uma variável dependente Y, diretamente associada à sua estimativa da probabilidade de ocorrência. Desse modo, a curva de regressão se ajusta em um intervalo de 0 (nenhuma probabilidade de ocorrência) a 1 (100% de probabilidade de ocorrência). Em outras palavras, a regressão logística adapta a regressão linear

à uma classificação binária aplicando uma função sigmóide à saída do modelo (JERGENSEN *et al*., 2020). A função logística pode ser definida como:

$$p(X) = \frac{e^{\beta 0 + \beta 1 X}}{1 + e^{\beta 0 + \beta 1 X}} \quad (6)$$

onde, *p(X)* é a probabilidade de ocorrência variando entre 0 e 1 e $\beta\eta$ são os coeficientes estimados do modelo. Em vista disso, a aplicação do modelo de regressão logística no presente estudo pretende predizer, a partir das propriedades físicas de determinada tempestade a probabilidade da ocorrência de relâmpagos. Dessa forma, ao aplicar o algoritmo, é possível classificar como sendo potencialmente elétrica ($P > 0{,}5$) ou não ($P < 0{,}5$).

2.5.3 Modelo de Árvore de Decisão

Os modelos de árvores de decisão permitem dividir um conjunto de dados em várias seções com níveis de importância associados. Isto significa que o algoritmo identifica quais características das variáveis preditoras são mais relacionadas a determinada classe (variável qualitativa). Esse modo de aprendizagem ocorre com base em "ganhos de informações", estabelecendo que o modelo realize as divisões de decisão a partir das características mais relevantes que predizem a classificação de certa ordem (JAMES *et al.*, 2013). Tais divisões são definidas dentro do conceito de árvore de decisão, como: Nó, Raiz, Ramos e Folhas. Os Nós são os valores ou qualidades que dividem a árvore em atributos; a Raíz é o primeiro nó da árvore de decisão, isto é, o parâmetro que fornece o maior ganho de informação para a previsão; os Ramos são as possibilidades dos Nós e as Folhas são as decisões finais a partir dos ramos.

A Figura 9 mostra um exemplo do conceito de árvore de decisão. Neste exemplo fictício é fornecido um conjunto de dados com algumas condições meteorológicas e, ao final é desejado saber se irá chover ou não. Dessa forma, o algoritmo aprenderá quais características do conjunto de dados são mais importantes para predizer se haverá ou não a ocorrência de precipitação. O parâmetro que possui o maior ganho de informação é aquele que diz se o céu está nublado ou não, dessa forma esse é o primeiro nó ou raiz da árvore de decisão. Os ramos (sim ou não) abrem possibilidades para outros nós ("nível de instabilidade") ou folhas ("não irá chover") que abrem novas possibilidades até chegarem na decisão final que determinará a resposta final do

modelo.

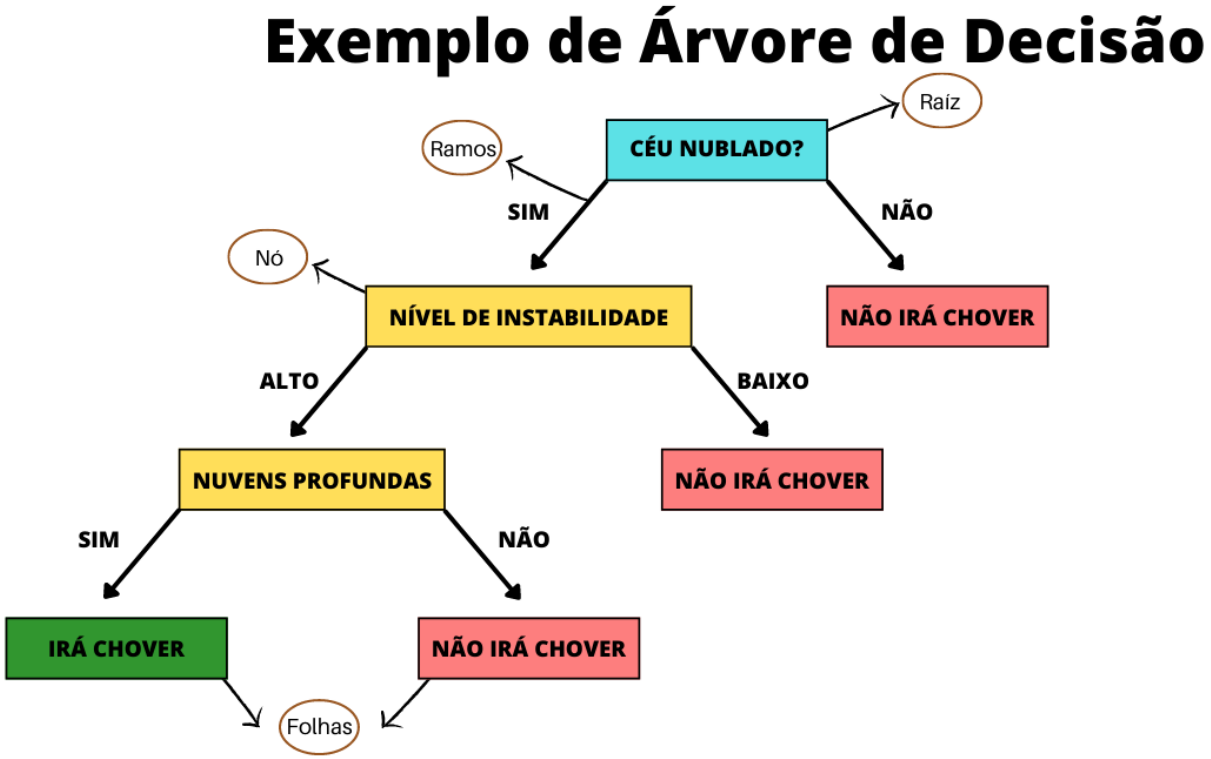

Figura 9 - Exemplo explicativo do conceito de árvore de decisão formada pela raiz (característica com maior ganho de informação, ramos (possibilidades de decisão), nós (características secundárias) e folhas (últimos desdobramentos que definirão a tomada de decisão).

Contudo de acordo com James *et al.* (2013) o uso de árvores de decisão possui algumas vantagens:

- ➔ São mais fáceis de explicar do que modelos de regressão.
- ➔ Dependendo da aplicação a tomada de decisão pelo método de árvores é mais eficiente do que métodos de regressão.
- ➔ As árvores podem manipular preditores qualitativos sem a necessidade de criar variáveis hipotéticas.
- ➔ São fáceis de explicar devido à sua estrutura gráfica de classificação e tomadas de decisão.

E possui algumas desvantagens:

- ➔ Normalmente as árvores de decisão não possuem o mesmo nível de precisão do que outros métodos de regressão e classificação.

➔ Para dados mais complexos, o emprego de modelos de floresta aleatória, as quais utilizam *n* árvores de decisão se torna mais eficiente do que a utilização de uma única árvore.

2.5.4 Modelo de Floresta Aleatória

Os modelos de Floresta Aleatória são definidos como um conjunto de árvores de decisão. No entanto, cada vez que uma divisão de uma árvore é considerada, é escolhida uma amostra aleatória de parte dos preditores (η) em relação ao total do número de preditores (σ). A amostra de η preditores é aproximadamente igual a raíz quadrada do total σ. Essa é uma forma crucial para evitar que as raízes (preditor mais forte do conjunto de dados) das árvores de decisões geradas sejam iguais em todos os casos, uma vez que isso não reduziria significativamente a variância diminuindo, dessa forma a precisão do prognóstico (JAMES *et al.*, 2013). A Figura 10 mostra um exemplo de como funciona as tomadas de decisões por florestas aleatórias, onde para esse caso, foram utilizadas duas árvores de decisão. As raízes das árvores são diferentes para atribuir ganhos de informação variados que, por sua vez, ditam as tomadas de decisões.

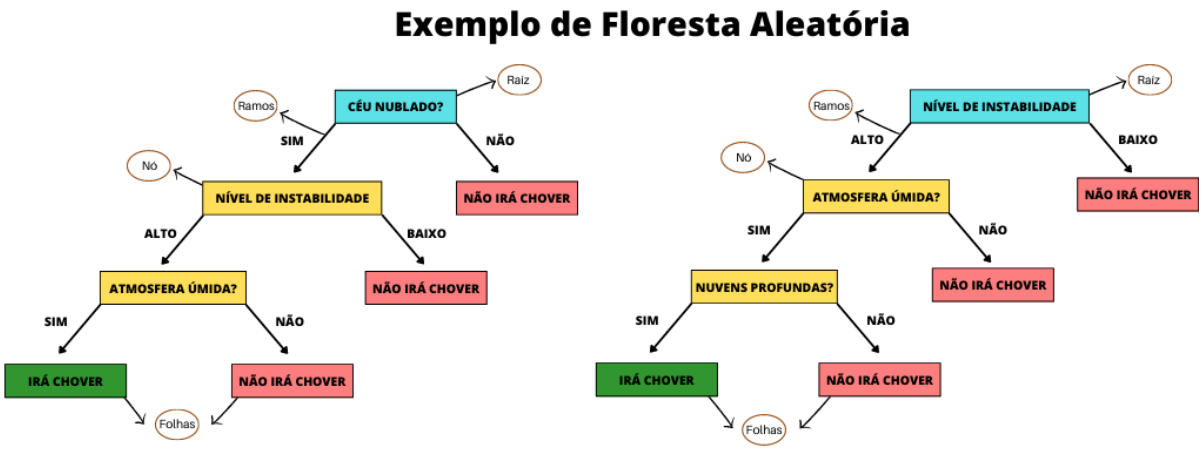

Figura 10 - Exemplo explicativo do conceito de floresta aleatória com duas árvores de decisão. Os primeiros nós ou raízes de cada árvore de decisão possuem pesos diferentes, assim, o ganho de informação varia atribuindo uma maior generalidade ao modelo.

2.6 Avaliação dos modelos de *Machine Learning*

A fim de qualificar a acurácia e precisão dos modelos de *ML* foram aplicadas métricas de avaliação para cada modelo. A utilização de métricas de avaliação adequadas é fundamental para entender o grau de sucesso e possíveis incertezas que modelos de *ML* podem apresentar.

Além disso, a avaliação da saída de um algoritmo depende do contexto e dos objetivos da análise tornando extremamente importante qual tipo de métrica qualificativa utilizar (FERRARI; SILVA, 2017). Em vista disso, as métricas utilizadas para os modelos de regressão diferem daquelas utilizadas para modelos de classificação. Ambas serão descritas nas seções seguintes.

2.6.1 Métricas para modelos de regressão

As métricas para modelos de regressão, de forma geral revelam o quanto o valor previsto desviou do valor real (conhecido também como valor observado). Sendo assim, consegue-se realizar uma estimativa quantitativa dos erros presentes no modelo. Existem cinco tipos de métricas mais utilizadas para avaliar modelos de regressão: Erro residual, Erro Absoluto Médio (da sigla em inglês, MAE), Erro Quadrático Médio (da sigla em inglês, MSE), Raiz do Erro Quadrático Médio (da sigla em inglês, RMSE) e R-quadrado ou R^2.

O Erro Residual (ER) consiste na diferença entre o valor previsto pelo modelo (y) e o valor real (ŷ). Sendo assim, os resíduos indicam a variação natural dos dados.

$$\text{RE} = yi - \hat{y}i \tag{7}$$

O Erro Absoluto Médio (MAE) é a média dos valores do ER, ou seja, fornece o erro médio do modelo. Em razão disso, quanto maior o valor do MAE pior é o modelo. No entanto, essa métrica não desconsidera severamente os *outliers* do modelo, diferentemente do MSE e RMSE.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} (yi - \hat{y}i) \tag{8}$$

O Erro Quadrático Médio (MSE) é a média dos valores do ER elevado ao quadrado. Semelhantemente ao MAE, quanto maior o valor do MSE pior é o modelo.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (yi - \hat{y}i)^2 \tag{9}$$

A Raíz do Erro Quadrático Médio (RMSE) é expressa pela raiz quadrada do MSE. A vantagem em utilizar o RMSE é que embora ele também apresente a acurácia do modelo, o resultado é expresso na mesma dimensão da variável analisada.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \quad (yi \; - \; \hat{y}i)^2} \tag{10}$$

O R-quadrado (R^2) é uma medida estatística que mostra o quão próximos os dados estão da linha de regressão ajustada, ao passo que expressa em porcentagem indicando a compreensão do modelo em relação a variabilidade dos dados em torno da média:

$$R^2 = \frac{SSE}{SST} \tag{11}$$

Onde, SSE é a soma dos quadrados explicada e SST é soma dos quadrados totais. Por exemplo, o valor de R^2 igual a 100% estabeleceria que o modelo explica toda a variabilidade dos dados ao redor da média.

2.6.2 Métricas para modelos de classificação

As métricas aplicadas em modelos de classificação se diferem daquelas aplicadas em modelos de regressão, uma vez que um modelo classificador procura decidir em qual classe uma observação melhor se encaixa. Dessa forma, as métricas para modelos de classificação buscam mensurar o quanto o modelo está divergente da classificação perfeita. No presente estudo serão aplicados seis métricas principais: Matriz de Confusão, Acurácia, Precisão, Sensibilidade (no inglês, *Recall*), Especificidade e F1-Score.

A matriz de confusão é amplamente utilizada para avaliar o comportamento de modelos de classificação supervisionados (CAELEN, 2017). É definida como uma matriz quadrática 2X2 (Tabela 1) que fornece o número de: Verdadeiros Positivos (VP), Falsos Positivos (FP), Falsos Negativos (FN) e Verdadeiros Negativos (VN).

As categorias (VP, FP, FN e VN) da matriz de confusão foram explicadas por Caelen (2017) e está baseada em um conjunto de dados de teste T = (Xi,Yi), onde i é a posição dos dados da amostra, a fim de prever determinado parâmetro (Ŷ) que é definido pela classe 0 (negativo) ou 1 (positivo). Com base nisso, o autor define uma função de perda que fornece a qualidade das previsões: $\Delta\, Y\; x\; \hat{Y} \rightarrow \; (VP, FP, FN \; e \; VN)$, onde Y é a classe observada (real) e Ŷ é a classe predita. Sendo assim, define-se:

- ➔ Quando Ŷ = 1 e Y = 1, então: Verdadeiro Positivo.
- ➔ Quando Ŷ = 1 e Y = 0, então: Falso Positivo.
- ➔ Quando Ŷ = 0 e Y = 1, então: Falso Negativo.

➔ Quando $\hat{Y} = 0$ e $Y = 0$, então: Verdadeiro Negativo.

Tabela 1 - Representação de uma matriz de confusão.

	OBSERVAÇÃO	
PREVISÃO	**SIM**	**NÃO**
SIM	Verdadeiro Positivo (VP)	Falso Positivo (FP)
NÃO	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Os parâmetros “Sim” e “Não” são referentes às classes das previsões (valores previstos) e às classes dos observados (valores reais). Se a classe real for a mesma que a predita, será considerada então um Verdadeiro Positivo ou Verdadeiro Negativo. Porém, se a classe real se diferenciar da classe predita, será considerada um Falso Positivo ou Falso Negativo. Em vista disso, algumas métricas de classificação utilizam as quantidades de VP, FP, FN e VN para apresentar outros valores de avaliação dos modelos.

A métrica denominada de Acurácia (do inglês, *accuracy)* fornece o quanto foi classificado corretamente pelo modelo. Por exemplo, se em um total de 100 observações, 75 foram classificadas de forma correta, a acurácia do modelo é de 0,75 ou 75%. Dessa forma, a acurácia é calculada através da razão entre a soma do número de acertos do modelo e a soma de todas as categorias:

$$\text{Ac} = \frac{VP + VN}{VP + VN + FP + FN} \tag{12}$$

A Precisão ou Taxa de Sucesso é uma métrica definida pela razão entre o número de previsões classificadas corretamente como positivos (VP) e o total de predições classificadas como positivas (VP e FP). Dessa forma, é possível observar o quanto das predições positivas (“sim”) foram classificadas corretamente.

$$\text{Pr} = \frac{VP}{VP + FP} \tag{13}$$

Já a Sensibilidade (no inglês, *Recall*), embora seja parecida com a métrica citada anteriormente permite identificar a taxa de acertos quando a observação foi “sim”. Desse modo,

é possível notar o quanto foi classificado corretamente como positivo levando em consideração o erro por FN.

$$\text{Rec} = \frac{VP}{VP + FN} \tag{14}$$

A Especificidade possibilita verificar a capacidade do modelo em prever com sucesso resultados classificados "não". Desse modo, é possível notar o quanto foi classificado corretamente como negativo levando em consideração o erro por FP.

$$\text{Esp} = \frac{VN}{VN + FP} \tag{15}$$

Por fim, a métrica F1 - *Score* (também conhecida como: *F-measure* ou *F-score*) é uma média harmônica que relaciona as métricas de Precisão e Sensibilidade. Em razão disso, se os valores de Precisão e Sensibilidade forem próximos a zero, implicará em um F1-*Score* baixo. Por outro lado, um valor de F1-*Score* alto sugere que o modelo é capaz de prever corretamente as predições consideradas positivas (precisão e sensibilidade alta).

$$\text{F1-Score} = \frac{2 * Prec * Rec}{Prec + Rec} \tag{16}$$

3. Resultados e discussões

Nesta seção será descrito os principais resultados encontrados sobre a relação entre as propriedades físicas dos sistemas convectivos com e sem relâmpagos, e também sobre o desempenho e eficiência da aplicação dos algoritmos de *ML* em prever as características elétricas das tempestades.

3.1 Análise da distribuição das propriedades das tempestades com e sem relâmpagos totais

Os resultados e análises a seguir foram realizadas considerando dois grupos de tempestades, os quais são representados pela cor vermelha (sem relâmpagos) e a cor preta (com relâmpagos). O primeiro grupo possui uma amostra de 33062 eventos e o segundo grupo, aquele com a presença de relâmpagos, 24384 eventos.

3.1.1 Raio Efetivo

O raio efetivo (km) é uma medida linear que considera o raio de um círculo, cuja área é a mesma do *cluster*, isto é, da área média de *pixels* conectados. Essa variável é utilizada devido à vasta variedade de formas que os sistemas convectivos apresentam (MACHADO *et al.*, 1998). Desse modo, a Figura 11 mostra o comportamento do raio efetivo em tempestades com e sem a ocorrência de relâmpagos. O gráfico à esquerda (Figura 11a) mostra a distribuição de frequências da propriedade física e o gráfico à direita (Figura 11b) mostra a dispersão dos dados (*boxplot*) e sua distribuição (violinoplot) ao redor dos valores da mediana.

A Figura 11a mostra que sistemas convectivos sem a presença de relâmpagos tendem a ser menores, uma vez que foram observados máximos valores de frequência relativa entre 0 a 50 km de raio efetivo. Além disso, observa-se uma área média de 38 km para esses sistemas convectivos enquanto para tempestades com relâmpagos a média é de 65 km. Esse fato, também pode ser constatado ao analisar a assimetria positiva na distribuição de frequências para tempestades com relâmpagos que evidencia a predominância desse grupo quando o raio efetivo é superior à 175 km. A Figura 11b evidencia que as maiores frequências dos valores de raio efetivo das tempestades em geral se encontram abaixo da mediana, aproximadamente 30 km para sistemas convectivos sem relâmpagos e 48 km para sistemas convectivos com relâmpagos. Em adição, nota-se que a amplitude do intervalo interquartílico, é maior para tempestades com relâmpagos evidenciando uma maior dispersão dos dados que pode ser em razão do grande número de *outliers* (valores que se diferenciam drasticamente dos outros).

Dessa maneira, nota-se que 75% do total de tempestades que produziram relâmpagos possuem raio efetivo abaixo de 75 km (valor do terceiro quartil). Sugerindo assim que grandes tempestades espacialmente desenvolvidas estão associadas com a ocorrência de relâmpagos. Além disso, nota-se que a predominância de sistemas convectivos com relâmpagos acima do limiar de 175 km sugere que quanto maior a tempestade maior a probabilidade de eletrificação da nuvem. Mattos (2009) estudou a relação entre as propriedades físicas e elétricas das tempestades no estado de São Paulo entre 2005 a 2007, e encontrou resultados semelhantes; i.e., mostrando que sistemas convectivos de grande desenvolvimento, em geral estão mais associados com a ocorrência de relâmpagos.

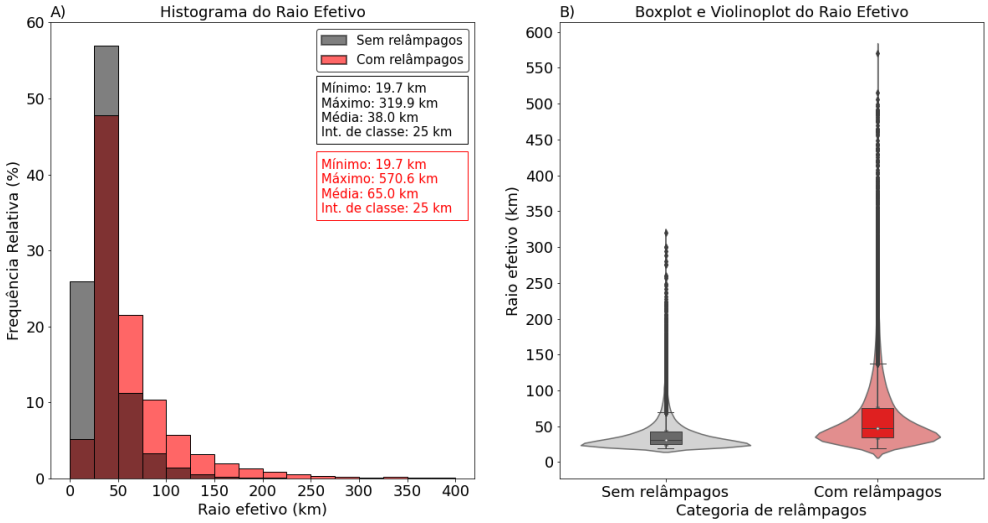

Figura 11 - Distribuição do raio efetivo (km) em sistemas convectivos com (cor vermelha) e sem (cor cinza) relâmpagos, sendo: A) histograma de frequência relativa (%) e B) gráfico *boxplot* e *violinoplot*.

3.1.2 Taxa de expansão em tempestades

A taxa de expansão normalizada (10^{-6}. s^{-1}) é um parâmetro utilizado para indicar o crescimento ou decaimento relativo de um sistema convectivo a partir da variação da sua área média em um intervalo de tempo (MACEDO *et al.*, 2004). A Figura 12 mostra a taxa de expansão para sistemas convectivos com e sem relâmpagos. A Figura 12a sugere que sistemas convectivos com relâmpagos possuem taxa de expansão mais positivas, indicando a existência de fortes processos dinâmicos que promovem o crescimento inicial das tempestades que acabam adquirindo características elétricas. Por outro lado, sistemas convectivos sem relâmpagos estão mais associados a valores negativos de taxa de expansão, sugerindo que os processos dinâmicos de formação dessas nuvens não são intensos o suficiente para promover uma abrupta expansão da zona convectiva. Essas diferenças podem ser verificadas ao analisar os valores médios $66{,}5x10^{-6}.s^{-1}$ e $96{,}6x10^{-6}.s^{-1}$ para tempestades sem e com relâmpagos, respectivamente.

Complementando os resultados discutidos anteriormente, é possível notar na Figura 12b o deslocamento da distribuição dos valores de taxa de expansão referente as tempestades sem relâmpagos (com relâmpagos) para valores negativos (valores positivos). Sendo assim, sugere-se que a taxa de expansão é uma importante propriedade física que se diferencia significativamente entre tempestades com e sem relâmpagos. Isso está associado à dinâmica de crescimento dos sistemas convectivos. É possível notar que, em média aquelas tempestades com relâmpagos possuem uma taxa de crescimento maior do que àqueles sem relâmpagos,

sugerindo que fortes processos de desenvolvimento propiciam condições ideais para o surgimento das descargas elétricas.

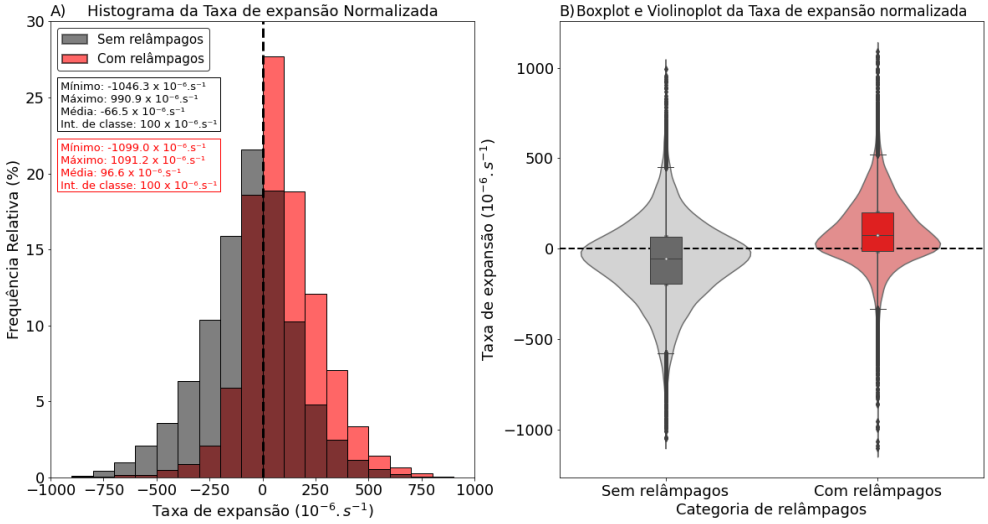

Figura 12 - Distribuição da taxa de expansão de normalizada ($10^{-6}.s^{-1}$) em sistemas convectivos com (cor vermelha) e sem (cor cinza) relâmpagos, sendo: A) histograma de frequência relativa (%) e B) gráfico *boxplot* e *violinoplot.*

3.1.3 Temperatura Média

A Temperatura Média corresponde à média da temperatura entre todos os *pixels* pertencente ao topo de um mesmo sistema convectivo. As maiores temperaturas são observadas em sistemas convectivos sem relâmpagos (Figura 13a, cor cinza), apresentando um máximo de 17% na frequência de ocorrência para a classe de temperatura de -43 ºC. Em contrapartida, os sistemas convectivos que apresentaram relâmpagos (Figura 13a, cor vermelha) possuem temperaturas mais baixas. Esse fato é corroborado ao analisar o valor médio: -49,3 ºC e -43,9 ºC para sistemas convectivos com e sem relâmpagos, respectivamente.

A Figura 13b mostra que a distribuição associada aos valores da mediana (2º quartil) não se sobrepõem, indicando uma distinção significativa em relação à temperatura média entre as tempestades com e sem relâmpagos. Adicionalmente, nota-se uma alta dispersão dos valores médios de temperatura em tempestades com relâmpagos, evidenciando uma distribuição relativamente mais uniforme quando comparada ao outro grupo. Nota-se uma maior concentração de temperatura média (aproximadamente -42,5 ºC) em sistemas convectivos sem relâmpagos ocorre quando há frequências expressivas em tempestades com relâmpagos. Esse fato sugere que somente a temperatura não explica essencialmente a formação de relâmpagos, mas sim, a união das características físicas e dinâmicas que promovem o início dos mecanismos de eletrificação das nuvens.

Portanto, é possível notar que temperaturas médias mais negativas são observadas em sistemas convectivos com relâmpagos. Isso ocorre devido a intensificação de correntes ascendentes dentro da nuvem que possibilitam um rápido crescimento vertical que pode ultrapassar, até mesmo a troposfera atingindo a camada atmosférica denominada de tropopausa (camada de transição entre a troposfera e a estratosfera) fazendo com que haja uma alta eficiência na produção de gelo no interior das nuvens, ao mesmo tempo que, os topos de tais sejam extremamente frios.

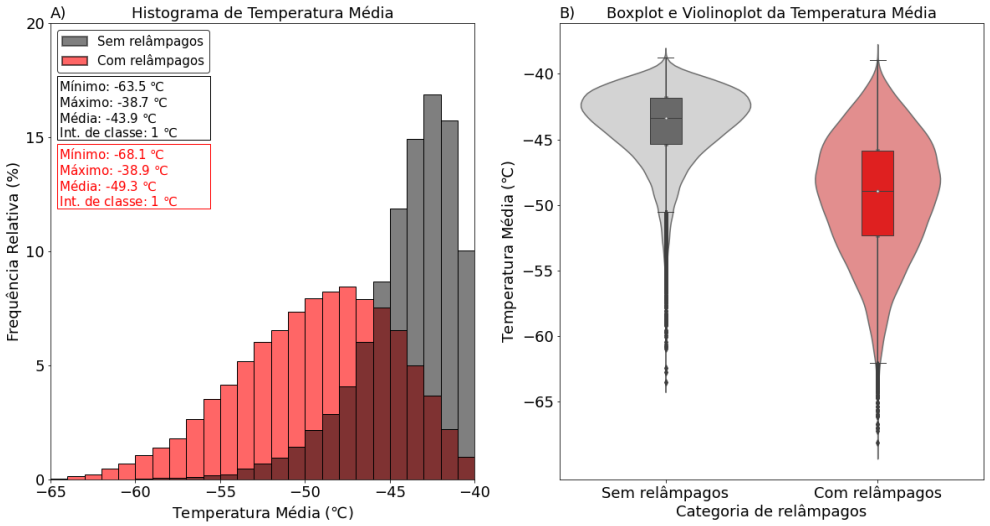

Figura 13 - Distribuição da temperatura média (ºC) em sistemas convectivos com (cor vermelha) e sem (cor cinza) relâmpagos, sendo: A) histograma de frequência relativa (%) e B) gráfico *boxplot* e *violinoplot.*

3.1.4 Variação da Temperatura Média

A Figura 14 apresenta as distribuições da variação da temperatura média (ºC/min) em sistemas convectivos com e sem relâmpagos. A Figura 14a evidencia que em sistemas convectivos sem relâmpagos, a variação da temperatura média tende a ser positiva quando comparada a sistemas convectivos com relâmpagos, onde as frequências para uma variação de temperatura negativa se mostraram elevadas, a partir de -2 ºC. Esse fato revela que, entre imagens de satélites consecutivas, o topo dos sistemas convectivos com relâmpagos diminui de temperatura mais intensamente quando comparados a sistemas convectivos sem relâmpagos. Estes resultados indicam a presença de fortes correntes ascendentes esteja associada com o desenvolvimento vertical da nuvem promovendo uma rápida queda de temperatura de seus topos.

A Figura 14b sugere que essa rápida variação negativa de temperatura ocorreu em torno de 50% (mediana próxima a 0 ºC) do total de sistemas convectivos com relâmpagos, em

contrapartida, para sistemas convectivos sem relâmpagos, esse número cai para 25% (1º quartil próximo a 0 ºC). No entanto, é possível notar que as medianas não se diferem amplamente. Isso mostra que, em caráter médio, a variação da temperatura média em sistemas convectivos com e sem relâmpagos não se diferenciam acentuadamente entre si, embora sistemas convectivos com relâmpagos apresentem maiores variações negativas com mais frequência. Dessa maneira, nota-se que essa variável não possui uma forte correlação com a ocorrência de relâmpagos (esse resultado também pode ser observado na Figura 22).

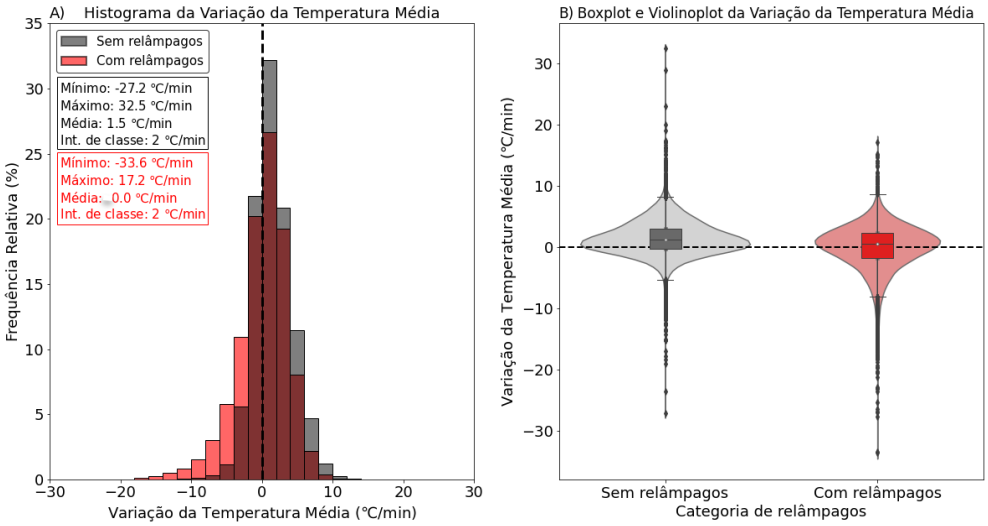

Figura 14 - Distribuição da variação temperatura média (ºC/min) em sistemas convectivos com (cor vermelha) e sem (cor cinza) relâmpagos, sendo: A) histograma de frequência relativa (%) e B) gráfico *boxplot* e *violinoplot*.

3.1.5 Temperatura Mínima

A Temperatura Mínima (ºC) corresponde ao mínimo valor dessa variável entre todos os pixels pertencentes ao topo de um mesmo sistema convectivo. A Figura 15a indica que a presença de relâmpagos está intimamente associada a menores temperaturas mínimas devido ao grande desenvolvimento vertical das nuvens. Nota-se essa predominância a partir de -56 ºC, podendo observar um aumento da frequência relativa de acordo com o decréscimo da temperatura mínima para sistemas convectivos com relâmpagos. Complementando os resultados encontrados na figura anterior, a Figura 15b mostra que as maiores frequências estão concentradas em torno de -49 ºC (-60 ºC) para sistemas convectivos sem relâmpagos (com relâmpagos). Além disso, percebe-se uma distribuição uniforme das frequências para tempestades com relâmpagos entre os valores de -41 ºC (limite superior) e -83 ºC (limite inferior).

Nota-se que na maior parte dos eventos, os sistemas convectivos sem relâmpagos possuem temperaturas mínimas mais elevadas do que aqueles com relâmpagos. Os resultados encontrados estão em consonância com o trabalho de Reynolds *et al*. (1957), os quais observaram que a atividade elétrica nas nuvens está intimamente relacionada à topos mais frios. Sendo assim, percebe-se que o desenvolvimento vertical das nuvens ligado a menores temperaturas em seus topos, é um importante indicativo probabilístico de uma eventual ocorrência de relâmpagos.

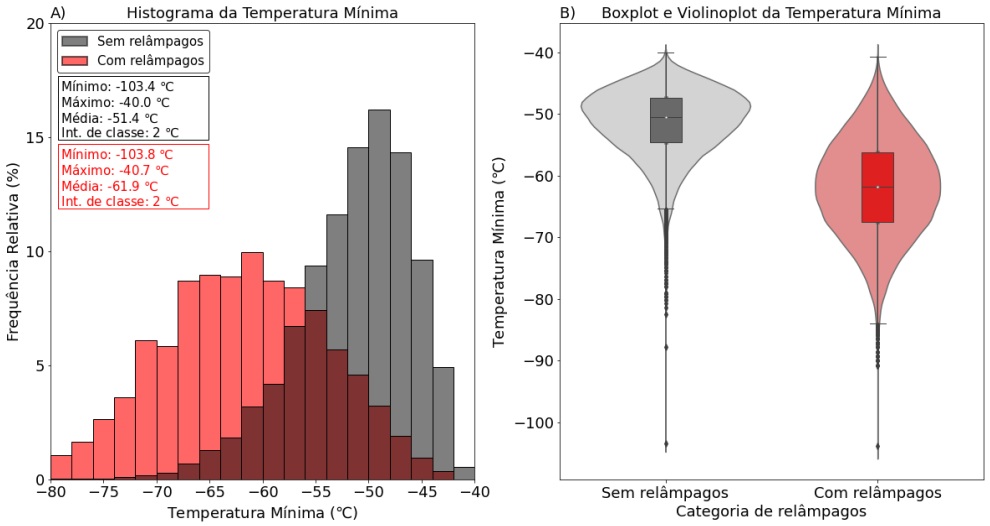

Figura 15 - Distribuição da temperatura mínima (ºC) em sistemas convectivos com (cor vermelha) e sem (cor cinza) relâmpagos, sendo: A) histograma de frequência relativa (%) e B) gráfico *boxplot* e *violinoplot*.

3.1.6 Variação da Temperatura Mínima

A Figura 16 mostra a Variação da Temperatura Mínima (ºC/min) em sistemas convectivos com e sem relâmpagos. A Figura 16a ressalta as variações de temperatura mínima positivas (negativas) em relação a sistemas convectivos sem relâmpagos (com relâmpagos), uma vez que, as médias também se mostram positivas (negativas). A Figura 16b mostra que embora as distribuições de ambos os grupos sejam aproximadamente simétricas, a categoria com relâmpagos possui 50 % do total de casos apresentando variações negativas de temperatura mínima, enquanto a categoria sem relâmpagos, apresenta apenas 25%.

A predominância de variação de temperatura positiva nos sistemas convectivos pode indicar que um consecutivo aumento dessa propriedade promove uma menor eficácia na produção de gelo e a consequente eletrificação da nuvem impedindo, dessa maneira, a formação e ocorrência de relâmpagos (MATTOS, 2009)

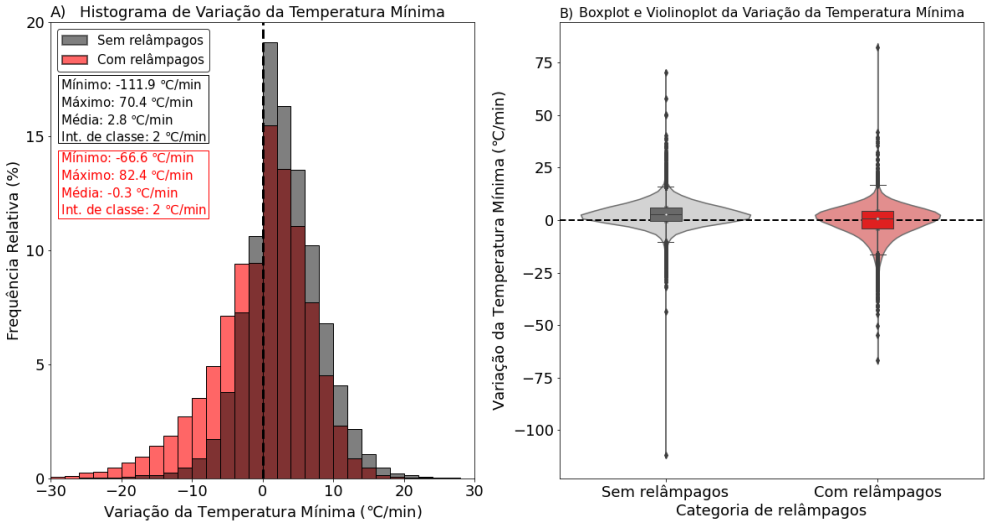

Figura 16 - Distribuição da variação da temperatura mínima (ºC/min) em sistemas convectivos com (cor vermelha) e sem (cor cinza) relâmpagos, sendo: A) histograma de frequência relativa (%) e B) gráfico *boxplot* e *violinoplot*.

3.1.7 Temperatura Mínima do *Kernel* de 9 *pixels*

A Temperatura Mínima do *Kernel* de 9 *pixels* ou Temperatura Mínima Média de Brilho do *Kernel* de 9 *pixels* (ºC) é definida como a média da temperatura entre os nove *pixels* mais frios do topo do sistema convectivo (Figura 17). Semelhantemente aos gráficos de temperatura, a frequência de ocorrência de TMIN9 para sistemas convectivos com relâmpagos mostram-se mais elevadas para temperaturas inferiores a -55 ºC, enquanto para sistemas convectivos sem relâmpagos as frequências mais altas são observadas entre -42ºC e -52 ºC. Também é observado que a média do grupo com relâmpagos (-59,5 ºC) é menor que do grupo sem relâmpagos (-49,2 ºC), ao mesmo tempo que o máximo observado para o último respectivo grupo, alcançou valores positivos (6,7 ºC). Isso indica que tempestades sem a presença de relâmpagos não possuem, na maioria das vezes, grande desenvolvimento vertical, implicando em topos de nuvens relativamente mais quentes.

A Figura 17b reforça a observação de que em média a TMIN9 é menor para sistemas convectivos com relâmpagos. O 3º quartil para este grupo mostra que 75% do total das tempestades estudadas apresenta topos com temperaturas inferiores a -55 ºC, contrastando com a temperatura de -45 ºC no 3ª quartil para tempestades sem relâmpagos. Importante ressaltar que, embora o intervalo interquartílico seja maior para o grupo com relâmpagos observou-se um grande número de *outliers* para o outro grupo, tendo uma quantidade significativa abaixo do limiar de -65 ºC e acima do limiar de -35 ºC, sendo que para este último foram observados

alguns *outliers* positivos de temperatura que podem estar associados a erros presentes na identificação dos sistemas convectivos.

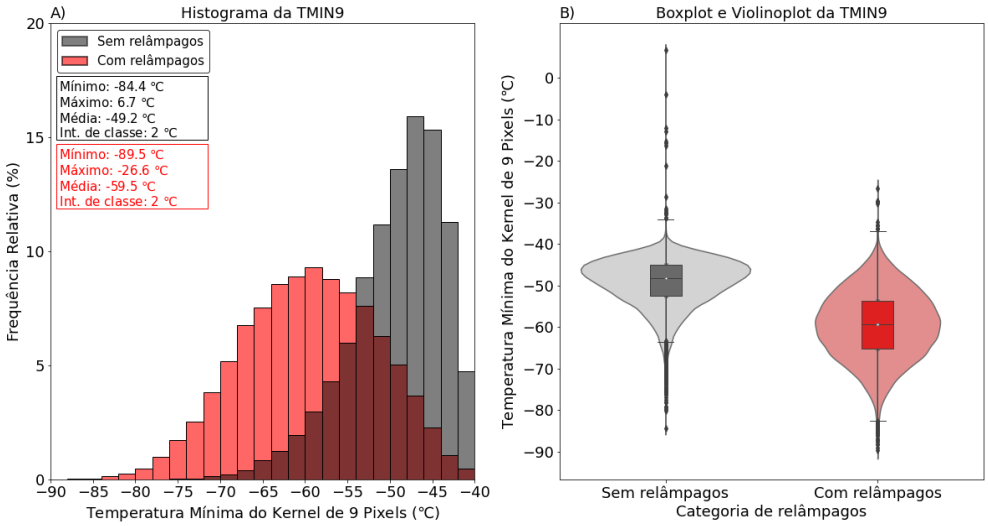

Figura 17 - Distribuição da temperatura mínima do *kernel* de 9 *pixels* (ºC) em sistemas convectivos com (cor vermelha) e sem (cor cinza) relâmpagos, sendo: A) histograma de frequência relativa (%) e B) gráfico *boxplot* e *violinoplot.*

3.1.8 Variação da Temperatura Mínima do Kernel de 9 *pixels*

A Figura 18 mostra a Variação Temperatura Mínima do Kernel de 9 *pixels* (ºC) em sistemas convectivos com e sem relâmpagos. Os resultados observados nas Figuras 18a e 18b são semelhantes aos observados nas Figuras 14 e 16, isto é, variações positivas (negativas) de TMIN9 possuem maior frequência relativa em sistemas convectivos sem (com) relâmpagos.

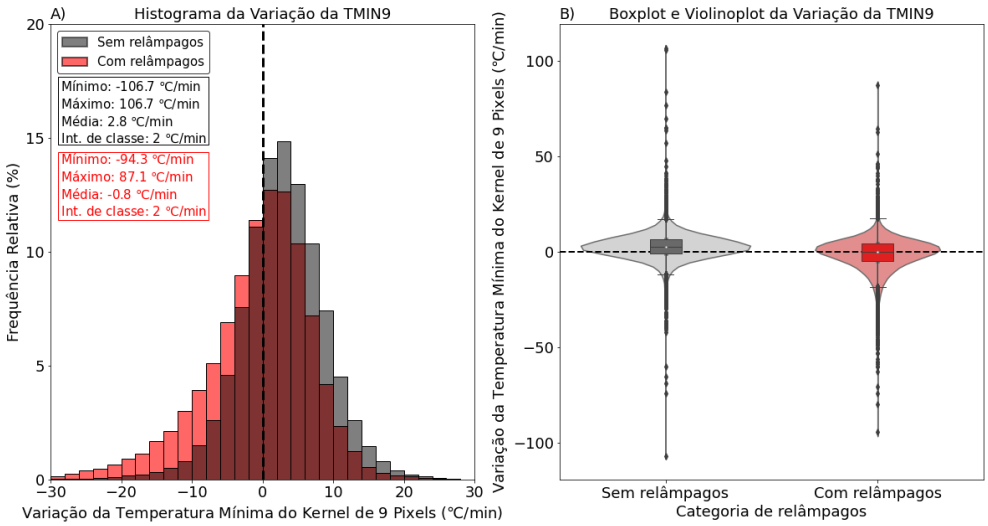

Figura 18 - Distribuição da variação da temperatura mínima do *kernel* de 9 *pixels* (ºC/min) em sistemas convectivos com (cor vermelha) e sem (cor cinza) relâmpagos, sendo: A) histograma de frequência relativa (%) e B) gráfico *boxplot* e *violinoplot*.

3.1.9 Relação entre relâmpagos e raio efetivo

A Figura 19 mostra a relação de dispersão entre a quantidade (quantidade de relâmpagos por SCM em 30 min) e densidade (quantidade de relâmpagos por quilômetro quadrado em 30 min) de relâmpagos em função do raio efetivo (km). Nota-se a existência de uma relação linear positiva entre a quantidade média de relâmpagos e o raio efetivo (Figura 19a), ou seja, conforme o tamanho da tempestade aumenta a quantidade total de relâmpagos também aumenta. É possível notar que entre 175 e 270 km existe uma rápida taxa de crescimento na quantidade de relâmpagos, aumentando de 1200 para 3300 para um aumento de 95 km no raio efetivo. Já a maior quantia média observada foi de 4100 relâmpagos para tempestades com raio efetivo de 360 km.

A Figura 19b mostra que as maiores concentrações de relâmpagos são observadas em sistemas convectivos menores, evidenciando uma tendência linear negativa entre as duas variáveis analisadas. O maior valor de densidade observado (0,026) ocorreu em sistemas convectivos que possuem raio efetivo de 19,7 km, enquanto o menor valor (~0,005 relâmpagos/30min*km^2 foi observado em sistemas convectivos com 500 km de raio efetivo. Esses resultados sugerem que tempestades menores tendem a ser mais eficientes em produzir relâmpagos do que aquelas que possuem uma vasta área contendo nuvens estratificadas. Machado e Rossow (1993) analisaram as propriedades de sistemas convectivos sobre a faixa tropical do continente africano e do Oceano Atlântico a partir de dados dos satélites Meteosat, GOES e GMS. Seus resultados mostraram que grandes sistemas convectivos estão associados a grandes coberturas por nuvens estratificadas, as quais não são nuvens produtoras de relâmpagos. Sendo assim, esse fato pode implicar nas menores concentrações de relâmpagos observados na Figura 19b para sistemas convectivos com extensões horizontais elevadas, as quais contrastam com os resultados encontrados na Figura 18a. Desse modo, pode-se dizer que sistemas convectivos maiores produzem, em média mais relâmpagos, no entanto, sistemas convectivos menores são mais eficientes.

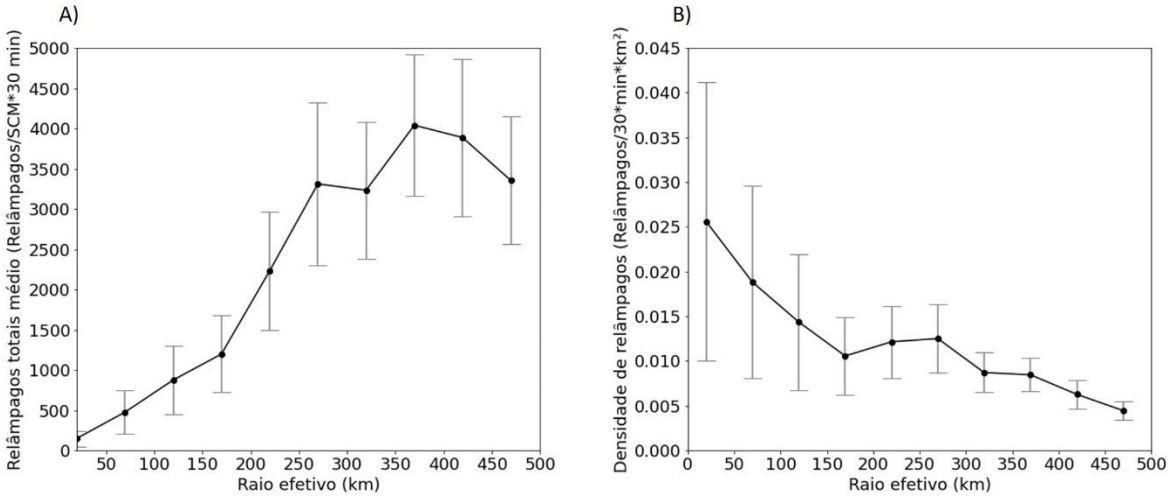

Figura 19 - Relação de dispersão entre o A) total (relâmpagos/30 min*SCM) e B) densidade (relâmpagos/30 min*km²) de relâmpagos e o raio efetivo (km).

3.1.10 Relação entre relâmpagos e taxa de expansão

A Figura 20 mostra a relação entre a quantidade e densidade de relâmpagos em função da taxa de expansão normalizada ($10^{-6}s^{-1}$). Nota-se que a fase de decaimento (crescimento), onde os valores de taxa de expansão são negativos (positivos) ocorre uma diminuição (ampliação) da quantidade de relâmpagos (Figura 20a). Na Figura 20b, é possível observar que os maiores (menores) valores de densidade de relâmpagos são observados quando a taxa de expansão é positiva (negativa). Os mínimos valores de densidade de relâmpagos (< 0,006 relâmpagos/30min*km^2) ocorreram em dois momentos, onde a taxa de expansão foi de -625x10^{-6}s^{-1} e a partir de -900x10^{-6}s^{-1} evidenciando que, durante a fase de decaimento os sistemas convectivos diminuem o rendimento da produção de relâmpagos. Em contrapartida, a partir da fase de maturação (onde o valor da taxa de expansão é igual a zero), a densidade de relâmpagos aumenta, com máximo valor (0,53) em aproximadamente 325x10^{-6}s^{-1} indicando que é durante o crescimento físico da tempestade que a taxa de produção de relâmpagos atinge seu ápice.

Dessa maneira, a associação entre essas propriedades sugere que é durante a fase de crescimento da tempestade que a eficiência em produzir relâmpagos é maior. Fisicamente, esses resultados tornam evidente que as correntes ascendentes presentes nos estágios iniciais das nuvens são responsáveis tanto pelo seu crescimento quanto pelo desenvolvimento das características que corroboram para sua eletrificação.

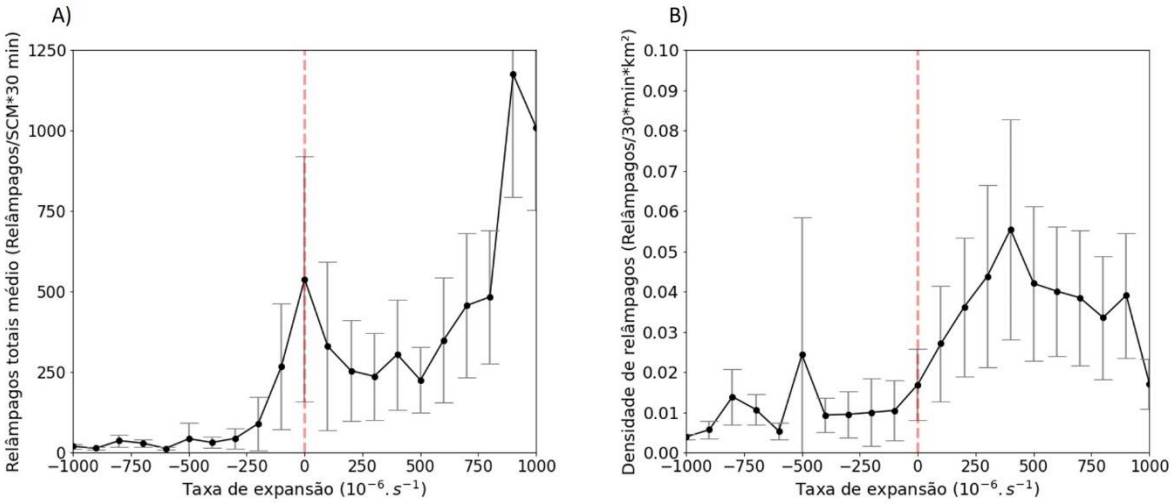

Figura 20 - Relação de dispersão entre o A) total (relâmpagos/30 min*SCM) e B) densidade (relâmpagos/30 min*km²) de relâmpagos e a taxa de expansão de normalizada ($10^{-6}s^{-1}$). A linha vertical tracejada vermelha representa o valor de 0,0 $10^{-6}s^{-1}$.

3.1.11 Relação entre relâmpagos e temperatura

A Figura 21a mostra uma tendência linear negativa entre a quantidade de relâmpagos e as temperaturas (média, mínima e mínima média do *kernel* de 9 *pixels*), ou seja, um decréscimo de temperatura implica em um aumento do número de relâmpagos. A temperatura média apresenta um aumento gradual do número de relâmpagos a partir de -50 ºC (Figura 21a), atingindo um valor máximo de aproximadamente 2200 relâmpagos em -70 ºC. A temperatura mínima e a TK9 mostram um comportamento semelhante. Entre os limiares de -25 ºC a -65 ºC observa-se um número constante de relâmpagos da ordem de dezenas de eventos. No entanto, a partir de -70 ºC, nota-se um crescimento significativo da quantidade de relâmpagos para ambas as propriedades à medida que as temperaturas vão se tornando mais negativas. Além disso, é possível observar que ao considerar o mesmo valor de temperatura entre -60 ºC e -87 ºC no eixo das abcissas a TK9 está relacionada a uma maior quantidade de relâmpagos quando comparada com a TMIN, onde essa última, ultrapassa a quantidade de relâmpagos em relação a TK9, a partir de -90 ºC.

A Figura 21b mostra a existência de uma relação inversamente proporcional entre as variáveis analisadas, isto é, à medida que ocorre um decréscimo nos valores das temperaturas observa-se um aumento na quantidade de relâmpagos. Para a temperatura média, a máxima densidade (0,065 relâmpagos/30 min*km²) observada foi aproximadamente -64 ºC (menor temperatura média da distribuição), enquanto para a temperatura mínima (0,042 relâmpagos/30 min*km²) foi em torno de -76 ºC e para a TK9 (0,041 relâmpagos/30 min*km²) foi em aproximadamente -74 ºC).

Esses resultados sugerem que a temperatura do topo dos sistemas convectivos é fortemente relacionada com a quantidade e a taxa de produção de relâmpagos, ao passo que, temperaturas mais negativas induzem condições termodinâmicas favoráveis a uma maior formação de partículas e cristais de gelo dentro da nuvem que corroboram para a sua eletrificação e consequente formação de relâmpagos (REYNOLDS *et al*., 1957; WALLACE; HOBBS, 1977). Além disso, importante ressaltar que se verificou que existem possíveis limiares de temperaturas mínimas dos topos das tempestades que resultam em uma maior eficiência na produção de relâmpagos. Esses resultados correspondem aos encontrados por Dotzek *et al*. (2005), o qual estudando SCM a partir de imagens do satélite do GOES observaram um aumento notável de relâmpagos abaixo do limiar de -70 ºC. Por outro lado, Goodman e Macgorman (1986) observaram que, em CCM, esse aumento ocorreu a partir de -53 ºC.

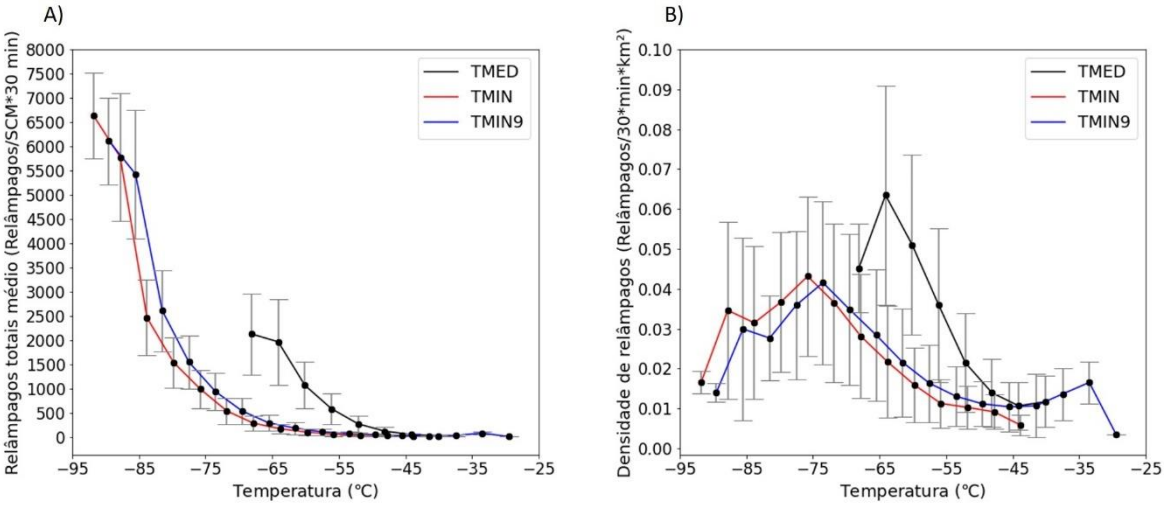

Figura 21 - Relação de dispersão entre o A) total (relâmpagos/30 min*SCM) e B) densidade (relâmpagos/30 min*km²) de relâmpagos e a temperatura média (linha na cor preta), mínima e média (linha na cor vermelha), mínima do *kernel* de 9 *pixels* (linha na cor azul).

3.2 Correlação de Pearson entre os parâmetros físicos e relâmpagos

A correlação de Pearson é um método estatístico utilizado para medir a relação entre duas variáveis contínuas, isto é que podem possuir qualquer valor numérico entre um intervalo. Sendo assim, valores positivos (negativos) implicam em variáveis diretamente (inversamente) proporcionais ao mesmo tempo que valores próximos a 1 ou -1 indicam forte relação. Nesse contexto, será avaliado a relação entre as propriedades físicas e a quantidade/densidade de relâmpagos através de mapa de calor (do inglês, *HeatMap*, Figura 22).

Analisando as relações encontradas entre as propriedades físicas e a quantidade de relâmpagos (“Relâmpagos Totais”) é possível observar significativos valores de correlação para as propriedades: raio efetivo (0,49) e as temperaturas média (-0,35), mínima (-0,42) e mínima do *kernel* de 9 *pixels* (-0,40). Em consistência aos resultados encontrados nas seções 3.1.9 e 3.1.11 esses valores de correlação indicam que a quantidade de relâmpagos aumenta conforme há um aumento do raio efetivo ou a partir de um decréscimo de temperatura e vice-versa. Em contrapartida, os valores para taxa de expansão (0,03) e as variações de temperatura média (-0,09), mínima (-0,07) e mínima do *kernel* de 9 *pixels* (-0,06) entre duas imagens de consecutivas não mostraram forte correlação (valores próximos a zero). No entanto, segundo Mattos e Machado (2011) os relâmpagos estão mais associados há valores positivos de taxa de expansão, os quais são observados durante a fase de crescimento progressiva do sistema. Dessa forma, como a correlação de Pearson considera a média entre os valores de desvio padrão entre as variáveis analisadas, a representatividade para a taxa de expansão não possui alta significância ao serem englobados os valores negativos da variável no cálculo.

Os resultados encontrados em relação à densidade de relâmpagos mostram que embora a magnitude da correlação seja menor, ainda é possível observar uma associação, principalmente entre as variáveis: taxa de expansão (0,22), temperatura média (-0,25), mínima (-0,20) e mínima do *kernel* de 9 *pixels* (-0,18). Além disso, os parâmetros associados a variação de temperatura também se mostraram mais relacionados (-0,34, -0,26 e -0,24, respectivamente a ordem supracitada) à densidade do que à quantidade de relâmpagos.

É importante ressaltar algumas observações interessantes adicionais na Figura 22. Por exemplo, nota-se uma notável correlação negativa (entre -0,3 e -0,54) entre o tamanho e a temperatura do topo dos sistemas convectivos, sugerindo que temperaturas mais frias são observadas em sistemas convectivos maiores. Semelhantemente, associação semelhante pode ser observada entre a taxa de expansão e as variações de temperatura entre duas imagens consecutivas, ao passo que, sistemas convectivos com topos mais frios apresentam valores de taxa de expansão mais positivos. Esse resultado condiz com a literatura (GOODMAN; MACGORMAN, 1986). Os resultados da literatura indicam que devido a intensos processos dinâmicos nos estágios iniciais (taxa de expansão positiva) existe um elevado desenvolvimento vertical da nuvem, colaborando dessa maneira, tanto para a evolução das temperaturas em seu topo para valores ainda mais negativos quanto para o aumento da produção de partículas e cristais de gelo em seu interior.

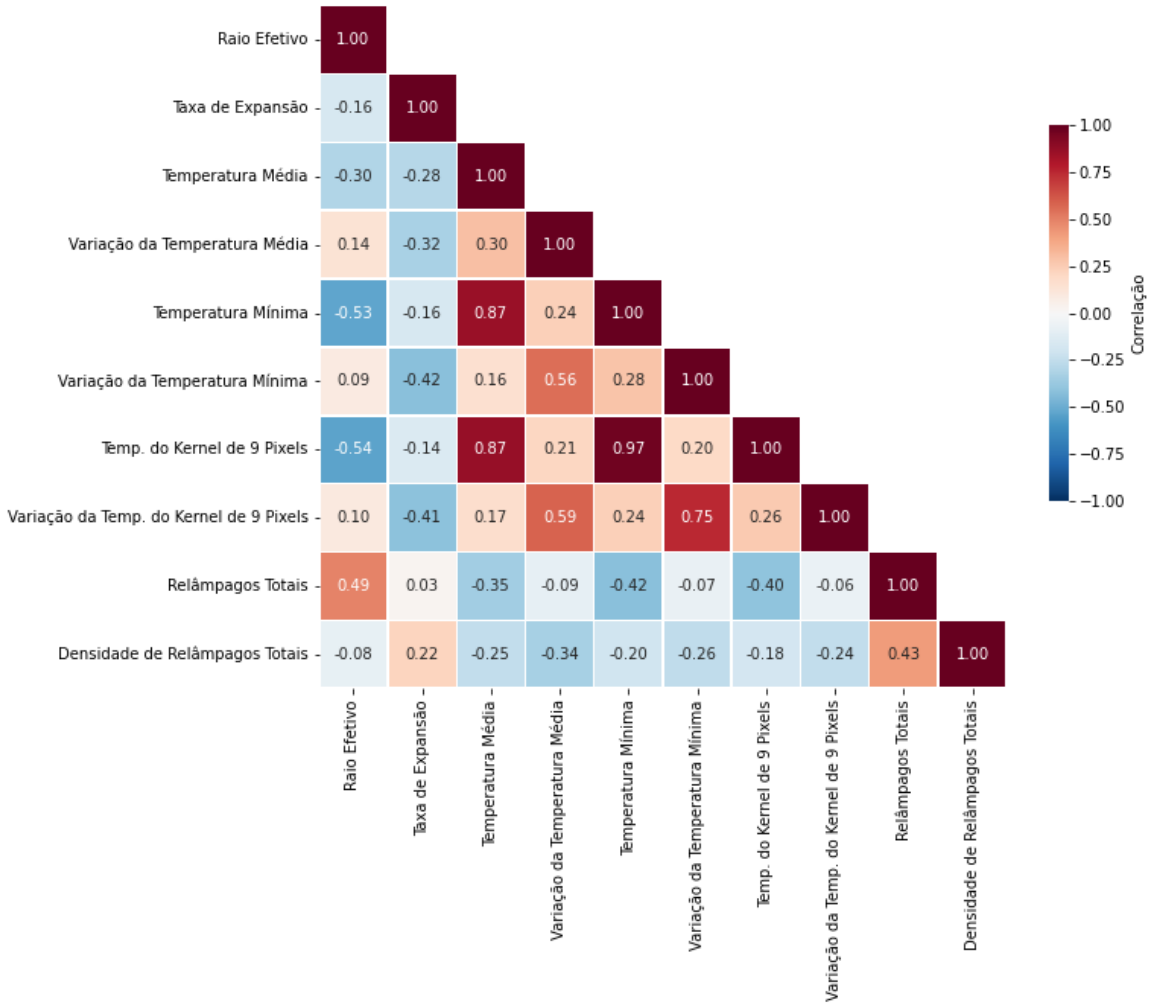

Figura 22 - Gráfico *heatmap* evidenciando a correlação de Pearson entre os parâmetros físicos das tempestades e os relâmpagos (ocorrência total e densidade).

3.3 Aplicação e avaliação dos algoritmos de *Machine Learning*

A partir dos resultados obtidos nas seções 3.1 e 3.2 foram selecionados os parâmetros que estão mais bem correlacionados com os relâmpagos, que são: raio efetivo, taxa de expansão normalizada e temperatura mínima dos sistemas convectivos. Dessa maneira, a partir da metodologia adotada na seção 2.5 foram aplicados os diferentes tipos de modelos de aprendizado de máquina (regressão linear, regressão logística, árvore de decisão e floresta aleatória) e, em seguida aplicadas as métricas de avaliação que serão discutidas e analisadas nas próximas seções.

3.3.1 Modelo de regressão linear

O modelo de regressão linear foi aplicado utilizando as variáveis supracitadas como preditoras (isto é: raio efetivo, taxa de expansão e temperatura mínima) a fim de inferir a

quantidade de relâmpagos que um sistema convectivo poderia apresentar. A equação 17 mostra a função de regressão encontrada pelo modelo aplicado:

$$Y = 4805{,}46208022 + 6{,}78060133 * (REFE(KM)) + 0.27221796 * (DSIZE) - 23{,}39577697 * (TMIN) \quad (17)$$

Dessa maneira, a Tabela 2 mostra que o erro residual mínimo associado a quantidade de relâmpagos é relativamente baixa (-3,5 relâmpagos), ou seja, em média, o modelo de regressão subestimou o número de relâmpagos em relação ao valor real observado. No entanto, o maior erro residual encontrado foi de aproximadamente 13444,9 relâmpagos, evidenciando a grande variação possível na predição de relâmpagos. O Erro Absoluto Médio (MAE) apresentou um valor moderadamente alto (368 relâmpagos), indicando que, em média, há um erro de 368 relâmpagos nas previsões realizadas pelo modelo. Em contrapartida, o Erro Quadrático Médio (MSE) apresentou um valor bastante alto (548571,5) revelando que há uma grande quantidade de *outliers* ou que a distribuição há um alto desvio padrão associado, influenciando no aumento quadrático do erro. Aplicando a raiz quadrada no MSE, temos o valor da métrica Raíz do Erro Quadrático Médio (RMSE), o qual resultou no valor de 740,7, permitindo compreender de forma mais simplificada a influência direta do número de *outliers* na precisão do modelo de regressão para prever o número de relâmpagos. Por fim, o coeficiente de determinação R-Quadrado ou R^2 torna evidente que o modelo de regressão linear aplicado não explica 70% da variância dos dados, isto é, 70% dos valores preditos estão distantes do valor médio central indicando que a eficiência do modelo de regressão em predizer a quantidade de relâmpagos a partir das propriedades físicas das tempestades é baixa.

Dessa forma, com base nos resultados supracitados nota-se que a aplicação de modelos de regressão linear para a previsão da quantidade de relâmpagos em sistemas convectivos não se mostrou ser um modelo apropriado. Esse fato pode estar associado à complexidade física e matemática envolvida nos processos físicos e microfísicos das tempestades responsáveis por promover a ocorrência de relâmpagos (MACGORMAN; RUST, 1998; RAKOV; UMAN, 2003). Sobre o ponto de vista estatístico, embora haja certa relação entre as propriedades físicas em relação à quantidade de relâmpagos, nota-se que não é uma relação expressivamente linear que possa ser explorada em caráter de previsão. Além disso, como discutido por James *et al.* (2013) alguns fatores podem se tornar potenciais problemas na performance de um modelo de regressão linear, como por exemplo: Não linearidade (significativa) entre as variáveis preditoras

e a variável de resposta, grande número de *outliers* e colinearidade, isto é, quando as variáveis preditoras estão intimamente associadas entre si.

Tabela 2 - Métricas de avaliação do modelo de regressão linear aplicados para as propriedades físicas das tempestades: Erro residual médio, Erro residual máximo, Erro Absoluto Médio (da sigla em inglês, MAE), Erro Quadrático Médio (da sigla em inglês, MSE), Raiz do Erro Quadrático Médio (da sigla em inglês, RMSE) e R-quadrado ou R^2.

Métricas	Valores
Erro residual médio	-3,5
Erro residual máximo	13444,9
MAE	368
MSE	548571,5
RMSE	740,7
R^2	0,3

3.3.2 Modelos de classificação

A Tabela 3 mostra os valores encontrados para as contingências da matriz de confusão (verdadeiro positivo, falso positivo, falso negativo e verdadeiro negativo) em relação à performance de cada modelo de classificação: regressão logística, árvore de decisão e floresta aleatória. Primeiramente, ao analisar os resultados para os modelos de classificação de forma geral é observado que os valores das classes apresentadas como “verdadeiros” (“falsos”) representam a maioria (minoria) dos resultados. Para o modelo de regressão logística observa-se que 80,1% (32219 tempestades) das predições foram classificadas corretamente (verdadeiro positivo + verdadeiro negativo) enquanto apenas 19,9% (7993 tempestades) foram classificadas erroneamente (falsos positivos + falsos negativos). Sendo assim, entre os três modelos, a regressão logística foi o que obteve a maior (menor) taxa de acertos (erros).

Para o modelo de árvore de decisão 72% (28955 tempestades) das predições foram classificadas corretamente e 28% (11257 tempestades) de forma errônea, ao passo que foi o modelo que apresentou as maiores taxas de falsos positivos e negativos, indicando a fragilidade

para falsos alarmes (previu que haveria relâmpagos e não houve) e erros grosseiros (previu que não haveria relâmpagos e houve). No entanto, a taxa de falsos negativos foi a menor entre as contingências dos próprios modelos, apresentando uma diferença dos valores de falsos positivos de 1,2%, contrastando aos 3,1% observados tanto para regressão logística quanto para floresta aleatória. Por fim, para o modelo de floresta aleatória, 79,1 % (31787 tempestades) foram classificadas corretamente, enquanto 20,9% (8425 tempestades) classificadas incorretamente. Para este modelo, observa-se que seus resultados foram bastante próximos aos encontrados para o modelo de regressão logística.

Sendo assim, entre os três modelos de classificação aplicados, o modelo que obteve a pior performance, isto é, apresentou menor taxa de acertos (72%) e maior taxa de erros (28%) foi o de árvore de decisão. Esse fato pode estar associado a imprecisão por trás desse método que utiliza apenas uma raiz por detrás da classificação da predição em questão, tornando o ganho de informações subsequentes tendencioso (JAMES *et al.*, 2013). Dessa forma, como as variáveis preditoras são contínuas e não essencialmente lineares, uma árvore de decisão pode não conseguir descrever a natureza de predição dos dados. Por outro lado, modelos de floresta aleatória baseiam-se em inúmeras árvores de decisão para diminuir essas tendências e aumentar o nível de classificação. Já o modelo de regressão logística, como fornece uma probabilidade de ocorrência, onde valores maiores (menores) que 0,5 referem-se a tempestades que (não) ocorrerão relâmpagos, conseguiu determinar a classificação com mais precisão.

Tabela 3 - Matriz de confusão para avaliação dos modelos de classificação aplicados para as propriedades físicas das tempestades. São mostrados os resultados para os modelos: Regressão Logística, Árvore de Decisão e Floresta Aleatória.

Matriz de confusão	Regressão Logística	Árvore de Decisão	Floresta Aleatória
Verdadeiro positivo	12317 (30,6 %)	11587 (28,8 %)	12173 (30,3%)
Falso positivo	3207 (8 %)	5879 (14,6 %)	3595 (8,9 %)
Falso negativo	4786 (11,9 %)	5378 (13,4 %)	4830 (12 %)

Verdadeiro negativo	19902 (49,5 %)	17368 (43,2%)	19614 (48,8%)

A Tabela 4 apresenta as métricas de avaliação das previsões realizadas para os três modelos de classificação: regressão logística, árvore de decisão e floresta aleatória. Tais métricas consideram os valores encontrados nas matrizes de confusão da Tabela 3, e realizam penalizações estatísticas a respeito de acertos e falhas com o intuito de fornecer valores que pronunciam a qualidade e desempenho final dos modelos. Desse modo, a acurácia ou a taxa de previsões corretas obtida pelos modelos foi de 80%, 72% e 79% para regressão logística, árvore de decisão e floresta aleatória, respectivamente. Esses valores condizem aos resultados encontrados na Tabela 3, onde foram observados altos valores de acertos associados aos modelos. Para a especificidade; parâmetro este que permite inferir o quanto da classe observada como "não" foi prevista corretamente, apresentou valores de 0,86, 0,74 e 0,85 para os respectivos modelos supracitados. Estes resultados mostram que os modelos de regressão logística e floresta aleatória tendem a classificar com mais eficiência de acerto quando há tempestades sem relâmpagos.

As métricas seguintes foram subdivididas em duas categorias: tempestades com relâmpagos e sem relâmpagos. A precisão, isto é, a taxa de previsões positivas que foram corretamente previstas, para o grupo com (sem) relâmpagos foi de 79%, 64% e 78% (81%, 76% e 80%). Embora tenha sido observada pouca divergência entre os grupos (exceto para o modelo de árvore de decisão), esses valores sugerem que a taxa de acerto para tempestades sem relâmpagos é maior, ou seja, os modelos se ajustam melhor e, consequentemente, conseguem prever com mais eficiência quando não haverá relâmpagos. Para a métrica sensibilidade, ou seja, a taxa de tempestades que realmente houve relâmpagos e foram corretamente previstas, encontrados para o grupo com (sem) relâmpagos foi de 72%, 67% e 71% (86%, 76% e 85%), respectivamente. Dessa forma, é possível notar que há um maior erro associado a maiores quantidades de falso negativo para o grupo com relâmpagos, revelando uma menor eficiência de previsão para este tipo de tempestade, embora as métricas tenham apresentado valores próximos a 70%, o qual é relativamente satisfatório. Já para o grupo sem relâmpagos, foi observado uma taxa de acertos levando em consideração a classe "sim" para observados, no em torno de 85% sugerindo uma ótima gama de acertos quando não houve relâmpagos de fato.

Por fim, o F1-score dos modelos; parâmetro este que relaciona as métricas de precisão e sensibilidade, apresentou valores de 76%, 67% e 74% (83%, 76% e 82%) para sistemas convectivos com (sem) relâmpagos. Percebe-se que para os dois grupos, o F1-score foi acima de 65% em todos os casos, chegando até a 83% para o grupo sem relâmpagos associado ao modelo de regressão logística. Assim, percebe-se que os modelos conseguem prever com certo grau de qualidade a ocorrência ou não de relâmpagos em sistemas convectivos a partir de suas propriedades físicas.

De forma geral, o modelo que obteve o melhor desempenho, respectivamente foi: regressão logística, floresta aleatória e árvore de decisão, sendo que os dois primeiros apresentaram na maioria dos casos uma divergência média entre as métricas de apenas 1%. Além disso, foi observado que os modelos se ajustam com menor (maior) eficiência para predizer quando não (sim) haverá relâmpagos. Esses resultados podem estar associados à existência de alguns limiares mínimos de área, taxa de expansão e temperatura, que em conjunto ou não, relacionam-se à ocorrência ou não de relâmpagos. Outro fator que poderia contribuir seria em relação a dispersão dos dados (observados na seção 3.1) dos sistemas convectivos com e sem relâmpagos, uma vez que, para o primeiro grupo foi observada uma dispersão maior quando comparada com os dados das tempestades que não houve relâmpagos. Desse modo, uma maior dispersão implicaria em maiores incertezas dos modelos. Dessa forma, uma maior variância nos dados poderia implicar em maiores incertezas.

Tabela 4 - Métricas de avaliação para os modelos de classificação aplicados para as propriedades físicas para tempestades com e sem relâmpagos. São mostrados os resultados para Regressão Logística, Árvore de Decisão e Floresta Aleatória.

Métricas	Tipo de tempestade (*flags*)	Regressão Logística	Árvore de Decisão	Floresta Aleatória
Acurácia	Ambas	0,80	0,72	0,79
Especificidade	Ambas	0,86	0,74	0,85
Precisão	Sem raios	0,81	0,76	0,80
	Com raios	0,79	0,67	0,78

Sensibilidade	Sem raios	0,86	0,76	0,85
	Com raios	0,72	0,67	0,71
F1-score	Sem raios	0,83	0,76	0,82
	Com raios	0,76	0,67	0,74

4. Conclusão

Este estudo avaliou a relação entre as propriedades físicas e elétricas das tempestades no estado de São Paulo entre 2013 e 2017 através da utilização de dados de temperatura do canal infravermelho proveniente do satélite geoestacionário GOES-13 e dados de relâmpagos de rede de detecção em solo. Além disso foi analisado como as propriedades das tempestades podem auxiliar na aplicação de ferramentas de inteligência artificial que buscam predizer a ocorrência de relâmpagos. Dessa forma, devido ao grande número de sistemas convectivos individuais analisados (57446) durante o período de 5 anos, torna essa uma das primeiras pesquisas do país a estudar um amplo número de tempestades e a aplicabilidade de suas propriedades em inteligência artificial.

Para avaliar as propriedades físicas mais relacionadas com a ocorrência de relâmpagos realizou-se um estudo preliminar sobre o comportamento e distribuição do tamanho, taxa de expansão e temperatura em sistemas convectivos com e sem a presença de relâmpagos. Foi observado que sistemas convectivos com relâmpagos tendem a apresentar maiores áreas do que sistemas convectivos sem relâmpagos, ao passo que, foi observado a predominância de frequências acima do limiar de 175 km de raio efetivo para o primeiro grupo. Além disso, notou-se que 75% do total de tempestades que produziram relâmpagos possuem raio efetivo menor que 75 km, enquanto para o grupo sem relâmpagos esse limiar foi de 48 km. Em consistência, notou-se que a quantidade de relâmpagos possui uma tendência linear positiva em relação ao desenvolvimento do tamanho das tempestades. Em contrapartida, a densidade de relâmpagos mostrou maiores concentrações em sistemas convectivos menores. Já a taxa de expansão das tempestades, mostrou que a presença (ausência) de relâmpagos está associada a valores médios positivos (negativos) da variável. Foi visto que apenas 25% do total de tempestades com relâmpagos possuía valores negativos de taxa de expansão associados. Ao mesmo tempo, para o grupo sem relâmpagos, observou-se que aproximadamente 75% dos casos, estavam

associados a valores negativos. Dessa maneira, nota-se que sistemas convectivos com (sem) relâmpagos passam a maior parte do tempo do seu ciclo de vida crescendo (dissipando). Quando foi comparado com a quantidade e a densidade de relâmpagos, notou-se um crescimento expressivo quando o valor de taxa de expansão era maior que zero. Os campos de temperatura (média, mínima e mínima do *kernel* de 9 *pixels*) mostraram que o topo dos sistemas convectivos com (sem) relâmpagos tendem a ser mais frios (quentes). Além disso, em conformidade, observou-se que a quantidade e densidade de relâmpagos possui uma relação inversa com a temperatura dos sistemas convectivos, uma vez que foram observados maiores quantidades e concentrações em sistemas convectivos mais frios. Dessa forma, percebeu-se que tais propriedades são importantes indicativos que diferem a severidade dos sistemas convectivos no que tange a maiores tamanhos, forte taxa de crescimento em seus estágios iniciais e desenvolvimento vertical, características essas referenciadas na literatura, fundamentais para promover o surgimento das propriedades elétricas das tempestades.

Para mensurar o nível de relação entre as propriedades físicas e elétricas dos sistemas convectivos foi avaliada a correlação de Pearson entre as variáveis. Embora o valor de correlação para a taxa de expansão tenha sido baixo, os resultados anteriores evidenciaram que existe uma relação forte com a presença de relâmpagos, mas que depende do estágio (inicial, maduro ou dissipação) que a tempestade se encontra. Esses resultados sugerem que a análise das relações entre as características das tempestades possa se tornar mais efetiva ao analisar os estágios de desenvolvimento.

A última etapa do trabalho consistiu em avaliar modelos de aprendizado de máquina para as propriedades das tempestades. Os resultados mostraram que predizer a quantidade de relâmpagos por meio de algoritmos de regressão linear ainda necessitam de mais investigação devido à complexidade natural dos relâmpagos dependerem de características micro e macrofísicas. No entanto, a aplicação de algoritmos de classificação (se haverá ou não relâmpagos) demonstrou ser amplamente eficiente, uma vez que as taxas de acertos foram superiores a 70% para todos os modelos. Dentre esses, a regressão logística mostrou-se ser a mais eficaz entre os modelos, seguida por pouca diferença (menor que 1%) pelo modelo de floresta aleatória. O modelo de árvore de decisão, apesar de ter apresentado bons resultados foi o que apresentou o pior desempenho.

Portanto, este trabalho apresentou que a integração de informações provenientes de satélites e redes de monitoramento de relâmpagos em solo é uma importante ferramenta para caracterizar e estimar a atividade elétrica de sistemas convectivos. Além disso, a utilização de ferramentas como a inteligência artificial é imprescindível para um melhor entendimento físico

do comportamento dos relâmpagos e possibilita criar meios que auxiliem na previsão de curto prazo (*nowcasting*). Entretanto, vale a pena ressaltar, que os processos físicos associados a ocorrência de relâmpagos mostram-se extremamente complexos necessitando ainda de estudos mais detalhados. Para estudos futuros seria interessante combinar diferentes propriedades em diversas escalas (micro e macro) através da união de dados de radares meteorológicos, satélites e redes de monitoramento de relâmpagos em solo para representar a naturalidade dimensional envolvida na formação das tempestades. Além disso, torna-se importante fomentar estudos que visem a construção de redes neurais convolucionais de aprendizado profundo que se baseiam na captura de imagens de satélites geoestacionários e distribuição de relâmpagos com o intuito de extrair informações e vieses que ditem o comportamento das tempestades.

5. Referências bibliográficas

ABREU, Lizandro Pereira de. **Relâmpagos no Nordeste do Brasil: Ocorrência, Variabilidade espaço-temporal e relação com microfísica das nuvens**. 2018. Dissertação de Mestrado. Brasil.

ALENEZI, Hadeel S.; FAISAL, Maha H. Utilizing crowdsourcing and machine learning in education: Literature review. **Education and Information Technologies**, v. 25, n. 4, p. 2971-2986, 2020.

ANDERSON, Christopher J.; ARRITT, Raymond W. Mesoscale convective complexes and persistent elongated convective systems over the United States during 1992 and 1993. **Monthly Weather Review**, v. 126, n. 3, p. 578-599, 1998.

BOCHENEK, Bogdan; USTRNUL, Zbigniew. Machine Learning in Weather Prediction and Climate Analyses—Applications and Perspectives. **Atmosphere**, v. 13, n. 2, p. 180, 2022.

CAELEN, Olivier. A Bayesian interpretation of the confusion matrix. **Annals of Mathematics and Artificial Intelligence**, v. 81, n. 3, p. 429-450, 2017.

CARDOSO, I. et al. Lightning casualty demographics in Brazil and their implications for safety rules. **Atmospheric Research**, v. 135, p. 374-379, 2014.

CINTINEO, John L.; PAVOLONIS, Michael J.; SIEGLAFF, Justin M. ProbSevere LightningCast: A deep-learning model for satellite-based lightning nowcasting. **Weather and Forecasting**, 2022.

CULKIN, Robert; DAS, Sanjiv R. Machine learning in finance: the case of deep learning for option pricing. **Journal of Investment Management**, v. 15, n. 4, p. 92-100, 2017.

DARCY, Alison M.; LOUIE, Alan K.; ROBERTS, Laura Weiss. Machine learning and the profession of medicine. **Jama**, v. 315, n. 6, p. 551-552, 2016.

DEO, Rahul C. Machine learning in medicine. **Circulation**, v. 132, n. 20, p. 1920-1930, 2015.

Divisão de Impactos, Adaptação e Vulnerabilidades. ELAT (BrasilDAT), Centro de Ciência da Terra, Instituto Nacional de Pesquisas Espaciais, 2022. Disponível em: <http://www.ccst.inpe.br/projetos/brasildat/>. Acesso em 24 de maio de 2022.

DIXON, Matthew F.; HALPERIN, Igor; BILOKON, Paul. **Machine learning in Finance**. Springer International Publishing, 2020.

DOTZEK, N. et al. Lightning activity related to satellite and radar observations of a mesoescale convective system over Texas o 7-8 April 2002. Atmospheric Research, v. 76, p. 127-166, 2005.

DRUGAN, John. J.; PRESTON, Ari. D. Lightning Cessation Guidance Using Polarimetric Radar Data and Lightning Mapping Array in the Washington, D.C., Area. **Atmosphere** 2022, 13, x

Earth Networks Total Lightning Network. Disponível em: https://www.earthnetworks.com/lightning-detection/#Lightning-Detection-Network. Acesso em 25 de fevereiro de 2022.

FERRARI, DANIEL GOMES; SILVA, LEANDRO NUNES DE CASTRO. **Introdução a mineração de dados**. Saraiva Educação SA, 2017.

FIGUEIREDO FILHO, Dalson Britto; SILVA JÚNIOR, José Alexandre. Desvendando os Mistérios do Coeficiente de Correlação de Pearson (r). **Revista Política Hoje**, v. 18, n. 1, p. 115-146, 2009.

GOODELL, John W. et al. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. **Journal of Behavioral and Experimental Finance**, v. 32, p. 100577, 2021.

GROSS, Jurgen; GROß, Jürgen. **Linear regression**. Springer Science & Business Media, 2003.

GUNST, Richard F.; MASON, Robert L. **Regression analysis and its application: a data-oriented approach**. CRC Press, 2018.

HAHN, Rosamaria. **Estimativa da ocorrência e severidade de granizo no Rio Grande do Sul baseado em observações de radar meteorológico**. 2021. Tese de Doutorado. Universidade de São Paulo.

HANDELMAN, G. S. et al. eD octor: machine learning and the future of medicine. **Journal of internal medicine**, v. 284, n. 6, p. 603-619, 2018.

HOLMSTROM, Mark; LIU, Dylan; VO, Christopher. Machine learning applied to weather forecasting. **Meteorol. Appl**, p. 1-5, 2016.

HOFFERT, H. H. XV. Intermittent lightning-flashes. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, v. 28, n. 171, p. 106-109, 1889.

HOSMER, David W.; JOVANOVIC, Borko; LEMESHOW, Stanley. Best subsets logistic regression. **Biometrics**, p. 1265-1270, 1989.

HOUZE JR, R. B. Cloud dynamics. San Diego: Academic Press, 1993. 573 p.

HORVAT, Tomislav; JOB, Josip. The use of machine learning in sport outcome prediction: A review. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 10, n. 5, p. e1380, 2020.

JAMES, Gareth et al. **An introduction to statistical learning**. New York: springer, 2013.

JERGENSEN, G. Eli et al. Classifying convective storms using machine learning. **Weather and Forecasting**, v. 35, n. 2, p. 537-559, 2020.

KELLEY, Jason et al. Using machine learning to integrate on-farm sensors and agro-meteorology networks into site-specific decision support. **Transactions of the ASABE**, v. 63, n. 5, p. 1427-1439, 2020.

KUČAK, Danijel; JURIČIĆ, Vedran; ĐAMBIĆ, Goran. MACHINE LEARNING IN EDUCATION-A SURVEY OF CURRENT RESEARCH TRENDS. **Annals of DAAAM & Proceedings**, v. 29, 2018.

LI, Jiaming et al. Machine learning for solar irradiance forecasting of photovoltaic system. **Renewable energy**, v. 90, p. 542-553, 2016.

LIAKOS, Konstantinos G. et al. Machine learning in agriculture: A review. **Sensors**, v. 18, n. 8, p. 2674, 2018.

LIMA, Kellen Carla. **Descargas elétricas atmosféricas em sistemas convectivos de mesoescala no sul da América do Sul**. 2005. Dissertação de Mestrado. Universidade Federal de Pelotas.

LIU, Hongwei; YUE, Fange; XIE, Zhouqing. Quantify the role of anthropogenic emission and meteorology on air pollution using machine learning approach: A case study of PM2. 5 during the COVID-19 outbreak in Hubei Province, China. **Environmental Pollution**, v. 300, p. 118932, 2022.

MACEDO, Suzana R. et al. Monitoramento de sistemas convectivos de mesoescala atuantes no Brasil utilizando o FORTRACC (Forecast and Tracking of Active and Convective Cells). **Congr. Bras. Meteor**, v. 13, p. 2004, 2004.

MACEDO, Suzana Rodrigues et al. Monitoramento e evolução de descargas elétricas atmosféricas associadas a sistemas convectivos de mesoescala. **Boletim da Soc. Bras. Meteorologia**, v. 29, n. 3, p. 67-71, 2005.

MACGORMAN, Donald R.; RUST, W. David; RUST, W. David. **The electrical nature of storms**. Oxford University Press on Demand, 1998.

MACHADO, L. B. T., ROSSOW, W. B. Structural characteristics and radiative properties of tropical cloud clusters. Monthly Weather Review, v. 121, pp. 3234–3260, 1993.

MACHADO, L. B. T., ROSSOW, W. B; GUEDES, R. L; WALKER, B. W. Life cycle variations of Mesoscale Convective Systems over the Americas. Monthly Weather Review, v. 126, p. 1630-1653, 1998.

MACHADO, Luiz Augusto T.; LAURENT, Henri. The convective system area expansion over Amazonia and its relationships with convective system life duration and high-level wind divergence. **Monthly weather review**, v. 132, n. 3, p. 714-725, 2004.

MADDOX, Robert A. Mesoscale convective complexes. **Bulletin of the American Meteorological Society**, p. 1374-1387, 1980.

MARTINS, Jorge A. et al. Climatology of destructive hailstorms in Brazil. **Atmospheric Research**, v. 184, p. 126-138, 2017.

MATTOS, Enrique Vieira. Relações das propriedades físicas das nuvens convectivas com as descargas elétricas. **Mestrado em meteorologia, Instituto Nacional de Pesquisas Espaciais, São José dos Campos**, 2009.

MATTOS, Enrique V.; MACHADO, Luiz AT. Cloud-to-ground lightning and Mesoscale Convective Systems. **Atmospheric Research**, v. 99, n. 3-4, p. 377-390, 2011.

MECIKALSKI, John R. et al. Regional comparison of GOES cloud-top properties and radar characteristics in advance of first-flash lightning initiation. **Monthly weather review**, v. 141, n. 1, p. 55-74, 2013.

MOORE, David S.; KIRKLAND, Stephane. **The basic practice of statistics**. New York: WH Freeman, 2007.

MORAES, Flávia Dias de Souza. Ambiente atmosférico favorável ao desenvolvimento de Complexos Convectivos de Mesoescala no Sul do Brasil. **Programa de Pós-Graduação em Geografia, Universidade Federal do Rio Grande do Sul, Instituto de Geociências**, 2016.

NAIR, Arshad Arjunan et al. Machine Learning Uncovers Aerosol Size Information From Chemistry and Meteorology to Quantify Potential Cloud-Forming Particles. **Geophysical Research Letters**, v. 48, n. 21, p. e2021GL094133, 2021.

NEWTON, Chester W. Structure and mechanism of the prefrontal squall line. **Journal of Atmospheric Sciences**, v. 7, n. 3, p. 210-222, 1950.

ODA, Paula SS et al. An initial assessment of the distribution of total Flash Rate Density (FRD) in Brazil from GOES-16 Geostationary Lightning Mapper (GLM) observations. **Atmospheric Research**, v. 270, p. 106081, 2022.

Observing Systems Capability Analysis and Review Toll 2022. Disponível em: https://www.wmo-sat.info/oscar/satellites/view/149. Acesso em 21 de maio de 2022.

PETERSON, Michael; RUDLOSKY, Scott; ZHANG, Daile. Changes to the appearance of optical lightning flashes observed from space according to thunderstorm organization and

structure. **Journal of Geophysical Research: Atmospheres**, v. 125, n. 4, p. e2019JD031087, 2020.

PINTO Jr., O.; PINTO, I. R. C. B. Tempestades e relâmpagos no Brasil - São José dos Campos: Instituto Nacional de Pesquisas Espaciais, 2000.

PINTO JR, Osmar. **A arte da guerra contra os raios**. Oficina de Textos, 2005.

POCKELS, F. Ueber das magnetische Verhalten einiger basaltischer Gesteine. **Annalen der Physik**, v. 299, n. 13, p. 195-201, 1897.

PURDOM, JAMES FW; MENZEL, W. PAUL. Use in Meteorology. **Historical Essays on Meteorology, 1919-1995: The Diamond Anniversary History Volume of the American Meteorological Society**, p. 99, 1996.

RAJKOMAR, Alvin; DEAN, Jeffrey; KOHANE, Isaac. Machine learning in medicine. **New England Journal of Medicine**, v. 380, n. 14, p. 1347-1358, 2019.

RAKOV, Vladimir A.; UMAN, Martin A. **Lightning: physics and effects**. Cambridge university press, 2003.

REYNOLDS, S. E.; BROOK, M.; GOURLEY, Mary Foulks. Thunderstorm charge separation. **Journal of Atmospheric Sciences**, v. 14, n. 5, p. 426-436, 1957.

RIBEIRO, Bruno Zanetti. Linhas de instabilidade no Sul do Brasil. **Doutorado em Meteorologia, Instituto Nacional de Pesquisas Espaciais. São José dos Campos**, 2018.

RICHTER, Chris; O'REILLY, Martin; DELAHUNT, Eamonn. Machine learning in sports science: challenges and opportunities. **Sports Biomechanics**, p. 1-7, 2021.

ROSSI, Alessio; PAPPALARDO, Luca; CINTIA, Paolo. A narrative review for a machine learning application in sports: an example based on injury forecasting in soccer. **Sports**, v. 10, n. 1, p. 5, 2021.

SALIO, Paola; NICOLINI, Matilde; ZIPSER, Edward J. Mesoscale convective systems over southeastern South America and their relationship with the South American low-level jet. **Monthly Weather Review**, v. 135, n. 4, p. 1290-1309, 2007.

SAMUEL, Arthur L. Machine learning. **The Technology Review**, v. 62, n. 1, p. 42-45, 1959.

SAUNDERS, Clive. Charge separation mechanisms in clouds. In: **Planetary Atmospheric Electricity**. Springer, New York, NY, 2008. p. 335-353.

SCHER, Sebastian; MESSORI, Gabriele. Predicting weather forecast uncertainty with machine learning. **Quarterly Journal of the Royal Meteorological Society**, v. 144, n. 717, p. 2830-2841, 2018.

SHAH, Dhruvil et al. Exploiting the capabilities of blockchain and machine learning in education. **Augmented Human Research**, v. 6, n. 1, p. 1-14, 2021.

SIDEY-GIBBONS, Jenni AM; SIDEY-GIBBONS, Chris J. Machine learning in medicine: a practical introduction. **BMC medical research methodology**, v. 19, n. 1, p. 1-18, 2019.

SPERLING, V. B. Processos Físicos e Elétricos das Tempestades de Granizo na Região Sul do Brasil. 2018.

STIRNBERG, Roland et al. Meteorology-driven variability of air pollution (PM 1) revealed with explainable machine learning. **Atmospheric Chemistry and Physics**, v. 21, n. 5, p. 3919-3948, 2021.

TAKAHASHI, Tsutomu. Riming electrification as a charge generation mechanism in thunderstorms. **Journal of Atmospheric Sciences**, v. 35, n. 8, p. 1536-1548, 1978.

UMAN, Martin A.; KRIDER, E. Philip. Natural and artificially initiated lightning. **Science**, v. 246, n. 4929, p. 457-464, 1989.

VILA, Daniel Alejandro et al. Forecast and Tracking the Evolution of Cloud Clusters (ForTraCC) using satellite infrared imagery: Methodology and validation. **Weather and Forecasting**, v. 23, n. 2, p. 233-245, 2008.

VONNEGUT, B. How the external currents flowing to a thundercloud influence its electrification. In: **Annales geophysicae**. Copernicus, 1991. p. 34-36.

VOYANT, Cyril et al. Machine learning methods for solar radiation forecasting: A review. **Renewable Energy**, v. 105, p. 569-582, 2017.

WALLACE, JOHN M.; HOBBS, P. V. Atmosphere science-an introductory survey. **Atmosphere science-an introductory survey**, p. V, 1977.

WALLACE, John M.; HOBBS, Peter V. **Atmospheric science: an introductory survey**. Elsevier, 2006.

WILLIAMS, Earle R. The electrification of thunderstorms. **Scientific American**, v. 259, n. 5, p. 88-99, 1988.

WILLIAMS, Earle R. The tripole structure of thunderstorms. **Journal of Geophysical Research: Atmospheres**, v. 94, n. D11, p. 13151-13167, 1989.

WINSTON, Patrick Henry. **Artificial intelligence**. Addison-Wesley Longman Publishing Co., Inc., 1992.

YING, Li; HUAILIANG, Chen. Review of machine learning approaches for modern agrometeorology. **Revista de Meteorologia Aplicada** v. 31, n. 3, p. 257-266, 2020.

ZHOU, Kanghui et al. A deep learning network for cloud-to-ground lightning nowcasting with multisource data. **Journal of Atmospheric and Oceanic Technology**, v. 37, n. 5, p. 927-942, 2020.

ZHOU, Yong et al. A review on global solar radiation prediction with machine learning models in a comprehensive perspective. **Energy Conversion and Management**, v. 235, p. 113960, 2021.

ZIPSER, Edward J. The evolution of mesoscale convective systems: Evidence from radar and satellite observations. **Tropical Rainfall Measurements**, v. 159, p. 166, 1988.