UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
Programa de Pós-Graduação em Engenharia Ambiental
Campus Londrina e Apucarana

MARCOS PAULO GUIMARÃES GUERRA

SISTEMA INFORMATIZADO DE MINERAÇÃO PARA OBTENÇÃO DE DADOS REFERENTES A OCORRÊNCIAS DE TEMPESTADES DE GRANIZO NA REGIÃO SUL E ESTADO DE SÃO PAULO

LONDRINA
2023

MARCOS PAULO GUIMARÃES GUERRA

SISTEMA INFORMATIZADO DE MINERAÇÃO PARA OBTENÇÃO DE DADOS REFERENTES A OCORRÊNCIAS DE TEMPESTADES DE GRANIZO NA REGIÃO SUL E ESTADO DE SÃO PAULO

COMPUTERIZED MINING SYSTEM TO OBTAIN DATA REGARDING THE OCCURRENCES OF HAILSTORMS IN THE SOUTHERN REGION AND STATE OF SÃO PAULO

Trabalho de conclusão de curso de Dissertação apresentada como requisito para obtenção do título de Mestre em Engenharia Ambiental pelo programa de Pós-Graduação em Engenharia Ambiental da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Saneamento Ambiental.

Orientador: Prof. Dr. Jorge Alberto Martins.

Coorientador: Prof. Dr. Marcos Vinícius Bueno de Morais.

LONDRINA

2023

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Londrina

MARCOS PAULO GUIMARAES GUERRA

SISTEMA INFORMATIZADO DE MINERAÇÃO PARA OBTENÇÃO DE DADOS REFERENTES A OCORRÊNCIAS DE TEMPESTADES DE GRANIZO NA REGIÃO SUL E ESTADO DE SÃO PAULO

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Engenharia Ambiental da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Engenharia Ambiental.

Data de aprovação: 28 de Abril de 2023

Dr. Alessandro Botelho Bovo, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Edio Roberto Manfio, Doutorado - Faculdade de Tecnologia de Garça (Fatecga)

Dr. Marcos Vinicius Bueno De Morais, Doutorado - Universidade Tecnológica Federal do Paraná

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 28/04/2023.

Dedico este trabalho a minha querida mãe e pai, minha amada namorada Jéssica e minha irmã Bruna por toda paciência e apoio nos momentos mais difíceis.

.

AGRADECIMENTOS

Gostaria de expressar minha mais profunda gratidão à Universidade Tecnológica Federal do Paraná (UTFPR), ao meu estimado orientador Professor Dr. Jorge Alberto Martins, e ao meu coorientador Professor Dr. Marcos Vinicius por sua orientação, apoio e expertise inestimáveis durante minha jornada de pesquisa. Sem o comprometimento e dedicação deles, essa conquista não seria possível.

Também gostaria de estender minha sincera gratidão ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) processo (#423316/2018-4 e #381250/2020-2), e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa concedida. Seu apoio financeiro foi fundamental para o desenvolvimento do meu trabalho e para a realização deste objetivo.

Por fim, gostaria de agradecer a todos os colegas, amigos e familiares que me apoiaram e me encorajaram ao longo deste percurso, em especial a minha mãe, Maria, minha irmã Bruna e namorada e futura esposa Jessica. Vocês foram uma fonte constante de inspiração e motivação, e eu sou grato por ter compartilhado esta experiência com cada um de vocês.

"A mente que se abre a uma nova ideia
jamais voltará ao seu tamanho
original." - Albert Einstein.

RESUMO

As tempestades de granizo podem causar danos em várias regiões do mundo. A região sul do Brasil é apontada como uma das áreas com maior incidência de granizo destrutivo. Os danos causados pelo granizo destrutivo não se limitam a áreas rurais. Também afetam áreas urbanas, com veículos sendo um dos principais itens danificados. Algumas tempestades de granizo são tão desastrosas que forçam os municípios a declarar estado de emergência ou calamidade pública. A relação histórica entre danos causados pelo granizo e mudanças climáticas é mencionada por muitos autores, sendo a região da tríplice fronteira entre Paraná, Santa Catarina e Argentina apontada como um local de grande atividade relacionada a eventos de granizo e atmosfera potencialmente favorável para a formação de tempestades. Além disso, algumas projeções indicam um possível agravamento dos danos causados por tempestades de granizo. O conhecimento prévio estes eventos podem ajudar a se preparar para possíveis emergências, mas sua detecção é uma tarefa muito complexa. O avanço da tecnologia permitiu a criação/melhoria de vários dispositivos eletrônicos, como sensores, novos tipos de armazenamento de dados e o aprimoramento de equipamentos de processamento, resultando em um aumento na produção de dados de vários tipos, incluindo dados meteorológicos. Juntamente com isso, técnicas de mineração de dados e algoritmos para encontrar padrões frequentes em um conjunto de dados podem ser ferramentas essenciais para melhorar a análise de tempestades de granizo e outros eventos meteorológicos extremos. Com o objetivo de analisar regras de associações de eventos meteorológicos, este estudo utilizou técnicas de mineração de padrões, em dados que foram coletados de forma automatizada em três plataformas diferentes: REDEMET, IPMET/Unesp e S2ID. O método utilizado para realizar a coleta automatizada, foi utilizando técnicas de *web scraping*. Posteriormente, os dados coletados foram processados e salvos em um banco de dados MongoDB. Após uma quantidade significativa de dados centralizados das três plataformas selecionadas, esses dados foram preparados para uso com o algoritmo FP-Growth, responsável por extrair padrões frequentes em conjuntos de dados, resultando na criação das tabelas de regras de associação. No final do estudo, um painel (*dashboard*) foi desenvolvida para apresentar os dados coletados e as regras de associação entre eventos meteorologicos extremos e o granizo. Após a análise das regras e com base nos resultados obtidos, constatou-se que a centralização dos dados é fundamental para verificar a associação do granizo com outros eventos atmosféricos. Assim, o banco de dados desenvolvidos pode auxiliar na previsão de eventos de granizo futuro utilizando modelos de inteligência artificial.

Palavras-chave: eventos extremos; desenvolvimento de sistema/*software;* aprendizado de máquina supervisionado; mineração de dados; tempestades de granizo

ABSTRACT

Hailstorms can cause damage in various regions of the world. The southern region of Brazil is identified as one of the areas with the highest incidence of destructive hail. The damage caused by destructive hail is not limited to rural areas. It also affects urban areas, with vehicles being one of the main items damaged. Some hailstorms are so disastrous that they force municipalities to declare a state of emergency or public calamity. The historical relationship between hail damage and climate change is mentioned by many authors, with the tri-border region between Paraná, Santa Catarina, and Argentina identified as a location of high activity related to hail events and potentially favorable atmosphere for storm formation. In addition, some projections indicate a possible worsening of the damage caused by hailstorms. Prior knowledge of these events can help prepare for possible emergencies, but their detection is a very complex task. The advancement of technology has allowed for the creation/improvement of various electronic devices, such as sensors, new types of data storage, and the improvement of processing equipment, resulting in an increase in the production of various types of data, including meteorological data. Along with this, data mining techniques and algorithms for finding frequent patterns in a data set can be essential tools for improving the analysis of hailstorms and other extreme weather events. With the aim of analyzing rules of associations of meteorological events, this study used pattern mining techniques on data that were collected in an automated manner on three different platforms: REDEMET, IPMET/Unesp, and S2ID. The method used to perform automated collection was using web scraping techniques. Subsequently, the collected data was processed and saved in a MongoDB database. After a significant amount of centralized data from the three selected platforms, this data was prepared for use with the FP-Growth algorithm, responsible for extracting frequent patterns in data sets, resulting in the creation of association rule tables. At the end of the study, a dashboard was developed to present the collected data and the association rules between extreme weather events and hail. After analyzing the rules and based on the results obtained, it was found that centralizing the data is essential to verify the association of hail with other atmospheric events. Thus, the developed database can help in predicting future hail events using artificial intelligence models.

Keywords: extreme events; system/software development; supervised machine learning; data mining; hailstorms

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE SIGLAS E ABREVIATURAS

AC	Acurácia
AM	Aprendizagem de Máquina
ANS	Aprendizado não supervisionado
API	Application Programming Interface
AR	Análises de Regressão
ARIMA	Auto Regressive Integrated Moving Average
ARM	Regras de associação
BP	Back Propagation
Bits	Binary digits
CNDC	Centro Nacional de Dados Climáticos dos EUA
CSS	Cascading Style Sheets
DM	Mineração de Dados
ES	Engenharia de Software
et al.	Outros autores
FA	Floresta Aleatória
FP-Tree	Árvore de Frequência de Padrões
HTML	HyperText Markup Language
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
IDE	Integrated Development Environment
IHC	Interface Humano Computador
IPMET	Instituto de Pesquisas Meteorológicas
METAR	Meteorological Aerodrome Report
MVT	Model View Template
NoSQL	Not Only SQL
PR	Precisão
RAM	Random-access memory
REDEMET	Rede de Meteorologia do Comando da Aeronáutica
RL	Regressão Linear
RNA	Rede Neural Artificial
RNAs	Redes Neurais Artificiais
SB	Sensibilidade
SI	Sistema Informatizado
SGBD	Sistema Gerenciador de Banco de Dados
SQL	Structured Query Language
S2ID	Sistema Integrado de Informações sobre Desastres
UTFPR	Universidade Tecnológica Federal do Paraná
URL	Uniform Resource Locator
WC	Web Crawler
WS	Web Scraping

SUMÁRIO

1 INTRODUÇÃO

Eventos atmosféricos extremos podem causar danos em várias regiões do mundo. Um dos eventos atmosféricos extremos a destacar são as tempestades de granizo, que ocorrem em várias regiões do planeta. Botzen, Bouwer e Van den Bergh (2010) observaram uma relação histórica entre o aumento dos danos causados pelo granizo e mudanças climáticas.

Tempestades de granizo destrutivo (diâmetro maior que 2,5 cm), tornados, rajadas de vento e enchentes, tem aumentado, podendo causar grande impacto e destruição por onde passam (PREIN; HOLLAND, 2018a). Além disso, Mc Philips et al. (2018) destacam que tais eventos podem causar impactos sociais, ecológicos e danos em estruturas tecnológicas como *data centers*, redes de telefonia, *internet*, e rede elétrica.

Muitas regiões da Europa sofrem com consequentes tempestades de granizo destrutivo, causando danos a plantações, construções e veículos. Especificamente na Alemanha, o conhecimento sobre o local estimado das tempestades é bem limitado, então um estudo foi realizado observando dados coletados de um radar polarizado, do ano de 2005 até 2011 (PUSKEILER; KUNZ; SCHMIDBERGER, 2016). Na Itália, através da climatologia de granizo medidos por hailpads (MANZATO et al., 2022), demonstra que nos últimos anos houve menos ocorrências de tempestades de granizo, porém com pedras consideravelmente grandes.

No Brasil, a região Sul do país se destaca tendo grande incidência de granizo destrutivo (MARTINS et al., 2017). Além desta região, imagens de satélites indicam que o Sudeste da América do Sul é um dos locais de granizo mais intensos no mundo (BEAL et al., 2020a).

Na Tríplice fronteira do estado do Paraná, do estado de Santa Catarina e da Argentina, é possível observar grande atividade relacionada a eventos de granizo e atmosfera potencialmente favoráveis a formação de tempestades severas (Brooks et al., 2003, apud BEAL et al., 2020).

Especificamente, as tempestades de granizo são potenciais causadores de perdas e prejuízos em culturas agrícolas, destacando-se a destruição de plantas, o desfolhamento de espécies perenes, além de danos a grãos e quedas de frutos (CALDANA; NITSCHE; CARAMORI, 2019). Porém, os estragos não se limitam

apenas as regiões rurais. Isto porque, as áreas urbanas são igualmente afetadas (HOHL; SCHIESSER; KNEPPER, 2002). Algumas tempestades de granizo podem ser tão desastrosas que obrigam os municípios a decretarem estado de emergência ou calamidade pública, sendo que o primeiro se caracteriza como uma situação menos grave, com danos suportáveis e superáveis pela comunidade. Enquanto, na segunda situação, o dano também traz risco à vida, sendo necessário o auxílio do governo e órgãos externos.

Neste sentido, uma possível solução a ser estudada é a previsão antecipada de mencionados eventos, de modo a possibilitar a transmissão de informações com antecedência à população para que tomem as medidas necessárias, a fim de reduzir os consequentes impactos.

No entanto, esta não é uma tarefa simples e trivial, existindo uma grande complexidade em se detectar tempestades de granizo. Sabe-se que o granizo surge, especialmente, em locais com clima temperado (BEAL et al., 2020b). Entretanto, é preciso uma melhor compreensão sobre a sua formação para avançar nos estudos envolvendo previsão e antecipação dos danos e perigos ocasionados.

Estudar as tempestades de granizo exige coleta de muitos dados (RAUPACH et al., 2021), os quais são extremamente difíceis de se obter, principalmente em regiões muito especificas que não possuem radares, aeroportos ou estações meteorológicas, dificultando a criação de uma base de dados centralizada sobre o granizo e, consequentemente, a criação de modelos estatísticos robustos.

O avanço da tecnologia permitiu a criação/melhoria de vários dispositivos eletrônicos, como sensores, novos tipos de armazenamento e o aprimoramento de equipamentos de processamento. Consequentemente, consegue-se produzir e armazenar gigantescos volumes de dados advindos de diversas fontes, sendo os dados a matéria-prima principal para o aprimoramento e funcionamento dos algoritmos de Inteligência Artificial (IA) (PULLMAN et al., 2019). Juntamente a isso, técnicas de mineração de dados e algoritmos para encontrar padrões frequentes em um conjunto de dados podem ser ferramentas essenciais para melhorar a análise de tempestades de granizo e outros eventos climáticos extremos.

Atualmente, os computadores possuem grande poder de processamento, permitindo o desenvolvimento e aprimoramento de diversas ferramentas, das quais os algoritmos de aprendizagem de máquina (AM) podem prever variáveis relevantes

quanto aos eventos de granizo (MARTIUS et al., 2018), razão pelo qual se faz necessário o desenvolvimento de um sistema/*software* próprio.

Observar o granizo e suas associações com outros fenômenos podem ser um conhecimento relevante para futuras análises meteorológicas e previsões estatísticas. A utilização de algoritmos de AM (RASCHKA, 2015) pode possibilitar a criação de tabelas de associações entre fenômenos climáticos, tais como o granizo acompanhado de eventos como chuvas, tempestades, tornados e diversos outros fenômenos.

Assim faz-se necessário, utilizar técnicas de coleta automatizada como *web scraping* para*,* posteriormente, centralizar e organizar os dados obtidos. Após uma quantidade significativa de dados centralizados, pode ser possível extrair padrões frequentes em conjuntos de dados, resultando na criação das tabelas de regras de associação. Com a criação das tabelas de associação, é de suma importância a criação de uma ferramenta que possibilite a visualização de tais dados extraídos, os sintetizando de forma fácil para exibir informações relevantes.

Manfio et al. (2018) desenvolveram um projeto que utilizava técnicas de AM e coleta de dados de forma automatizada, chamado de projeto Solar-Sima Prototype, que incorporou uma dashboard de monitoramento, da qual permitia visualizar e controlar as informações referentes a um sistema fotovoltaico. Outros projetos envolvendo energias renováveis, foram conduzidos, utilizando conceitos de monitoramento por dashboard, utilizando conceitos de Interface Humano-Computador (IHC) (MANFIO; GUERRA; MORENO, 2016). Manfio, Guerra, Mancuzo, (2022) utilizaram uma dashboard para monitorar comandos por voz enviados para um inversor de tensão, através da dashboard era possível verificar a situação do inversor, se ele havia recebido o comando por voz e já tinha o executado.

Uma dashboard de monitoramento para observar as informações geradas pelas bases centralizadas é de suma importância, levando em consideração que quando os dados estão centralizados, poderá ser possível observar informações diferentes em relação as bases de dados separadas.

2 REFERENCIAL TEÓRICO

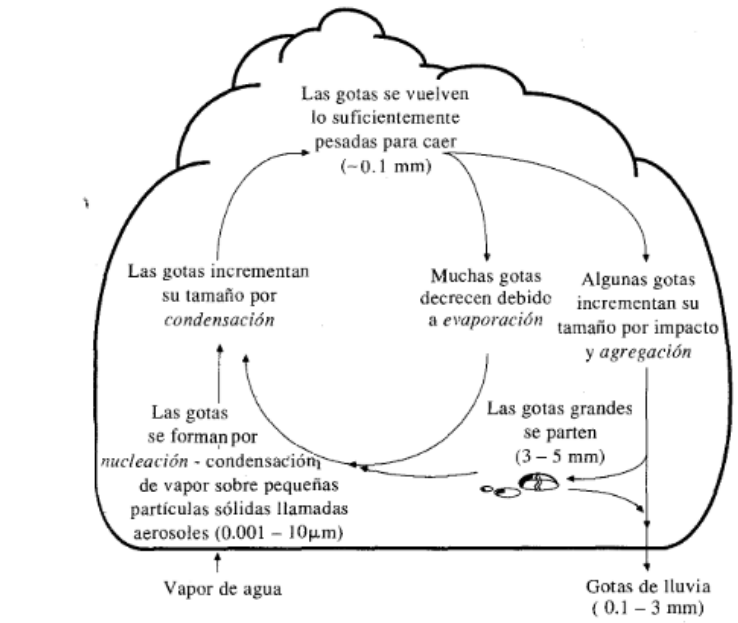
A princípio, serão abordadas questões relativas à formação do granizo e seus impactos. Ao final, alguns aspectos das tecnologias empregadas no desenvolvimento do projeto passarão a ser estudadas, bem como alguns trabalhos realizados envolvendo coleta e mineração de dados empregando a utilização de técnicas de *web scraping* e a utilização do algoritmo FP-Growth.

2.1 Impactos do granizo

Vários estudos são conduzidos a fim de se entender como o granizo surge, sua formação, onde poderá ocorrer o próximo evento, entre outros pontos. Para tanto, a tecnologia tem sido muito utilizada para tentar responder tais questionamentos.

Uma precipitação se forma a partir da condensação do ar que sobe e resfria (Figura 1). Se a temperatura na altura da nuvem se encontra abaixo do ponto de congelamento, há formação de cristais de gelo. Para acontecer este processo de condensação é necessário que exista um núcleo de condensação para permitir a união ou nucleação das moléculas de água. Algumas partículas presentes no ar, conhecidas como aerossóis, atuam como núcleo de condensação.

Figura 1 - Formação de precipitação no interior das nuvens



Fonte: Chow et al. (1994)

As tempestades de granizo ocorrem devido a instabilidades atmosféricas e geralmente estão associadas a nuvens do tipo Cumulonimbus. As pedras de granizo são formadas nas partes superiores das nuvens onde a temperatura é baixa, causando a transformação de gotas de água em partículas de gelo (FELIPE et al., 2019). O granizo pode surgir, especialmente, em locais com clima temperado (BEAL et al., 2021a). A obtenção de dados observacionais de granizo é extremamente recente (MARTINS et al., 2017), além das bases de dados encontradas não estarem centralizadas, dificultando o acesso aos mesmos para pesquisas cientificas.

Imagens de satélite podem sugerir regiões em que as taxas de ocorrência de tempestades de granizo são mais elevadas (GALLO et al., 2019). Estudos indicam que a formação do granizo é resultado de um complexo processo químico (MICHAEL; STUART, 2009; SELAN et al., 2014; MA et al., 2013). Beal et a., (2021b) avaliaram a composição química de amostras de granizo na América do Sul, incluindo a concentração de íons, metais e metaloides. As amostras foram limpas com água ultrapura para remover contaminantes. Os resultados mostraram que as concentrações de íons seguiram uma sequência específica, com Ca2+ sendo o mais abundante (BEAL et al., 2021b).

Como as tempestades de granizo são eventos localizados, sua medição e detecção são desafiadoras, o que significa que muitos eventos passam despercebidos. Uma das maneiras de se medir as ocorrências de granizo é através dos *hailpads*, que em geral são placas de isopor coberto com algum material, como tinta ou papel alumínio, permitindo a aferição do impacto deixado pelas pedras de granizo (GRIESER; HILL, 2019; MARCOS et al., 2021; DIELING; SMITH; BERUVIDES, 2021). As características de pedras de granizo são exploradas utilizando tanto observações de superfície, como conjuntos de dados observacionais de detecção remota, também conhecidos como técnicas de sensoriamento remoto (ALLEN et al., 2020). Tempestades de granizo ocorrem em várias regiões do planeta, algumas com mais intensidade que outras. O tamanho, intensidade e duração das tempestades de granizo determinam o grau de gravidade dos impactos (FELIPE et al., 2019).

Na Argentina, entre os meses de outubro de 2018 e abril de 2019, foram realizadas as campanhas de campo, conhecidos como RELÂMPAGO e CACTI, que coletaram dados através dos *hailpads* e outros instrumentos, como radar Doppler,

imagens de satélite e sondagens atmosféricas (BECHIS et al., 2022). Durante o período de observação, em novembro de 2018, foi relatado granizo de 4 cm de diâmetro em amplas áreas da província de Mendoza (BECHIS et al., 2022).

Processos referentes ao desenvolvimento e à distribuição das chuvas de granizo a nível mundial têm sido revistas por meio da análise das características físicas e microfísicas, a fim de entender os motivos que levam a formação do granizo e o que determina como ele cai ao chão (ALLEN et al., 2020). Na China, por exemplo, eventos de granizo mostraram uma diminuição entre o período de 1981 a 2010, e um aumento entre 2011 e 2021 (BIAN et al., 2023). Na Itália, nos últimos anos, nota-se uma diminuição na frequência de tempestades de granizo, porém um aumento no tamanho das pedras (MANZATO et al., 2022b). Na região mediterrânea da Europa, Kim et al., (2022) mostram que houve um aumento no número de eventos de granizo entre 2010 e 2021, comparado ao período de 1999-2010. Os impactos do granizo a culturas agrícolas se estendem também em regiões da Europa. Hermida et al. (2013) observaram que as chuvas de granizo ocorrem principalmente no verão-outono europeu, especialmente nas áreas continentais.

Allen et al., (2020) mostram que os dados de observação de granizo registrados pelo Centro Nacional de Dados Climáticos dos EUA (CNDC) é o conjunto de dados mais amplo disponível. Porém, os relatórios de granizo deste centro são limitados em relação ao local, pois a maior parte dos dados meteorológicos são relacionados ao seu país de origem.

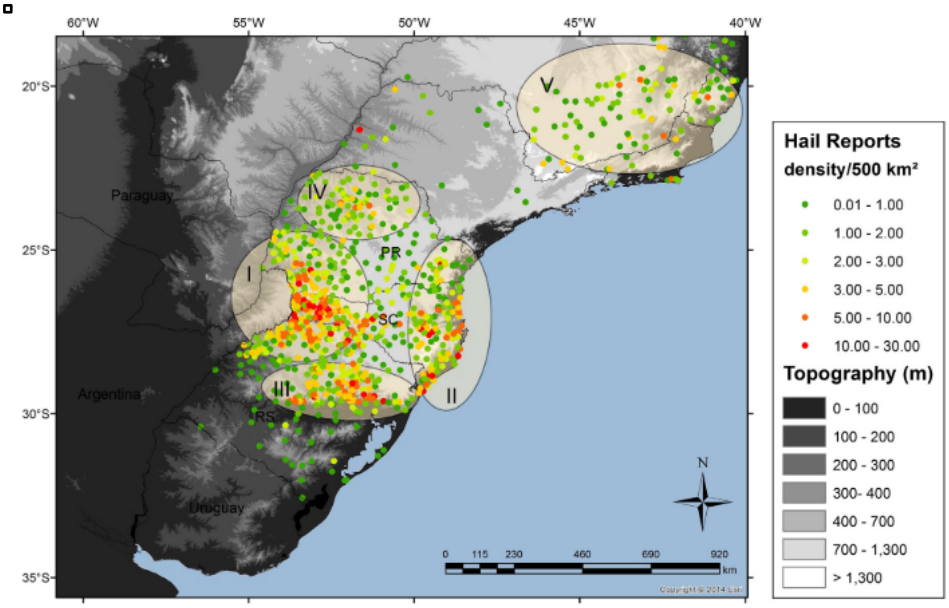
Observações recentes de alta resolução espacial do granizo sugerem tamanhos de granizo bem maiores do que aqueles especificados nos mencionados relatórios dos Dados de Tempestades do CNDC, assim como tem se tornado forte tendência de granizos gigantes em áreas com maior densidade populacional (ALLEN; TIPPETT; SOBEL, 2015).

Prein; Holland, (2018b) relatam um aumento das perdas econômicas causadas por fenômenos meteorológicos extremos, como granizo, tornados, rajadas de vento e inundações, e a dificuldade em avaliar a relação entre esses eventos e as mudanças climáticas devido às limitações em observações, modelagem e compreensão dos processos físicos.

No Brasil, as regiões com forte predominância de culturas agrícolas sofrem muitos prejuízos com o granizo destrutivo, implicando em possíveis perdas de safras

inteiras ou potencialmente parte delas. Em geral, tempestades de granizo são mais frequentes no final da tarde, início da noite, no período de transição entre o inverno e a primavera (MARTINS et al., 2017). Com relação aos granizos considerados destrutivos, as regiões mais afetadas no país são os estados da região Sul e Sudeste, conforme mostrado na Figura 2.

Figura 2 - Mapa que indica a notificação de granizos destrutivos entre o período de 1991 e 2012



Fonte: Martins et al., 2017

Carvalho et al. (2020) analisaram os impactos de chuvas de granizo na produção agrícola brasileira, entre os anos de 2002 e 2018. Os autores verificaram que granizos e outros eventos climáticos extremos, provocaram grande prejuízo, reduzindo a produção agrícola e, consequentemente, podendo trazer sérias consequências ao acesso dos alimentos à população e a economia brasileira (DE CARVALHO et al., 2020).

Shah et al., (2021) identificaram prejuízos em culturas agrícolas, abordando que a variabilidade climática e eventos extremos causados pelo aquecimento global impactam negativamente a produção agrícola, aumentando a fome global e as crises alimentares, principalmente em regiões mais vulneráveis. Mudanças bruscas de temperatura, aumento da umidade e eventos climáticos não extremos também podem afetar o desenvolvimento das plantações, a produção e a renda dos agricultores (SHAH; HELLEGERS; SIDERIUS, 2021).

Felipe et al., (2019) identificaram que os eventos de granizo na região

Mesorregião Centro-Sul Paranaense (MRCSP) são localizados e altamente variáveis devido a fatores como latitude, longitude, variabilidade de precipitação, topografia e altitude. Áreas com diferenças significativas de altitude em curtas distâncias, especialmente nas encostas norte e sul do Rio Iguaçu, foram as mais afetadas pelo granizo (FELIPE et al., 2019).

Hermida et al. (2013) relatam que dado os danos causados pelas tempestades de granizo, em especial àquelas consideradas destrutivas, a antecipação da ocorrência destes eventos faz-se necessária.

Como as características dinâmicas da formação de granizo são extremamente difíceis de se prever, o uso de ferramentas estatísticas pode ser uma alternativa para auxiliar na previsão destas tempestades. Assim, é importante conhecer as técnicas e as ferramentas de aprendizagem de máquina.

2.2 Inteligência artificial

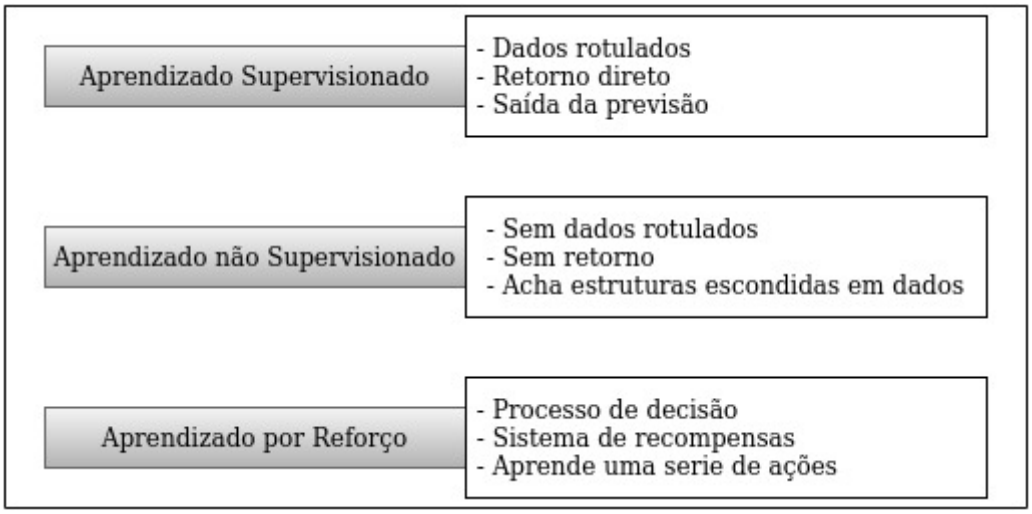
2.2.1 Aprendizagem de máquina (AM)

Por volta da segunda metade do século XX, a AM passou a ser um subcampo da Inteligência Artificial (IA). Em adição a esse fato, um dos recursos mais abundantes do século XXI é a grande quantidade de dados gerados diariamente.

Inicialmente, dados agrupados poderiam não significar nenhuma informação relevante. Algoritmos de AM tentam retirar informações relevantes e de difícil interpretação de grandes volumes de dados, utilizando-os, como por exemplo, para a criação de modelos, que posteriormente poderiam ser utilizados para realizar previsões (RASCHKA, 2017a).

Raschka (2017a) divide AM em três tipos: Aprendizado supervisionado; Aprendizado não supervisionado; e Aprendizado por reforço, sendo que cada um possui características diferentes e geralmente sendo empregados a tipos de problemas distintos. A Figura 3 destaca as principais características.

Figura 3 - Os três tipos de AM.



Fonte: Raschka (2017a)

2.2.1.1 Aprendizado supervisionado

O objetivo do aprendizado supervisionado (AS) é treinar um modelo de previsão com base em dados já rotulados, ou seja, é criado um modelo baseado em uma divisão dos dados, no qual uma parte é destinada para o treinamento e o restante usado para avaliar o modelo criado.

No treino, os dados são rotulados e assim o algoritmo de AS já sabe como classificá-los e cria um modelo de previsão, ou seja, o resultado da classificação já é conhecido, daí o termo supervisionado. A outra parte dos dados que foi anteriormente separada, passa pelo modelo de previsão criado que tenta classificá-los e, após a classificação, avalia-se os erros e acertos da previsão (RASCHKA, 2017).

O modelo criado no treino pode ser multirrótulos ou de natureza binária. A título de exemplificação, novos e-mails que passarem em um modelo binário serão classificados em apenas duas situações (p. ex. Spam e não spam), enquanto o modelo multirrótulos possui várias classificações distintas, como no caso de tentar identificar quais números estão em uma imagem (RASCHKA, 2017).

A classificação é bem complexa, pois o modelo treinado precisa saber reconhecer e classificar números de 0 a 9, pois se reconhece apenas números de 0 a 8, em uma situação que tenha o 9, o número em questão não será reconhecido, uma vez que o modelo o desconhece. Assim, é necessário treinar um novo modelo onde os dados de treino contenham o número faltante (RASCHKA, 2017).

Existe um segundo tipo de AM supervisionada, a análise de regressão (AR). Na AR, o treino do modelo utiliza basicamente duas variáveis: uma chamada de explicativa e outra de resposta (ou exploratória), e a relação entre essas duas variáveis permite o modelo realizar previsões (MARSLAND, 2014; RASCHKA, 2017).

2.2.1.2 Aprendizado não supervisionado

No aprendizado não supervisionado (ANS) não se tem uma variável resposta, isto é, os dados não são rotulados e geralmente possuem estruturas desconhecidas ou escondidas. Na ANS existem algumas técnicas que são normalmente utilizadas para classificação e exploração dessas estruturas, sendo as mais comumente utilizadas a clusterização e a redução de dimensionalidade (RASCHKA, 2017a).

A clusterização é uma técnica de análise de dados muito utilizada para exploração de grandes volumes de dados, que consegue dividi-los em subgrupos de acordo com seu grau de similaridade. Empresas tendem a utilizar muito a clusterização, com intuito de encontrar padrões entre grupos de clientes, sendo especialmente útil para criar estratégias focadas para cada grupo em específico (RASCHKA, 2017a).

Na redução de dimensionalidade, o principal objetivo é reduzir o ruído e a dimensionalidade das observações, isto é, cada observação pode possuir muitas mensurações atribuídas, não apenas duas variáveis, como por exemplo x1, x2, e sim x1, x2, x3, x4…, a fim de obter mais informações relevantes e limpas. Utiliza-se a técnica de redução, pois dois dos desafios da AM é a necessidade de muito poder computacional e grandes espaços de armazenamento (RASCHKA, 2017a).

2.2.1.3 Aprendizado por reforço

No Aprendizado por reforço, busca-se desenvolver um agente, também chamado de sistema. O agente tem como objetivo maximizar o desempenho do algoritmo, por meio de interações com o ambiente.

Por exemplo, em um jogo de xadrez, o agente executa uma ação no ambiente (p. ex. mover uma peça), por sua vez essa ação tem uma consequência e a consequência da ação gera um retorno para o agente, normalmente chamada de recompensa.

A recompensa serve de reforço para o aprendizado do algoritmo de acordo com as definições que o programador sugeriu. No caso do xadrez, ganhar o jogo é um reforço bom e perder o jogo é um reforço ruim. Se o objetivo é a vitória, então o algoritmo, com base nas recompensas, tende a evitar ações que levem à derrota e maximiza as ações que levem à vitória, isso tudo em um processo cíclico de erro e acerto, ou seja, o algoritmo aprende errando e tentando novamente, sempre com o objetivo de aperfeiçoar a interação com o ambiente (RASCHKA, 2017).

Existem diversos algoritmos que realizam previsões estatísticas e que poderiam ser utilizados para previsão do granizo, como as Redes Neurais Artificiais (HSU; GUPTA; SOROOSHIAN, 1995; MCGOVERN et al., 2017), Floresta Aleatória (BREIMAN, 2001) e Regressão Linear (RENCHER; SCHAALJE, 2008). Entretanto, é necessário obter dados e informações de ocorrência de granizo antes de qualquer aplicação de modelos. Neste trabalho, utilizaremos a técnica de *web scrapping*.

2.3 Web Scrapping

O desenvolvimento tecnológico está aumentando rapidamente e muitas informações estão circulando na internet. Atualmente, todas as informações podem ser facilmente encontradas por meio de mecanismos de busca como o Google e o Bing. No entanto, as informações fornecidas por esses sites ou aplicativos de mídia social às vezes são sobrecarregadas, o que dificulta a busca por informações relevantes para as necessidades do usuário (DEWI; MEILIANA; CHANDRA, 2019). Nos últimos anos, com o crescimento da big data, as informações na internet estão aumentando rapidamente, tornando-se importante ter tecnologias que ajudem o usuário a obter informações de forma eficiente e fácil (ALSHAMMARI et al., 2021).

É necessário um sistema que possa buscar informações relevantes, filtrar e apresentá-las ao usuário. O *web scraping* (WS) é uma das técnicas mais populares para extrair dados ou conteúdo da web (DEWI; MEILIANA; CHANDRA, 2019).

O WS também são conhecidos como web crawlers (WC), robôs ou spiders, são ferramentas para coletar o conteúdo específico definido pelo programador, usuário.

Pode-se dizer, portanto, que *crawlers* são rastreadores de dados (BATSAKIS; PETRAKIS; MILIOS, 2009). Existem basicamente três tipos de WC, o focado, por semântica, e por aprendizado.

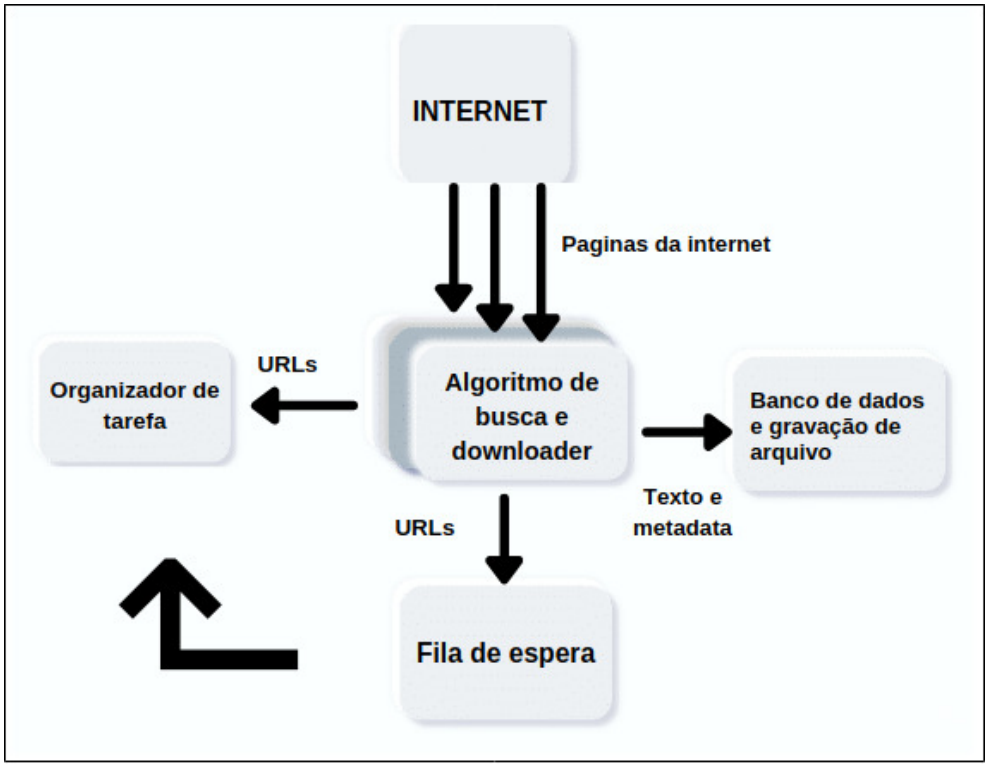
O WC focado indexa links e volta recursivamente salvando as páginas onde os links têm mais similaridade com os critérios estabelecidos pelo programador. Por exemplo: buscar por dados sobre tempestades de granizo, links que tenham granizo no nome serão salvos (BATSAKIS; PETRAKIS; MILIOS, 2009).

Similar ao WC focado, o WC por semântica possui a prioridade de download de páginas e é feita quando o conteúdo da página parece similar aos critérios estabelecidos pelo programador (BATSAKIS; PETRAKIS; MILIOS, 2009).

O WC por aprendizado utiliza o mesmo conceito de aprendizado por treinamento, onde inicialmente o WC é submetido a um treino e ele aprende quais páginas são relevantes, e quais não são. Inicialmente, o programador seleciona diversas páginas manualmente e as classifica para ensinar o algoritmo (BATSAKIS; PETRAKIS; MILIOS, 2009).

As etapas genéricas do funcionamento de um algoritmo de WC são demostradas na Figura 4 (BATSAKIS; PETRAKIS; MILIOS, 2009).

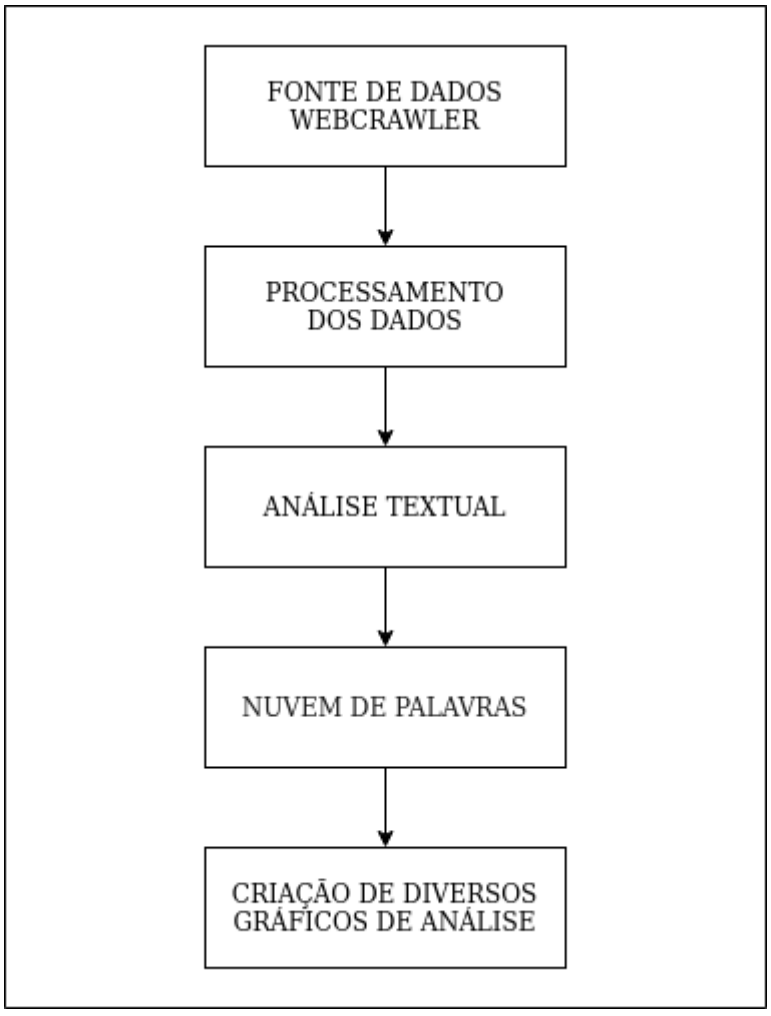
Figura 4 - Etapas genéricas do WC



Fonte: Adaptador de Batsakis; Petrakis; Milios., (2009)

Pesquisas nas mais diversas áreas tem utilizado o WC como ferramenta para busca de dados. Hu; Ge; Hou, (2014) realizaram um estudo que visava a análise de geo eventos, sobre um incidente na ilha Huangyan. Para isto, utilizou-se da coleta de dados por meio de um algoritmo de WC para coletar textos que eram discutidos em rede sociais, blogs, fóruns, tudo que remetia a assuntos sobre o incidente na ilha, que foi motivo de disputa territorial entre China e Filipinas. Com posterior análise desses dados, foi possível retirar muitas informações relevantes sobre como o mundo estava repercutindo o assunto (HU; GE; HOU, 2014). A Figura 5 demonstra o processo da etapa de coleta dos dados.

Figura 5 - Etapas estudo WC para análise de geo eventos



Fonte: Hu; Ge; Hou, (2014)

O estudo anterior utiliza a tecnologia do WC como ferramenta e tem a finalidade de buscar termos textuais e salvá-los em uma base de dados central, porém estudos com outras finalidades também são conduzidos utilizando tal tecnologia. Por exemplo, o monitoramento em tempo real de dados, não com a finalidade de

centralizá-los, e sim a de criar alertas em tempo real sobre assuntos que sejam relevantes de acordo com a necessidade da pesquisa.

Galaz et al., (2010) conduziram um estudo que objetivava utilizar o WC para realizar monitoramento em tempo real de dados na internet, que pudessem sugerir algum impacto ambiental, com a finalidade de auxiliar no monitoramento ecológico, já que é muito complicado para uma pessoa ficar 24 horas observando “sinais” que pudessem indicar algum dano ecológico ocorrendo em meio a um amontoado de informações pela internet. Esses sinais foram classificados de acordo com a necessidade de monitoramento, podendo ser, mas não se limitando a termos como, fogo, desmatamento e ou seca (GALAZ et al., 2010). A Figura 6 apresenta as etapas do estudo para monitoramento em tempo real.

Figura 6 - Etapas estudo monitoramento ecológico em tempo real

Fonte: Galaz et al., (2010)

Os dados coletados por WC geralmente são armazenados no estado que se encontram no momento da coleta, de modo que é muito importante a etapa de preparação de dados para o sucesso da utilização de inteligência artificial, destacando a necessidade de se obter dados limpos e organizados a partir de um grande volume de informações brutos (KUMAR; ROY, 2023).

Um bom exemplo de coleta de dados e preparação para utilização com técnicas de AM, foi um estudo realizado em 2023, onde o artigo trata sobre o problema da contaminação do ar e seu impacto na saúde humana, destacando a importância de monitorar a qualidade do ar para prevenir doenças.

Para isso, os autores utilizam a técnica de AM para criar um modelo preditivo de qualidade do ar, analisando tendências e padrões a partir de dados históricos e criando um modelo de previsão para valores futuros. O modelo proposto, chamado de Air Quality Prediction Model (AQPM), utiliza dados coletados do site do Central Pollution Control Board (CPCB) por meio da técnica de WC (KALAIVANI; KAMALAKKANNAN, 2023).

Alshammari et al., (2021) realizaram um estudo em 2021, onde implementaram um WC em Python para extrair dados de sites de destino e armazenar os dados coletados em um arquivo separado por vírgulas. Este sistema visava coletar informações sobre os produtos que os estudantes universitários precisam dos sites de destino e retorná-los aos usuários com uma página simples. Os usuários podem procurar o produto sobre o qual desejam obter informações e o WC extrai as seguintes informações (nome do produto, preço do produto, URL do produto) das seguintes lojas online: Amazon, eBay, Jarir e Extra, em seguida, armazenam-no no arquivo separado por vírgulas para análise de informações (ALSHAMMARI et al., 2021).

Outro estudo envolvendo a utilização de WC foi realizado por Naga Chandrika et al., (2020) que abordaram a importância de um WC para os mais diversos fins, muitas vezes precisamos buscar informações sobre cursos, reviews e dados para fornecer uma visão geral sobre eles. O estudo proposto apresenta um conceito de como integrar a extração de dados com PHP para fazer a atualização de informações executando o Python no servidor. O processo consiste em utilizar um script Python para extrair as informações necessárias de um site e, em seguida, enviar os dados para o servidor web hospedado na internet. O script PHP em execução no servidor recebe os dados do script Python e os armazena em um servidor MySQL. A cada execução do script Python, os dados do site são atualizados (NAGA CHANDRIKA et al., 2020).

Giambastiani et al., (2022) desenvolveram um WC com a finalidade de coleta de dados para pesquisas relacionadas a questões de ecossistema, avaliando a importância da presença de áreas verdes nas cidades e, em particular, das árvores, que fornecem múltiplos serviços ecossistêmicos. No entanto, as árvores estão sujeitas a estresses externos, como a decomposição e, consequentemente, a fenômenos de colapso. Para monitorar a estabilidade das árvores, foram realizadas numerosas investigações em relação aos aspectos fisiológicos e biomecânicos. O estudo dessa reconfiguração é uma análise complexa, e existem poucos trabalhos na literatura em

relação à grande variabilidade dada pela idade, espécie e posição das árvores. A fim de captar dados para o estudo, foram utilizadas as câmeras meteorológicas, aplicando técnicas de WC (GIAMBASTIANI et al., 2022).

Alguns tipos de dados são difíceis de se obter, dados meteorológicos por exemplo, dependem muito da localização de onde se deseja obtê-los, Fatmasari; et al., (2019), relatam a dificuldade de se obter dados meteorológicos detalhados e atualizados em algumas cidades da Indonésia, devido ao processo burocrático necessário para acessar as informações em diferentes instituições. Para resolver esse problema, os pesquisadores usaram a tecnologia de WC para coletar dados meteorológicos em tempo real de sites disponíveis na internet. Esses dados foram armazenados em um banco de dados para pesquisas futuras e para o desenvolvimento de um aplicativo de suporte à decisão baseado em dados meteorológicos (FATMASARI; KUNANG; PURNAMASARI, 2019).

No entanto, o acesso a fonte de grandes volumes de dados apresenta desafios, como a detecção de captcha, e necessidade de capacidade de computação intensiva e confiabilidade na extração de dados. A solução proposta por um estudo realizado por Chaulagain et al., (2017) buscou abordar a coleta de dados utilizando a biblioteca Python Selenium como técnica de WC, simulando um usuário real trabalhando com um navegador.

Ulfah; Najiah, (2023) discutiram a utilização da biblioteca Selenium como coletor de dados em sites de revistas científicas, o artigo tinha o objetivo de agilizar e facilitar a busca por artigos relevantes para uma pesquisa. O estudo foi realizado com o uso da linguagem de programação Python e de bibliotecas específicas, como Selenium e BeautifulSoup. Os dados coletados incluem título, imagem, perfil, link, número e classificação das revistas científicas (ULFAH; NAJIAH, 2023).

Manjari et al., (2020) abordou a técnica de sumarização automática de texto, que é amplamente utilizada para obter uma visão geral do conteúdo presente em vários documentos e para encontrar informações específicas na internet. O artigo propõe um novo processo para gerar um resumo de extração das informações com base na consulta do usuário, usando a extração de dados de vários sites na internet, com a ajuda do Selenium. Ao final da coleta, algoritmo *Term Frequency-Inverse Document Frequency* (TF-IDF) é aplicado para a sumarização do texto. (MANJARI et al., 2020).

A coleta de dados por meio de técnicas de WC, se estende para as mais

diversas áreas de pesquisas, Morais (2022) utilizou técnicas de mineração de texto e uma rotina em linguagem R para coletar os artigos científicos mais citados nas áreas de Psicologia e Educação que abordam o ensino de matemática para pessoas com Transtorno do Espectro Autista (TEA). Foram identificadas as frequências das palavras e realizados gráficos com os resultados obtidos.

Observou-se que os artigos científicos das duas áreas possuem diferenças e semelhanças em relação aos termos utilizados, com os artigos de Psicologia mais voltados ao ensino e aprendizagem e os de Educação mais voltados aos níveis do espectro autista. No entanto, ambos possuem o termo ‘estudante’ como o mais citado. É importante ressaltar que foi um estudo inicial, e foi sugerido que ele seja estendido para outras bases de dados científicas e áreas de pesquisa (MORAIS, 2022).

2.4 Algoritmo FP-Growth

O algoritmo fp-growth, utilizado para gerar modelos de regras de associação, foi proposto em por Han et al. (2000) onde utilizaram estrutura de dados da árvore de prefixos denominada árvore de frequência de padrões (FPTree). O algoritmo é muito utilizado em mineração de dados e análise de associação para encontrar padrões frequentes em um conjunto de dados.

O FP-Tree ou árvore de frequência, constrói uma árvore de frequência a partir do conjunto de dados de entrada, onde cada nó na árvore representa um item frequente no conjunto de dados. Em seguida, o algoritmo fp-growth usa a árvore para encontrar todos os subconjuntos frequentes dos dados.

O FP-Growth é uma técnica eficiente em relação a outros algoritmos de mineração de dados frequentes, como o Apriori, pois requer menos passagens pelo conjunto de dados e usa menos espaço de memória. O algoritmo Apriori é bem conhecido para descoberta de conjuntos de itens frequentes, porém ele requer várias varreduras de um banco de dados, o que é ineficiente na era do Big Data de hoje. Ao contrário do algoritmo Apriori, o algoritmo FP-growth requer apenas duas varreduras do banco de dados e não gera conjuntos de itens candidatos. Portanto, o algoritmo FP-growth é considerado mais eficiente que o algoritmo Apriori (BAGUI; DEVULAPALLI; COFFEY, 2020).

Muitos estudos são conduzidos utilizando o algoritmo FP-Growth, a fim de encontrar padrões frequentes nas mais diversas áreas, alguns estudos serão abordados a seguir.

Wang e Cheng (2018) discutem a importância da auditoria de comportamento do usuário para garantir a segurança do sistema de gerenciamento de informações elétricas, propondo um algoritmo para detectar regras de associação entre o comportamento do usuário e problemas do sistema (alertas ou erros).

O método proposto classifica primeiro o registro do usuário de acordo com sua relação com os problemas do sistema e, em seguida, utiliza o FP-Growth para detectar padrões frequentes que ocorrem antes dos problemas do sistema. Os resultados dos experimentos indicam que o método proposto é eficaz na detecção de comportamentos regulares que podem causar problemas de segurança do sistema (WANG; CHENG, 2018).

Bagui et al. (2020) abordaram a mineração de regras de associação (ARM), mais comumente usada em bancos de dados transacionais, e é um processo de descoberta de relacionamentos entre itens ou conjuntos de itens. As regras de associação são formadas a partir de conjuntos de itens frequentes, sendo esse, portanto, o primeiro passo da ARM, a descoberta de conjuntos de itens frequentes (BAGUI; DEVULAPALLI; COFFEY, 2020).

Qi; Guo, (2022) estudaram o aumento da escala de redes móveis domésticas e a cobertura de estações base 5G em grandes cidades, municípios e vilas. No entanto, a comunicação de alta frequência da rede 5G NR aumentou o consumo de energia dos equipamentos em 3-5 vezes em comparação com o 4G.

O custo de energia para os operadores de comunicação está aumentando rapidamente, tornando-se imperativo estudar a economia de energia e a redução de consumo das estações base 5G (QI; GUO, 2022).

Para alcançar economia de energia nas estações base de comunicação, os estudiosos realizaram trabalhos aprofundados de pesquisa, sendo que os métodos de economia de energia das estações base podem ser divididos em três tipos: o primeiro é a atualização de hardware; o segundo é um modelo linear simples ou a experiência de especialistas, que depende da experiência do especialista para julgar e operar; o terceiro é algoritmos de aprendizado de máquina, como árvore de regressão de levantamento de gradiente, rede neural e assim por diante (QI; GUO, 2022).

A prática mostra que o algoritmo de aprendizado de máquina tem melhor efeito de economia de energia do que a experiência tradicional de especialistas, e essa informação tende a ser muito importante no que se refere a tomadas de decisões em relação a utilização de algoritmos de AM inalterados (QI; GUO, 2022).

Em resumo, o estudo acima apresenta uma técnica de análise de cluster envolvendo dados da estação base, que utiliza a técnica de FP-Tree de onde o algoritmo FP-Growth se baseia, para então descobrir regras de associação da carga de serviço da estação base 4G/5G do operador. O método sugerido no estudo utiliza segmentação vertical, ressegmentação, mineração de sub-banco de dados e integração de resultados para melhorar ainda mais o algoritmo. Em conclusão, o estudo apresenta uma melhoria do algoritmo FP-Growth para identificar cenas de economia de energia em estações base 4G/5G, a confiabilidade da regra de associação e a eficácia do algoritmo de identificação de campo de economia de energia, melhora significativamente o potencial de economia de energia do equipamento, mantendo os índices KPI e KQI da rede 4G/5G dos operadores inalterados (QI; GUO, 2022).

Porém o algoritmo FP-Growth também possui algumas limitações, alguns estudos são conduzidos a fim de sugerir melhorias em sua eficácia e eficiência. Li; et al., (2011) utilizaram as restrições do FP-Growth como forma de melhorar a eficácia e eficiência do processo. Entende-se que as restrições são os parâmetros utilizados no algoritmo FP-Growth, sendo que os parâmetros mais comuns utilizados para reduzir o número de regras de associação geradas são o *support* e *confidence*, pois altos percentuais de *support* e *confidence*, tendem a gerar menos regras de associação, devido às restrições impostas.

O método proposto por Li; et al., (2011), consiste em três fases: i) primeiro, são gerados os frequentes itemsets (são os dados em si utilizados no estudo, se fosse um estudo envolvendo fenômenos climáticos, seriam por exemplo, itemsets como (granizo, chuva), (tempestade, chuva), etc.; ii) segundo, explora-se as propriedades das restrições dadas para podar o espaço de busca ou salvar a verificação de restrição nas bases de dados condicionais; iii) terceiro, para cada itemset possível de satisfazer a restrição, gera-se sua base de dados condicional e realiza-se as três fases recursivamente na base de dados condicional. Os resultados experimentais mostram que o método proposto supera o algoritmo FP-growth revisado, como o FP-growth+. (LI; CAO; GUO, 2011).

Borgelt, (2005) propõe implementações em C do algoritmo FP-growth, com o objetivo de melhorar ainda mais sua performance, devido a linguagem C ser comumente mais performática e que outras linguagens como Python e Java. O algoritmo FP-Growth proposto em C inclui duas variantes da operação principal de

computar uma projeção de uma árvore FP-tree (a estrutura de dados fundamental do algoritmo FP-growth), e as árvores FP-tree projetadas são podadas opcionalmente, removendo itens que se tornaram infrequentes devido à projeção (uma abordagem chamada de FP-Bonsai) (BORGELT, 2005).

O autor apresenta resultados experimentais comparando esta implementação do algoritmo FP-growth com outros três algoritmos de mineração de conjuntos de itens frequentes (Apriori, Eclat e Relim) que também foram implementados. Ele ainda realizou experimentos em cinco conjuntos de dados, BMS-Webview-1, T10I4D100K, censo, xadrez e cogumelos, e foram comparados os resultados dos algoritmos Apriori, Eclat e Relim com o FP-growth, que obteve o melhor desempenho em todos os conjuntos de dados, exceto no conjunto artificial T10I4D100K, onde foi superado pelo Relim (BORGELT, 2005).

As técnicas convencionais de mineração de regras de associação, que consistem em encontrar relações entre diferentes itens, expressas como $A \Rightarrow B$, onde A e B são conjuntos de itens, tem como objetivo encontrar todas as regras de associação acima de um suporte mínimo e uma confiança mínima especificados pelo usuário, porém um Wang et al., (2002), propõe uma nova abordagem chamada TD-FP-Growth, que usa uma busca top-down no FP-tree em vez de uma busca bottom-up como a abordagem padrão. Essa nova abordagem pode reduzir o espaço de busca e melhorar a eficiência (WANG et al., 2002).

As técnicas de mineração de dados (DM), como *clustering*, classificação e associação de regras (ARM), são amplamente utilizadas para a extração de conhecimento dos dados em outros domínios, como a saúde. Koukaras; tjortjis; Rousidis, (2022), abordam o uso crescente das redes sociais por pessoas, organizações, empresas e governos, principalmente durante a pandemia da COVID-19, que gerou grande quantidade de dados online relacionados a diversos aspectos.

A análise de conteúdo derivado dessas plataformas, como o Twitter, é um desafio devido à grande quantidade de dados esparsos e ruidosos que precisam ser analisados para extração de conhecimento, o que reforça a utilização de ferramentas e técnicas para lidar com grandes volumes de dados (KOUKARAS; TJORTJIS; ROUSIDIS, 2022).

Uma metodologia foi proposta para o estudo acima, visando identificar tópicos de discussões relacionados à COVID-19 no Twitter, que combina a técnica de extração de tópicos, com ARM para mitigar problemas de extrações genéricas e

reduzir o número de tópicos redundantes. O objetivo foi encontrar as regras de palavras mais fortes em tweets usando métricas como *support, confidence, lift e leverage*, essas métricas fornecem informações importantes sobre a força e a relevância das regras de associação descobertas (KOUKARAS; TJORTJIS; ROUSIDIS, 2022).

O estudo conclui que a metodologia proposta pode gerar menos tópicos, com inferências mais fortes, que representam as atitudes públicas em relação à pandemia da COVID-19 expressas por postagens no Twitter. Isso pode levar a uma melhor compreensão dos tópicos de discussão e a melhores tomadas de decisão pelos formuladores de políticas e outras partes interessada (KOUKARAS; TJORTJIS; ROUSIDIS, 2022).

3 OBJETIVOS

3.1 Objetivo geral

Desenvolver um sistema de monitoramento e associação de ocorrências de granizo no Brasil com outros fenômenos climáticos extremos. Isso será alcançado através da criação de uma metodologia de coleta automatizada e centralização de dados, bem como do desenvolvimento de um modelo de regras de associação baseado em fenômenos climáticos extremos com o granizo. Para atingir esse objetivo geral, os objetivos específicos são:

3.2 Objetivos específicos

- Estruturar um algoritmo responsável pela captação dos dados através da internet para a região Sul do Brasil e para o estado de São Paulo;
- Analisar as associações do granizo com outros fenômenos climáticos a partir dos padrões frequentes dos conjuntos de dados.
- Elaborar uma dashboard de monitoramento e disponibilizar as tabelas e resultados encontrados.

Observa-se que, para atingir os objetivos específicos, foi necessário o desenvolvimento de dois algoritmos escritos em linguagem Python, sendo eles: um para captação de dados; e um para processamento e decodificação dos dados obtidos. Na criação dos modelos de associação foi utilizado o algoritmo de fp-growth disponível publicamente.

4 METODOLOGIA

A metodologia proposta neste trabalho envolve três etapas fundamentais. A primeira etapa consiste no desenvolvimento do sistema de coleta e armazenamento. Para coleta de dados, foram utilizadas técnicas de WS (KHDER, 2021) baseados em três plataformas de dados selecionadas, sendo elas:

- Rede de Meteorologia do Comando da Aeronáutica (REDEMET): que tem como objetivo disponibilizar e integrar produtos meteorológicos voltados para aviação civil e militar, o acesso aos dados meteorológicos do padrão *Meteorological Aerodrome Report* (METAR), pode ser feito utilizando a própria API da REDEMET, previamente fazendo um cadastro e obtendo um token de acesso.

- Instituto de Pesquisas Meteorológicas (IPMET/Unesp): é um instituto de pesquisas vinculado a UNESP do município de Bauru, estado de São Paulo, e realiza o monitoramento de dados em estações meteorológicas.

- Sistema Integrado de Informações sobre Desastres (S2ID): é uma plataforma mantida pela defesa civil do Brasil/Governo Federal, que faz o registro de dados e prejuízos relacionados a fenômenos climáticos, com reconhecimento federal deles.

A seleção destas plataformas, consistiu no fato de possuírem dados estruturados de eventos atmosféricos extremos, que são constantemente atualizados para a região de estudo.

Posteriormente, após a coleta dos dados, os mesmos, foram processados e passaram por um procedimento de limpeza, e centralização. A centralização dos dados, é o processo de unificar as três plataformas de dados em uma única base de dados central, que tem como foco eventos de granizo, porém outros tipos de fenômenos climáticos também foram coletados.

O local de centralização, se trata de uma ‘coleção de dados’. O nome ‘coleção’ é uma nomenclatura padrão, muito utilizada para se referir aos conjuntos de dados de bancos não relacionais, também conhecidos como *Not Only SQL* (NoSQL). Assim sendo, para esse estudo, sempre que o termo “coleção” ou “coleções” forem utilizados, entende-se como os conjuntos de dados que foram captados, por exemplo dados da plataforma REDEMET, se torna coleção da REDEMET.

A segunda etapa consiste na preparação dos dados, a fim de utilizá-los com o algoritmo *FP-growth*, que realiza a associação entre os eventos atmosféricos extremos. Após a execução do algoritmo, foi gerado uma tabela de regras de associação e com base nessa tabela, foram realizadas as análises de associações em relação ao granizo, também foi feita uma análise individual das coleções de dados obtidos de cada uma das três plataformas selecionadas.

Com a tabela de associações, gráficos gerados e dados centralizados, iniciou-se a terceira etapa, que consistiu no desenvolvimento da dashboard de monitoramento.

4.1 Desenvolvimento do sistema de captura e centralização dos dados

4.1.1 Preparação do ambiente

4.1.1.1 Preparação do Linux e Linguagem de programação

O sistema operacional utilizado como base para implementação e desenvolvimento do sistema de coleta e dashboard de monitoramento, foi o Linux distribuição Ubuntu versão 21.0, e a linguagem utilizada foi a versão 3.7 do Python.

A *Integrated Development Environment* (IDE) utilizada para desenvolvimento do código foi o Visual Studio Code da Microsoft, versão 1.58.2.

A utilização do Python como linguagem principal se deve em razão de algumas importantes características, como o rápido desenvolvimento, tipagem de variáveis simplificada, sintaxe do código simples, e o mais importante, por possuir várias bibliotecas externas voltadas para ciência de dados e coleta automatizada.

Quadro 1 apresenta as bibliotecas externas que foram utilizadas no presente estudo.

Quadro 1 - Bibliotecas utilizadas

Biblioteca	Finalidade
Pandas	Visualização dos Dados obtidos
Mlxtend	Contém algoritmo fp-growth
Plotly	Visualização do Treemap
Numpy	Manipulação dos arrays de dados
Requests	Requisição na API (*Application Programming Interface*) da REDEMET
Selenium	Obtenção automatizada dos dados

Fonte: Autoria própria (2023)

- Pandas: Utilizada para manipulação de dados, pois permite uma variedade de operações envolvendo Data Frames, que são estruturas bidimensionais na forma de tabelas, e geralmente são compostos por duas ou mais series. As series, por sua vez, são matrizes unidimensionais rotuladas, capazes de armazenar dados de qualquer tipo como inteiro, string, float, objetos python, entre outros tipos, e são acessadas através dos rótulos de seus eixos, também chamados de índices.

- Mlxtend (machine learning extensions): Uma biblioteca de funções úteis para realizar tarefas que envolvam exploração de dados, muito utilizada em ciência de dados. Alguns dos algoritmos presentes na Mlxtend são relacionados a aprendizagem de máquina, como regressão linear, random forest etc. Para o presente estudo, apenas foi utilizado o algoritmo de associação FP-Growth.

- Plotly: É uma biblioteca *open source*, utilizada para geração de gráficos, por exemplo, gráfico de linha, bolhas, dispersão, box plots, histogramas etc. Nesse estudo foi utilizado para gerar gráficos de barras e *tree maps*.

- Numpy (*Numerical Python*): Numpy é utilizada para diversos fins, sendo muito empregada em ciência de dados e é geralmente em conjunto com a biblioteca Pandas. O Numpy facilita a realização de cálculos com *Arrays* multidimensionais (Matrizes), pois a manipulação de matrizes se torna muito mais simples, sendo possível uma variedade de cálculos matemáticos envolvendo duas ou mais matrizes ou *arrays* unidimensionais. Para o processamento dos diversos dados coletados nesse estudo, utilizou-se o Numpy, para percorrer as Series de dados em do Pandas, de forma mais performática, devido ao alto volume de dados.

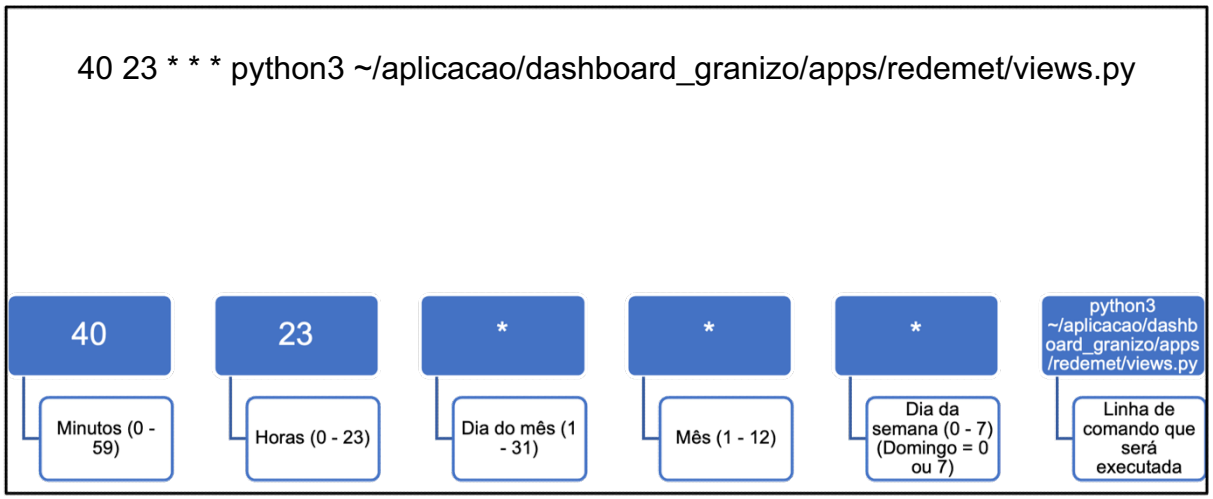
- Requests: Biblioteca utilizada para realizar operações de requisições de GET, POST, DELETE, OPTIONS, PUT, para endereços de URL (*Uniform Resource Locator*). De forma resumida, uma requisição é o ato de enviar uma solicitação através de um endereço de internet, geralmente contendo um servidor em uma ponta que envia a requisição e outro em outra ponta que recebe, processa e devolve o resultado do processamento para o servidor inicial.

- Selenium: A utilização da biblioteca Selenium é muito comum em projetos que envolvam automação de testes, principalmente em ambiente de desenvolvimento comercial, que já estejam em produção. Porém, em cenários que não é possível

coletar dados por uma API, utiliza-se a biblioteca Selenium para realizar a automação para preenchimento de formulários e clicks em botões, simulando a operação feita por um ser humano.

Outra configuração muito importante do ambiente é a utilização do CRON, um agendador de tarefas em sistemas Linux, que foi utilizado para automatizar o processo de coleta de dados nas plataformas REDEMET, IPMET e S2ID. O CRON é configurável, e com ele é possível definir um horário de execução de tarefas, uma tarefa pode ser uma operação utilizando o terminal de linhas de comando, para executar alguma ação, por exemplo, iniciar um programa, ou script. Porém para que a configuração e execução das tarefas ocorra de forma correta, é necessário seguir um padrão de sintaxe, a Figura 7 demonstra o padrão utilizado para configuração e funcionamento do CRON.

Figura 7 - Sintaxe padrão de configuração do CRON



Fonte: Autoria própria (2023)

Em resumo é possível criar rotinas em horários específicos para que o CRON as execute automaticamente, por exemplo, na Figura 7 o CRON é configurado para executar um código fonte Python todo dia as 23:40.

Abaixo segue um modelo de escrita do comando para o CRON, onde a cada 2 minutos, um código fonte Python é iniciado.

- ***/2 * * * python3 ~/aplicacao/dashboard_granizo/apps/metar/views.py**

Essa configuração executa o arquivo views.py da pasta metar. No arquivo views.py está contido o código que decodifica as mensagens no padrão METAR, e

desse ponto em diante será chamado de algoritmo de decodificação.

Devido ao alto custo de processamento necessário para realizar a decodificação, e para evitar travamentos do algoritmo, são processados 1000 registros a cada intervalo de duas horas.

Ressalta-se que o algoritmo de decodificação, trabalha especificamente em conjunto apenas com a plataforma da REDEMET, as plataformas da S2ID e IPMET/Unesp, não necessitam de decodificação, pois seus dados coletados não são inicialmente no padrão METAR.

No Apêndice A encontram-se detalhes da execução das rotinas em CRON das plataformas restantes.

4.1.1.2 Sistema Gerenciador de Banco de Dados (SGBD)

O SGBD tem a função de gerenciar e armazenar toda base de dados coletada e, para esse estudo, foi implantado o PostgreSQL versão 13.4 e o MongoDB 6.0.0.

O Postgres se trata de um banco de dados objeto relacional e foi utilizado apenas para controle de logins e autenticação de usuários.

Em relação ao MongoDB, ele é um banco de dados NoSQL, também conhecido como banco de dados não relacional, e foi utilizado para armazenar as coleções de dados que foram obtidos das plataformas. A escolha do MongoDB para armazenamento dos dados, se deu em função de sua melhor performance para lidar com grandes volumes de dados que não precisam se comunicar entre si.

No MongoDB foi criada uma base de dados principal chamada de ‘granizo’, que possui internamente 10 coleções de dados:

- aerodromos
- municípios
- datas_redemet
- datas_s2id
- datas_ipmet
- redemet
- s2id
- ipmet
- temp
- center_data

4.1.1.3 Instalação dos frameworks

Um framework pode ser entendido como um conjunto de códigos já prontos e testados, a fim de promover agilidade e produtividade no desenvolvimento de softwares.

Em relação a interface gráfica, a dashboard de monitoramento proposta como objetivo nesse estudo, foi desenvolvida utilizando o framework Django em sua versão 3.8 e framework Bootstrap versão 5.1.0, a fim de ternar a dashboard responsiva para diversos tamanhos telas de dispositivos distintos, como celulares, tabletes, monitores para desktop. O Django é um framework escrito em linguagem Python destinado a desenvolvimento de aplicações web, possui diversas ferramentas e funcionalidades que facilitam e aceleram a produtividade, e o Bootstrap é um framework web para desenvolvimento de componentes de interface e front-end voltados para sites e aplicações web em páginas HTML, que utilizam *Cascading Style Sheets* (CSS) e JavaScript.

Além de controlar e direcionar o fluxo de URLs que chegam ao servidor, o Django é o responsável por receber as solicitações de carregamento de uma página *HyperText Markup Language* (HTML) e retornar a página para ser exibida ao usuário.

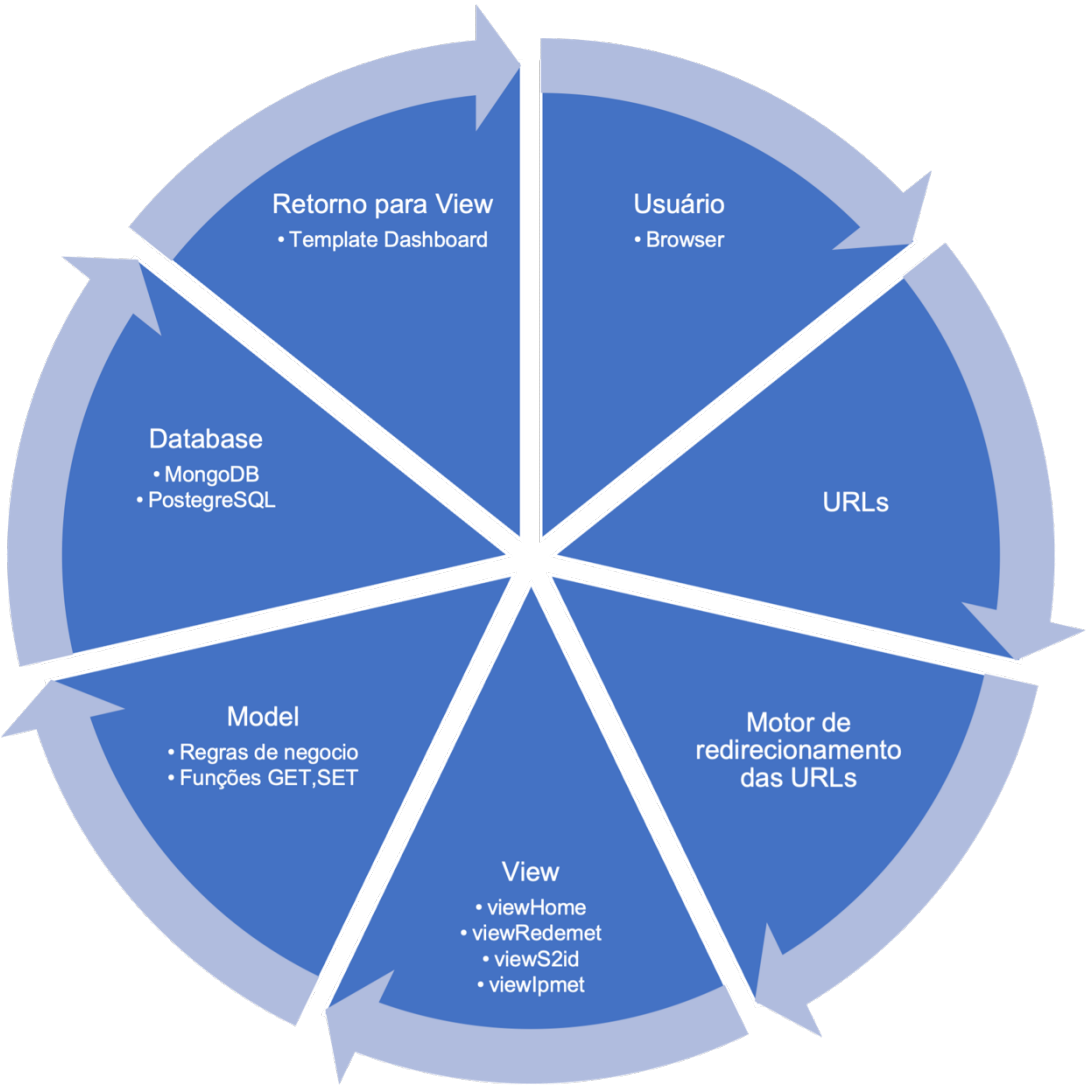
Uma das características principais do Django, é que sua arquitetura segue o padrão Model View Template (MVT) sendo essa uma tipologia dividida em camadas. Iniciando pela camada da Model, nela são escritas as classes, e as respectivas funções de cada classe, sendo que cada classe é um objeto que será posteriormente transformado em tabela relacional, e possui funções como GET, SET, entre outras, para manipular os dados de cada objeto.

A View serve como um controlador, é nessa camada que são organizadas as funções que solicitam dados para a camada das Models e essa por sua vez, fazem a ligação direta com o banco de dados, posteriormente a camada de Model retorna os dados solicitados para a camada da View, que renderiza um template HTML e finalmente exibi ao usuário.

A camada de Template, resumidamente é uma página HTML, geralmente é a última camada antes de chegar ao usuário.

Dessa forma, no estudo proposto, o funcionamento do Django e sua arquitetura, se resumem ao apresentado na Figura 8**Erro! Fonte de referência não encontrada.**.

Figura 8 - Ciclo MVT do Django



Fonte: Autoria própria (2023)

4.2 Métodos de coletas, processamento e centralização dos dados

O presente estudo utilizou três plataformas distintas para obtenção de dados, sendo que cada plataforma possui suas particularidades em relação ao método de coleta e processamento dos dados obtidos.

4.2.1 Construção do *web scraping* e plataformas utilizadas como fonte de dados

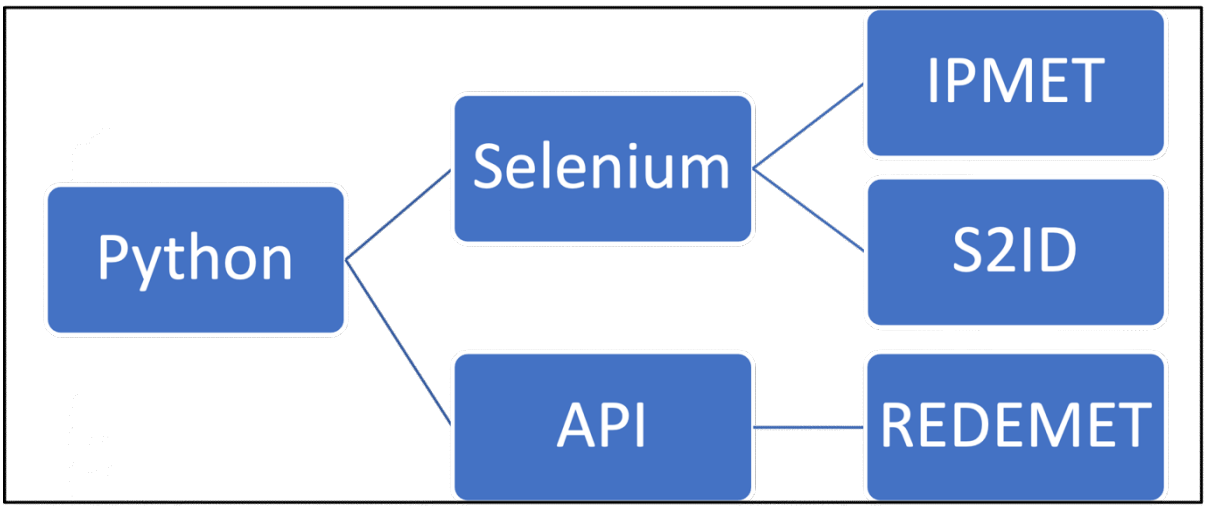
Dados a dificuldade em se obter dados observacionais de granizo, a coleta de informações relacionadas esse evento é um dos objetivos desse estudo.

Sendo assim, foi desenvolvido um algoritmo de WS focado em coleta de dados automatizada através de *internet*, utilizando duas estratégias distintas devido a necessidade de cada plataforma.

Na REDEMET, foram coletados dados no padrão METAR, necessitando de um algoritmo que executasse de forma cíclica e contínua, as requisições pela própria Application Programming Interface (API) da REDEMET.

Em relação as plataformas da IPMET/Unesp e S2ID, tais plataformas não possuem API própria. Então para se obter os dados, é preciso manualmente preencher um formulário, com datas e mais algumas informações em suas respectivas plataformas. Então foi necessário o desenvolvimento de um algoritmo de acesso direto a página HTML, simulando o preenchimento do formulário para obtenção de dados, esse processo é feito de forma autônoma por um algoritmo que foi programado para interagir com a página HTML e simular o comportamento de uma pessoa, pode-se chamar esse comportamento automático de bot. Para programar o bot, foi utilizada a biblioteca Python Selenium. Em resumo, os algoritmos de WS utilizados nas plataformas se dividem em dois tipos. A Figura 9, mostra de forma resumida cada estratégia de coleta de dados.

Figura 9 - Estratégia de coleta automatizada de cada plataforma



Fonte: Autoria própria (2023)

4.2.2 Coleção Municípios

A coleção de Municípios possui dados relacionados aos município da região Sul e estado de São Paulo do Brasil, e tem por finalidade servir como complemento de informações para as bases de dados do S2ID e IPMET. O Quadro 2 apresenta os dados contidos em cada objeto da coleção de Municípios.

Quadro 2 - Dados da coleção MUNICÍPIOS

Nome do dado	Tipo do dado	Descrição do dado
Id	ObjectId	Idenficador único do registro faz referência ao nome do município
Cod uf	String	Faz referência ao código do estado
Uf	String	Faz referência ao nome do estado
Sgl uf	String	Faz referência a sigla utilizada como padrão para o estado
Cod ibge	String	Faz referência ao código do IBGE do município, esse código é padrão
Nome município	String	Faz referência ao nome do município, é o mesmo dado do ID, porém a localização pelo ID se torna mais rápida pelo indexador do banco de dados
Cod_latitude	String	Latitude do município

Cod_longitude	String	Longitude do município
Sgl regiao	String	Sigla da região de localização do município, ex: N, S, L, O
Região	String	A descrição completa da região do município, ex: Norte, Sul, Leste, Oeste

Fonte: Autoria própria (2023)

Já a base REDEMET não teve seus dados agregados com dados dos munícipios, pois o este já utiliza os dados dos aeródromos como dados de localização.

4.2.3 Coleção Aeródromos

Os aeródromos (aeroportos grandes ou pequenos) coletam dados meteorológicos e os salvam no padrão METAR o Quadro 3 apresenta um exemplo:

Quadro 3 - Exemplo de mensagem no padrão METAR

METAR	SBGR 121500Z 15010KT 8000 RA SCT020 BKN040 OVC080 23/19 Q1015 TSRA
SBGR	Identificador ICAO do aeródromo (Aeroporto Internacional de Guarulhos, em São Paulo, Brasil)
121500Z	Data e hora em UTC (12º dia do mês, às 15 UTC)
15010KT	Vento de 150 graus a 10 nós
8000	Visibilidade de 8000 metros (boa visibilidade)
RA	Chuva (rain)
SCT020	Nuvens esparsas a 2000 pés (pés acima do nível do solo)
BKN040	Nuvens fragmentadas a 4000 pés
OVC080	Nuvens encobertas a 8000 pés
23/19	Temperatura de 23 graus Celsius, ponto de orvalho de 19 graus Celsius
Q1015	Pressão atmosférica de 1015 hPa
TSRA	Trovoada com chuva (thunderstorm with rain)

Fonte: Autoria própria (2023)

Este exemplo indica a presença de chuva e trovoadas no local durante a observação METAR.

Tais dados consistem em mensagens codificadas no formato de texto. Existem Aeródromos espalhados por todo o Brasil que observam e registram em um banco de dados as mensagens METAR de hora em hora.

Através da API da REDEMET, é possível coletar os dados registrados pelos aeródromos, sendo necessário informar o código do aeródromo que se deseja requisitar os dados.

A coleção chamada Aeródromos da base de dados Granizo, contém informações como o código e estado de cada aeródromo registrado no Brasil. Estes dados foram de estado, foram utilizados para filtrar apenas os aeródromos da região Sul e estado de São Paulo. Os dados contidos na coleção Aeródromos estão relacionados no Quadro 3.

Quadro 4 - Dados da coleção AERODROMOS

Nome do dado	Tipo do dado	Descrição do dado
Id	ObjectId	Idenficador único do registro faz referência a sigla que identifica o aeródromo, ex: “SBAE”
Id_ext	Int32	Id externo que referência também o aeródromo, porém utilizando uma nomenclatura de inteiros
Nome	String	Nome do aeródromo responsável pelo registro de observação, ex: “Aeroporto Internacional Bauru-Arealva / Moussa Nak”
Cidade	String	Cidade do aeródromo
Uf	String	Estado do aeródromo
País	String	País do aeródromo
Latitude	String	Latitude do aeródromo
Longitude	String	Longitude do aeródromo
Altitude metros	Int32	Altitude do aeródromo

Fonte: Autoria própria (2023)

Destaca-se que os registros de todos os aeródromos armazenados na coleção ‘Aerodromos’, podem ser visualizados no Apêndice B.

4.2.4 Coleções de datas

As coleções de datas são muito importantes, pois é possível associar a utilização de HASHs (algoritmo matemático para criptografia) como técnica para controle de duplicidade dos registros coletados. Assim, a data correspondente foi uma das variáveis utilizadas para gerar o HASH de cada registro.

Para cada plataforma selecionada, foi criado uma coleção correspondente de datas, por exemplo, DATAS_REDEMET, DATAS_S2ID e DATAS_IPMET. Cada coleção armazena como dado apenas a última data que foi executada uma coleta de dados pelo algoritmo de coleta em Python.

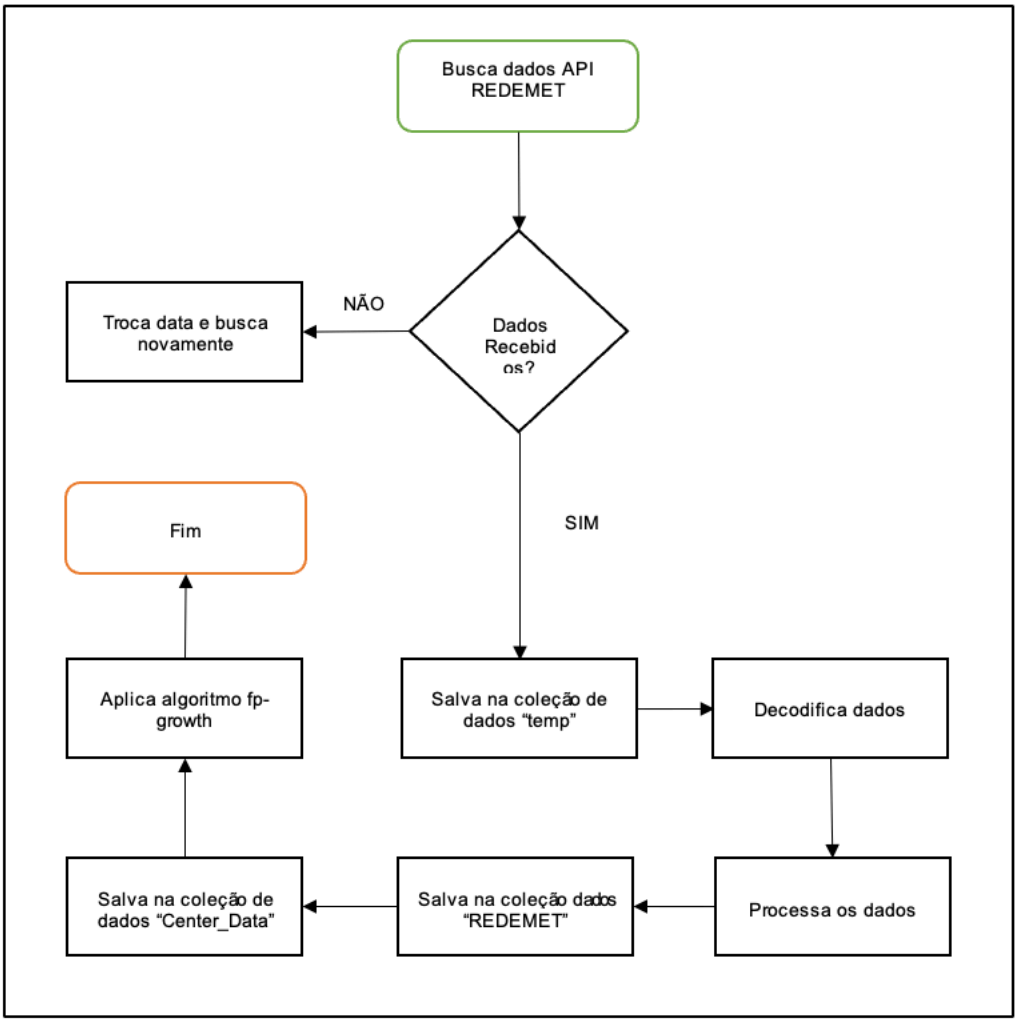
Dessa forma, garante-se que na próxima vez que o CRON (agendador de tarefas) realizar uma execução no algoritmo de coleta de dados, os dados coletados, não serão de datas repetidas.

A estrutura básica das coleções de datas, é apenas uma variável chamada “_id” que armazena a data já coletada. Esse dado possui valor numérico, pois as datas ficam organizadas na ordem decrescente, ou seja, a última data fica no topo da coleção. Quando um algoritmo de coleta inicia sua execução, ele busca qual foi a última data coletada e não a repete, dando sequência as próximas datas.

4.2.5 Metodologia da base REDEMET

O método utilizado para coleta e armazenamento dos dados na REDEMET é apresentado na Figura 9, que engloba desde a captação dos dados até o processamento e criação do modelo de regra de associação.

Figura 9 - Fluxograma de coleta, processamento e criação do modelo de associação da base de dados REDEMET



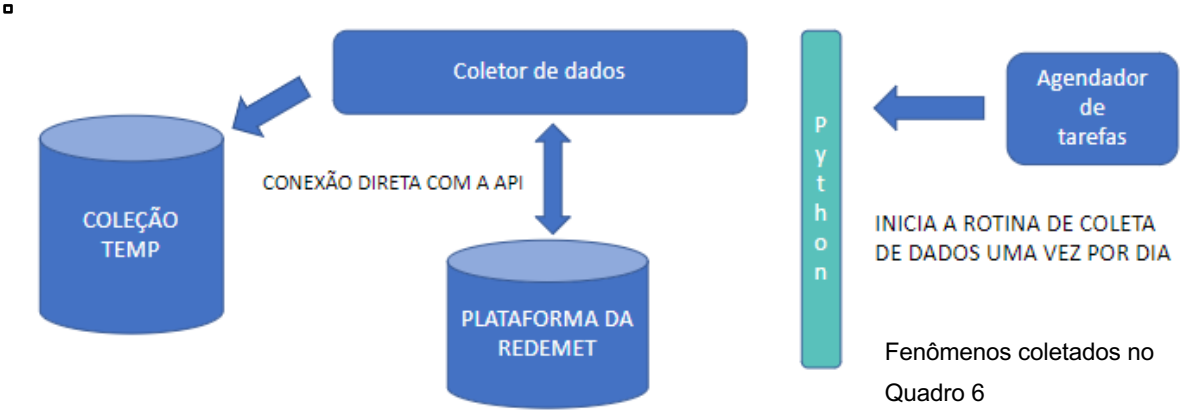
Fonte: Autoria própria

O fluxo de coleta inicia-se de forma automatizada, com o agendador de tarefas do Linux o CRON executando o algoritmo de coleta de dados.

Inicialmente o Linux teve seu relógio ajustado para o padrão universal time Coordinated (UTC), e todos os dias as 23 horas e 40 minutos, o agendador de tarefas iniciava a coleta. Para coletar os dados referentes ao dia em questão, e em caso de falha, o algoritmo foi programado para realizar até três tentativas antes de pular para o próximo dia.

O processo de coleta utilizou a própria API da REDEMET para obtenção dos dados, através de requisições diretas, a Figura 10 demonstra o processo de obtenção dos dados.

Figura 10 - Fluxo de coleta automatizada dos dados REDEMET



Fonte: Autoria própria (2023)

Com o algoritmo de coleta iniciado, os dados da REDEMET são coletados via API e salvos na coleção TEMP do MongoDB. A requisição é executada utilizado um GET na URL da API, sendo obrigatório informar quatro parâmetros:

• A sigla do aeródromo que se deseja obter os dados

• Data inicial e final que devem seguir um padrão (ano/mês/dia);

• Api token que é obtido no site da REDEMET, que é adquirido após preencher um pequeno cadastro;

• Tamanho do arquivo de paginação para determinar a quantidade de registros por página, pois há API da REDEMET limita o tráfego de dados.

Um modelo da URL utiliza para requisição de dados na REDEMET pode ser visualizado na Figura 11:

Figura 11 - Exemplo de requisição de dados da API do REDEMET

```
"https://api-redemet.decea.mil.br/mensagens/metar/{0}?api_key="+TOKEN_REDE-
MET+"&data_ini={1}&data_fim={2}&page_tam=200")
```

Fonte: Autoria própria (2023)

Para realizar a coleta de dados apenas das regiões Sul e estado de São Paulo, buscou-se na coleção Aerodromos, os aeródromos que fazem parte do universo de estudo, filtrando-os pelos estados em questão (SP, PR, SC e RS).

A API da REDEMET retorna até 200 mensagens METAR por requisição, porém, muitos dias os dados disponíveis para coleta, ultrapassam de 200 registros

por requisição. Então foram necessárias várias requisições para o mesmo dia, a fim de que todos os dados da data em questão fossem coletados.

O controle de paginação, utilizou uma variável de retorno da API da REDEMET, o *'next_page_url'* que indicava se ainda existiam dados a serem coletados para o dia em questão. A paginação indica a quantidade existente de páginas contendo 200 dados. Após o retorno das requisições, os dados eram salvos na coleção TEMP no MongoDB e inicialmente estavam no padrão METAR.

Quadro 5 apresenta a estrutura de dados da coleção TEMP.

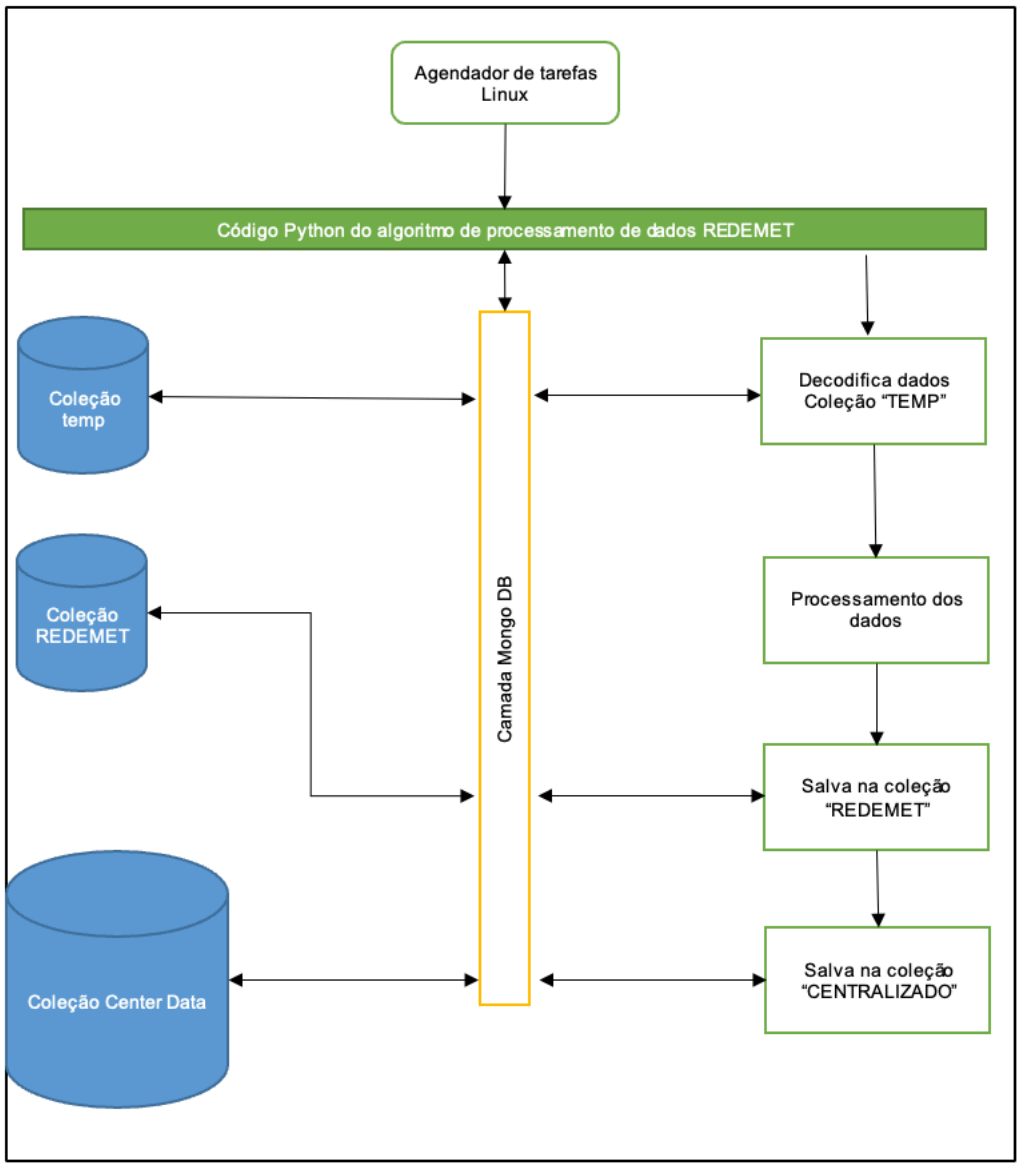
Quadro 5 - Dados da coleção TEMP

Nome do dado	Tipo do dado	Descrição do dado
Id	ObjectId	Idenficador único do registro
Fonte	Int32	Referência numérica da plataforma de coleta ex: 1 = Redemet, 2 = S2ID, 3 = IPMET
Localidade	String	Código do aeródromo, ex: "SBFI"
Recebimento	String	Data do recebimento no formato "YYYY-MM-dd Horas"
Metar	String	Mensagem padrão METAR
Is_visualizado	Boolean	Indica se o dado já foi processado pelos algoritmos de processamento de dados, a fim de evitar duplicidade de processamento

Fonte: Autoria própria (2023)

Após a coleta dos dados, inicia-se a etapa de processamento, que envolve a decodificação dos dados no padrão METAR, criação do HASH de controle, limpeza, padronização de textos e centralização dos dados na coleção CENTER DATA, que é a coleção de dados que será utilizada para centralizar os dados de todas as plataformas. A Figura 12 exibe um fluxo do processo realizado na etapa de processamento de dados.

Figura 12 - Método de processamento dos dados REDEMET



Fonte: Autoria própria (2023)

A primeira parte do fluxo de processamento dos dados é a execução automática pelo agendador de tarefas o CRON, pois é ele que também inicia o algoritmo de decodificação dos dados METAR. Ressalta-se que o CRON inicia dois tipos de algoritmos, um de coleta, como já citado anteriormente e um de decodificação, porém em horários distintos para balancear o uso dos recursos de hardware.

Os dados que estão na coleção TEMP, são divididos em três partes, a sigla do aeródromo, data e hora da coleta e a mensagem no padrão METAR. As mensagens do METAR nem sempre possuem o mesmo comprimento de caracteres. Então, o algoritmo de decodificação percorre toda a cadeia de caracteres buscando pelos fenômenos climáticos presentes na mensagem METAR que inicialmente estão codificados, e são descritos por siglas, como demonstrado no Quadro 6.

Quadro 6 - Siglas para decodificação da mensagem METAR

Sigla referente ao fenômeno e característica	Descrição
MI	baixo
PR	parcial
BC	banco
DR	flutuante
BL	soprada
SH	pancadas
TS	trovoada, raios e relâmpagos
FZ	congelante
DZ	chuvisco
RA	chuva
SN	neve
SG	grãos de neve
IC	cristais de gelo
PL	pelotas de gelo
GR	granizo
GS	pelotas de neve
+FC	tornado

Fonte: Autoria própria (2023)

O algoritmo termina sua execução quando todos os registros contidos na coleção “TEMP" forem percorridos, decodificados e salvos na coleção "REDEMET.

Os registros decodificados e salvos na coleção REDEMET seguem a estrutura apresentada no

Quadro 7.

Quadro 7 - Dados da coleção REDEMET

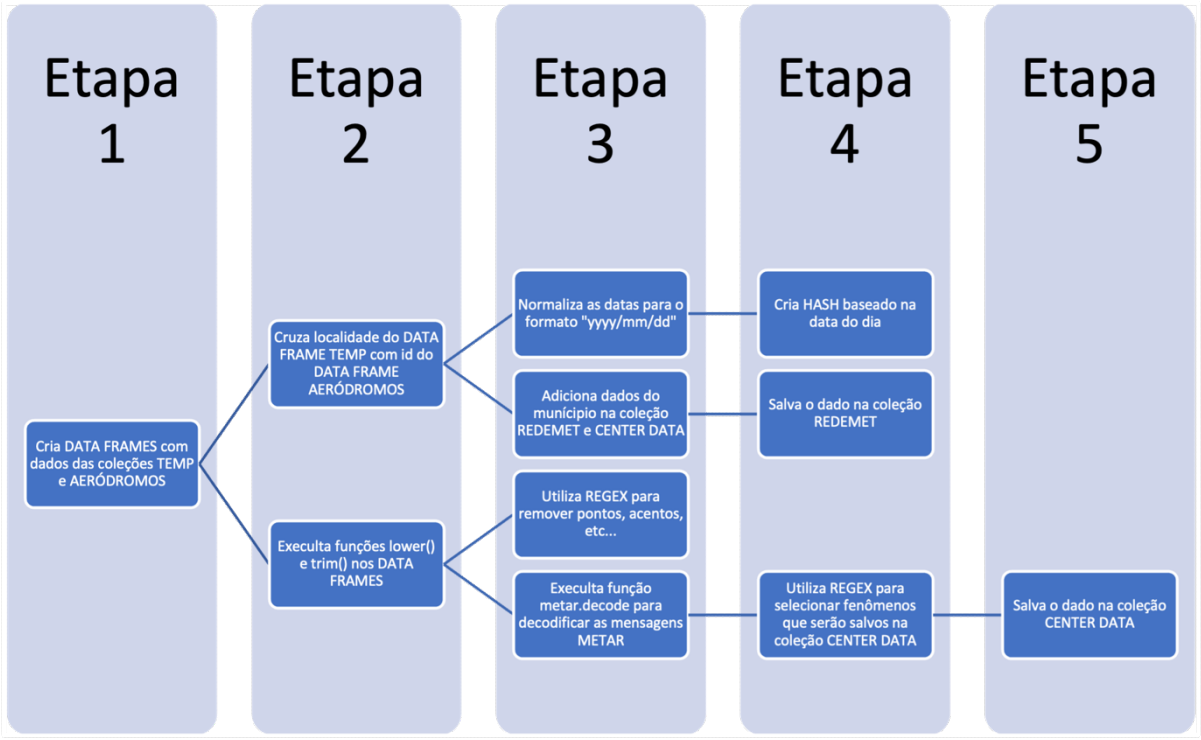
Nome do dado	Tipo do dado	Descrição do dado
Id	String	Identificador único do registro [hash gerado a partir dos dados (Figura 19)]
Data	String	Data do recebimento no formato “YYYY-MM-dd”
Aeroporto	String	Nome do aeródromo que coletou o dado METAR
Cidade	String	Cidade do aeródromo

Uf	String	Estado do aeródromo
País	String	País do aeródromo
Latitude	String	Latitude do aeródromo
Longitude	String	Longitude do aeródromo
Altitude	Int32	Altitude do aeródromo
Direcao_vento	String	Direção do vento em graus
Velocidade_vento	String	Velocidade do vento em knots
Rajada_vento	String	Rajada do vento em knots
Temperatura	String	Temperatura em grãos celsius
Fenomeno_legenda	Array	Lista de fenômenos em ordem de ocorrência ex: chuva, tempestade, granizo
Fenomeno_numero	Array	Lista com números que fazem referência aos fenômenos, ex: chuva = 1, tempestade = 2

Fonte: Autoria própria (2023)

No presente estudo, não foram coletados dados repetidos do mesmo dia, ou seja, um aeródromo que realiza observações de hora em hora pode conter a mesma observação ao longo do dia. Então, no dia em questão será salvo apenas um registro de tal fenômeno na coleção REDEMET. Esta limitação se faz necessária para balanceamento da coleção REDEMET. Sem essa limitação, as associações entre os eventos meteorológicos podem causar repetições e erros nos resultados obtidos nas associações. Por exemplo, seriam relacionados apenas os eventos mais comuns, como chuvas, chuvisco ou tempestades. Outros eventos considerados mais raros, como o granizo, quase não teriam relevância nas regras de associação. A Figura 13 demonstra por etapas um resumo do fluxo de processamento dos dados.

Figura 13 - Fluxo da etapa processamento de dados REDEMET



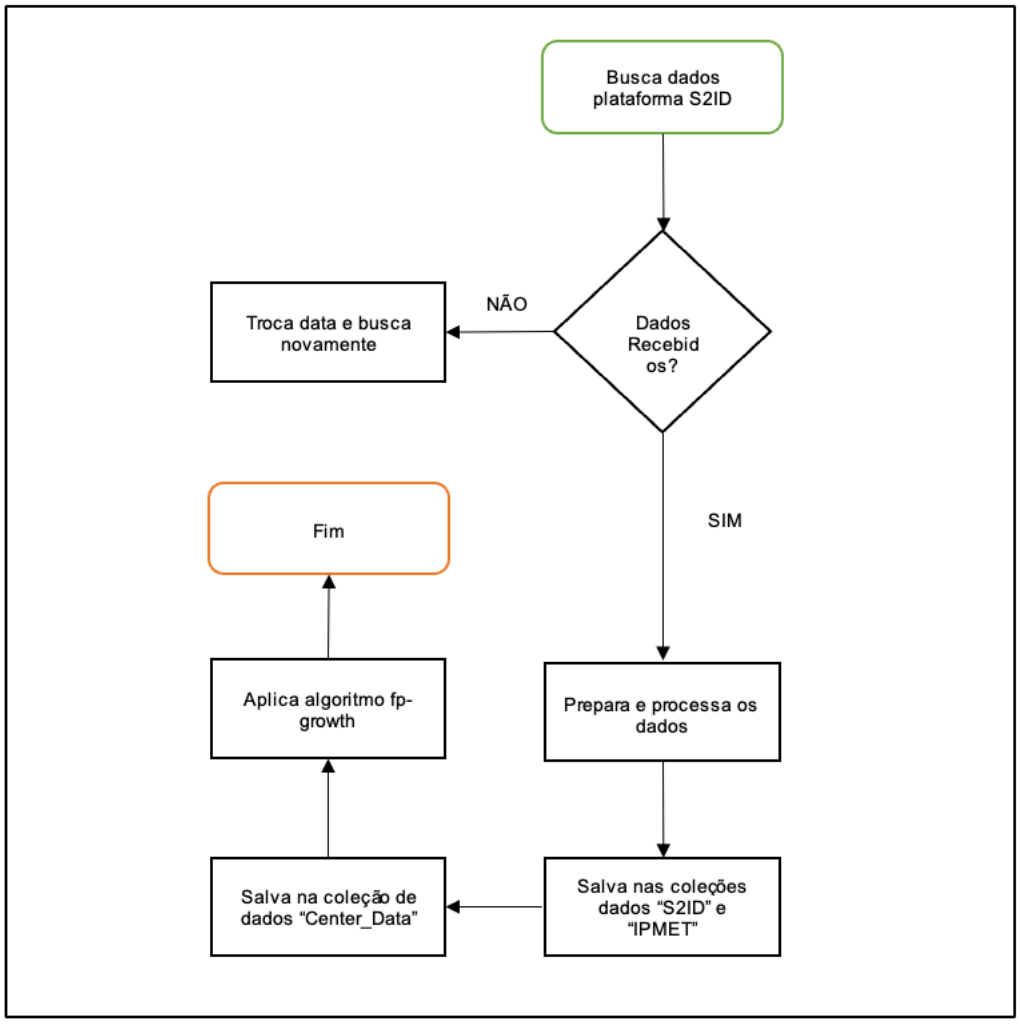
Fonte: Autoria própria (2023)

A etapa final é a centralização dos dados na coleção CENTER DATA que armazena os dados das três plataformas. Porém, antes de serem salvos na coleção CENTER DATA, eles passam por um filtro REGEX para selecionar apenas os fenômenos meteorológicos mais relevantes. A relação dos fenômenos pode ser observada no Quadro 12.

4.2.6 Metodologia das bases IPMET e S2ID

A Figura 14 apresenta o fluxograma de coleta, processamento e criação do modelo de associação da plataforma S2ID.

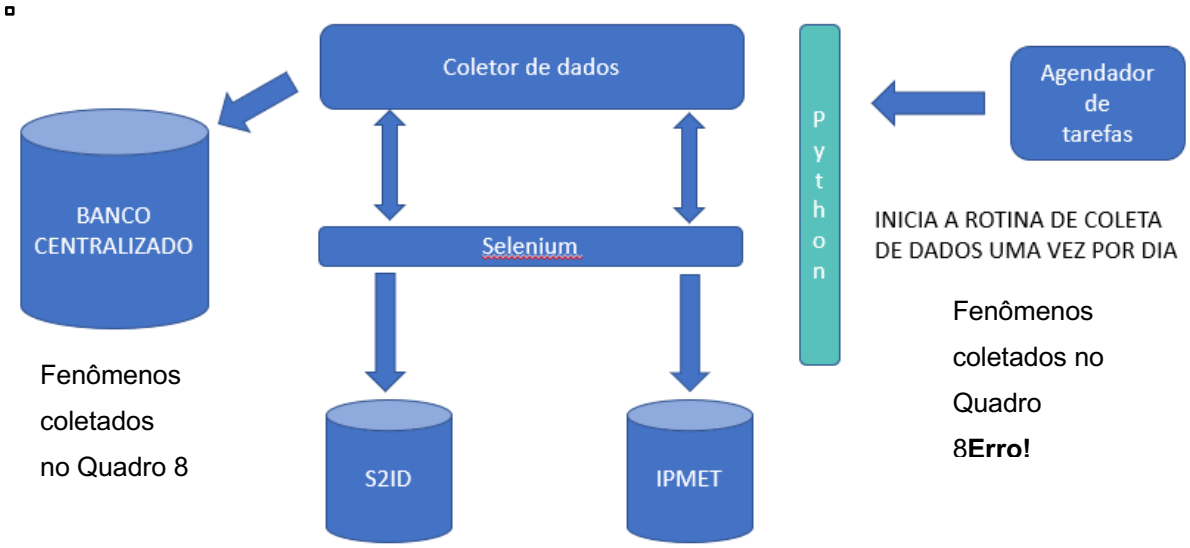
Figura 14 - Fluxograma de coleta, processamento e criação do modelo de associação da base de dados S2ID



Fonte: Autoria própria (2023)

O processo de coleta na plataforma S2ID, utiliza a biblioteca Selenium, que é uma ferramenta muito utilizada para automação de testes em aplicações web, mas também pode ser utilizada como coletora de dados em páginas da *web*. Para coletar dados utilizando o Selenium, é necessário identificar os principais elementos da página *web* de onde será realizada a coleta. É importante ressaltar que a coleta de dados de uma página da web pode estar sujeita a limitações legais e éticas, como a proteção de dados pessoais e a violação de termos de serviço de sites e serviços online. Na Figura 15, é apresentado o fluxo de coleta de dados das plataformas S2ID e IPMET/Unesp. Em geral, o método de coleta de ambas as plataformas é praticamente igual, mudando apenas no bloco "coletor de dados", pois nesse bloco, é onde são realizadas a programação específica de cada plataforma em relação aos elementos da página *web* de cada uma.

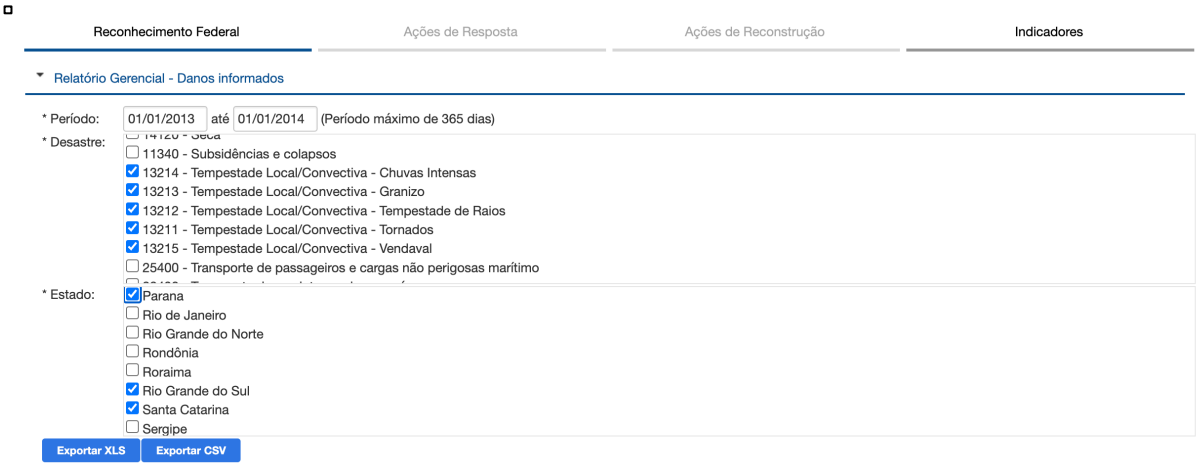
Figura 15 - Fluxo de coleta automatizada S2ID e IPMET



Fonte: Autoria própria (2023)

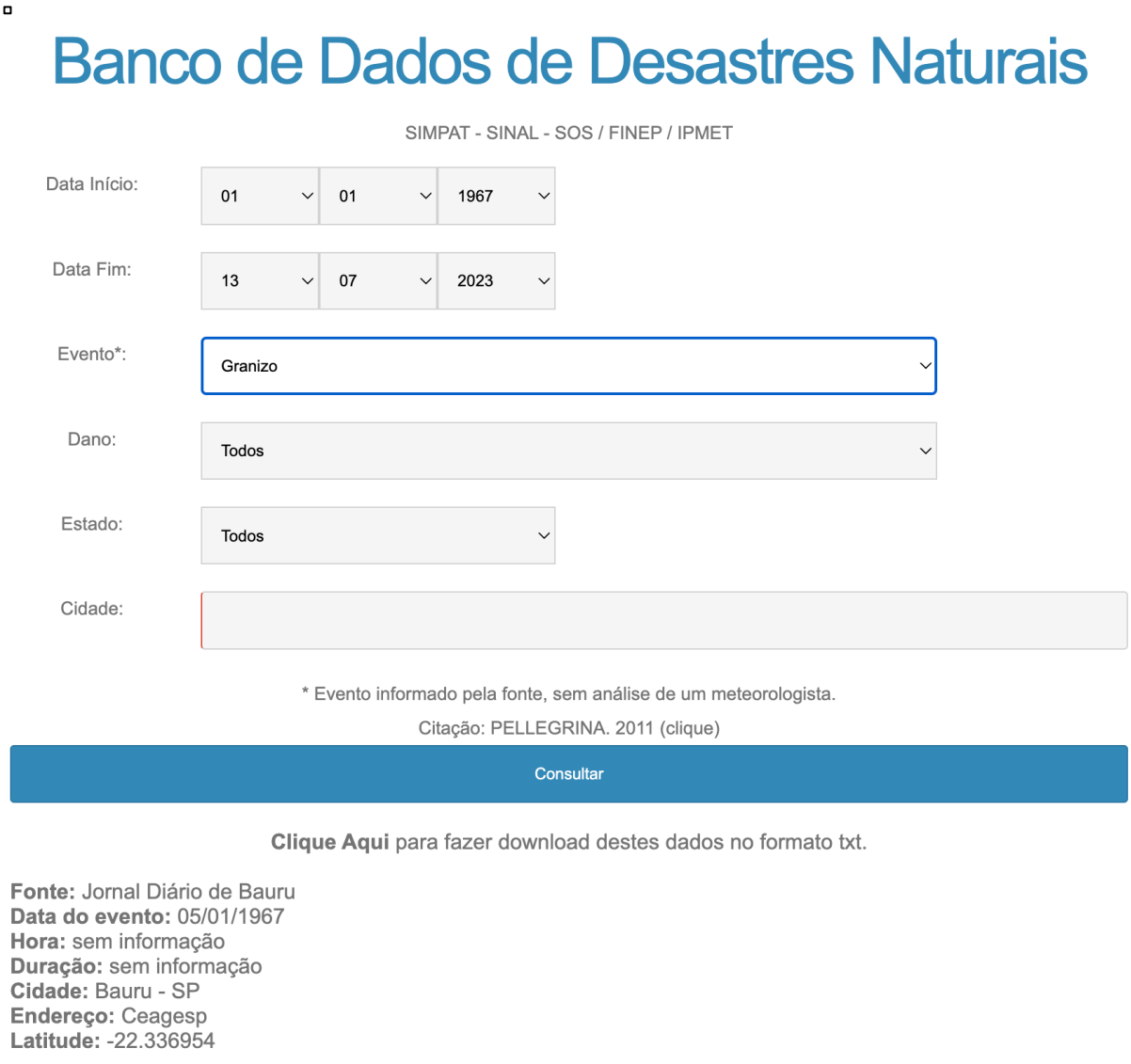
É no bloco “Coletor de dados”, que existe uma função chamada de busca_dados(), ela é responsável por navegar em uma página *web* utilizando a biblioteca Selenium, os exemplos das telas que o Selenium faz a seleção, são apresentados nas Figura 16 e Figura 17, sendo sobre as plataformas S2ID e IPMET respectivamente. Em relação aos eventos climáticos que são coletados na S2ID e IPMET, ambos são apresentados no Quadro 8.

Figura 16 - Exemplo da tela de seleção pelo Selenium da plataforma S2ID



Fonte: Autoria própria

Figura 17 - Exemplo da tela de seleção pelo Selenium da plataforma IPMET



Fonte: Autoria própria

Quadro 8 – Eventos climáticos coletados das plataformas S2ID e IPMET

Eventos climáticos	S2ID	IPMET
Granizo	x	x
Chuva	x	
Tempestade de Raios	x	
Vendaval	x	
Tornado	x	x
Ventos fortes/vendaval		x
Chuvas fortes		x
Raio		x
Ciclone		x
Frente fria/chuvas contínuas		x
Chuvas moderadas		x
Geadas		x
Estiagem		x

Fonte: Autoria própria (2023)

Após a execução da função busca_dados() chegar ao fim, ela retorna uma variável flag que pode ser verdadeira ou falsa, e os dados baixados no formato zip. Então o arquivo zip é descompactado, gerando um arquivo .csv que é lido utilizando a biblioteca “Pandas”, e assim é gerando um dataframe dos dados contendo eventos climáticos. Caso a variável flag retornada da função busca_dados() seja falsa, a função registra uma mensagem em um log de aviso, informando que nenhum dado foi coletado na data e a execução do algoritmo se encerra, além disso, a data também é salva na coleção datas, mesmo que nenhum registro tenha sido encontrado, para evitar que o algoritmo de coleta, percorra a mesma data repetidas vezes.

O dataframe gerado é percorrido em forma de *looping*, procurando remover as linhas que contêm valores nulos ou inválidos. Em seguida, as funções set_data_collection_hash() e set_date() são executadas.

Por fim, o algoritmo de coleta deleta o arquivo csv e zip, anteriormente baixados, e registra uma mensagem de log, informando que a coleta de dados diária foi finalizada.

Em resumo, o algoritmo de coleta é responsável por orquestrar todo o processo de coleta e inserção dos dados nas respectivas coleções do banco de dados,

o Quadro 9 e Quadro 10 exibem os dados que são coletados nas plataformas S2ID e IPMET.

Quadro 9 - Dados da coleção S2ID

Nome do dado	Tipo do dado	Descrição do dado
Id	String	Identificador único do registro, [hash gerado a partir dos dados (Figura 20)]
Status	String	Indica a situação da observação do registro
Data	String	Data do recebimento no formato “YYYY-MM-dd”
Cidade	String	Cidade referente a observação do registro
Uf	String	Estado referente a observação do registro
País	String	País referente a observação do registro
Fonte	String	Referência a plataforma de onde o registro foi coletado, ex: S2ID
População	Int32	Indica a população da cidade referente a observação do registro
Cobrade	String	Indica os fenômenos observados em forma de uma única mensagem
Fk_id	String	Faz referência ao Id, é utilizado para referenciar o registro na tabela “Center Data”
Fenomeno_legenda	Array	Lista de fenômenos em ordem de ocorrência ex: chuva, tempestade, granizo
Fenomeno_numero	Array	Lista com números que fazem referência aos fenômenos, ex: chuva = 1, tempestade = 2

Fonte: Autoria própria (2023)

Quadro 10 - Dados da coleção IPMET

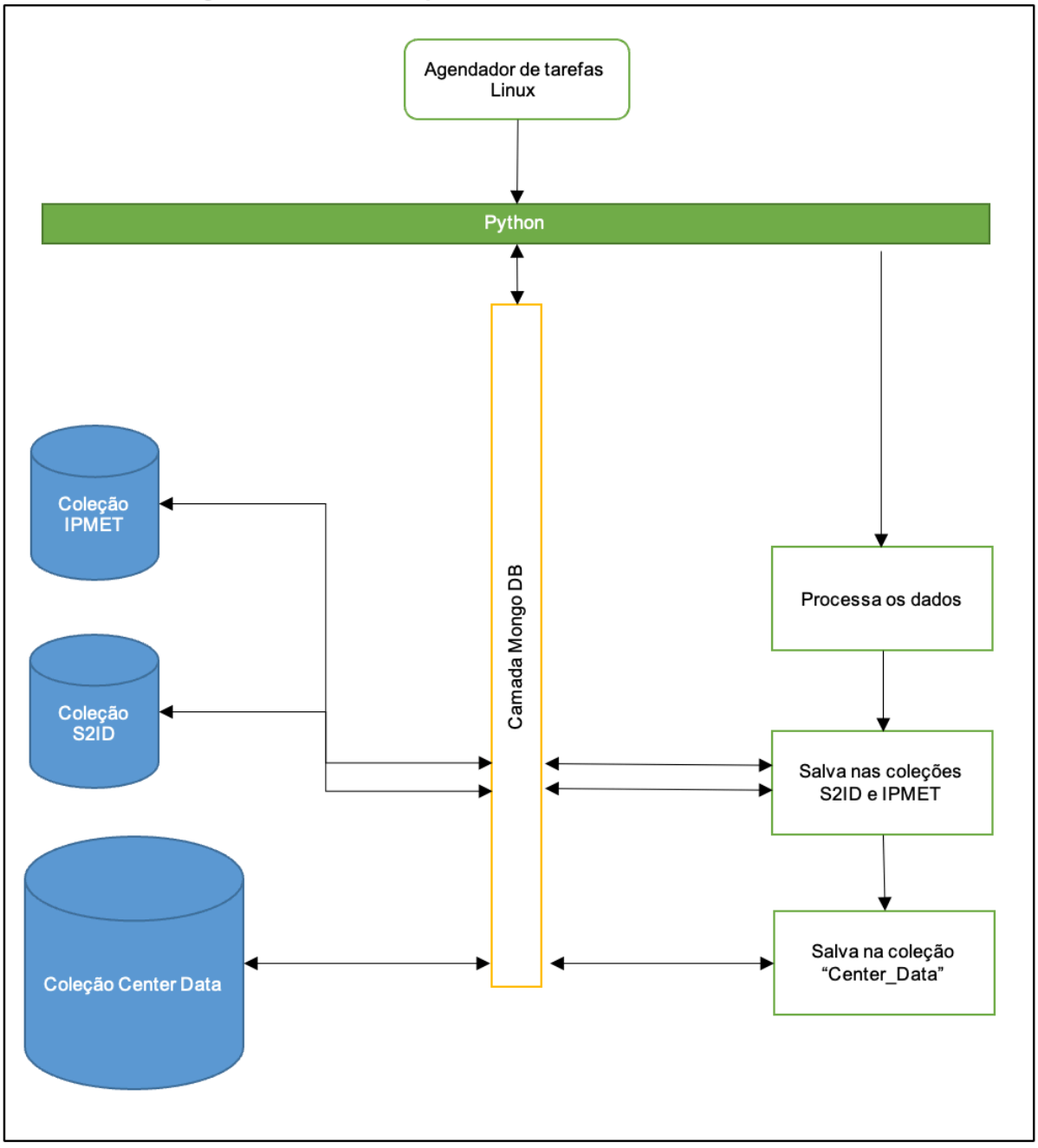
Nome do dado	Tipo do dado	Descrição do dado
Id	String	Identificador único do registro. [hash gerado a partir dos dados (Figura 21)]
Data	String	Data do recebimento no formato “YYYY-MM-dd”
Hora	String	Hora do recebimento
Endereço	String	Endereço do local de observação do registro

Cidade	String	Cidade referente a observação do registro
Uf	String	Estado referente a observação do registro
Longitude	String	Longitude do local referente a observação do registro
Latitude	String	Latitude do local referente a observação do registro
Fonte	String	Referência a plataforma de onde o registro foi coletado, ex: IPMET
Fk_id	String	Faz referência ao Id, é utilizado para referenciar o registro na tabela “Center Data”
Fenomeno_legenda	Array	Lista de fenômenos em ordem de ocorrência ex: chuva, tempestade, granizo
Fenomeno_numero	Array	Lista com números que fazem referência aos fenômenos, ex: chuva = 1, tempestade = 2

Fonte: Autoria própria (2023)

A etapa de processamento das plataformas S2ID e IPMET envolve a criação do HASH de controle, padronização de texto, centralização dos dados na coleção CENTER DATA, e a limpeza dos dados. A Figura 18 exibe um fluxo do processo realizado na etapa de processamento de dados.

Figura 18 - Método de processamento dos dados S2ID e IPMET



Fonte: Autoria própria (2023)

O processamento dos dados das plataformas S2ID e IPMET também são muito parecidos, apenas se diferenciando nos tipos e quantidade de dados coletados, pois cada plataforma possui tipos de dados distintos, sendo assim, são necessários ajustes específicos em cada algoritmo de processamento.

Cada algoritmo tem sua inicialização individual também pelo CRON, assim é possível realizar fluxos de coleta e processamento em paralelo, sendo que o CRON executa diariamente os algoritmos da S2ID as 23:50 e da IPMET as 23:30.

Visando um melhor controle de redundância de dados, foi implementado uma

solução baseada em HASHs. Para geração do HASH são utilizados alguns dados em forma de strings, que são concatenadas, formando uma única linha de caracteres, servindo de entrada de dados para geração do HASH:

- HASH REDEMET (Figura 19)

Figura 19 - String utilizada para criação do hash REDEMET

```
hash = utils.set_md5_hash(str(obs['code'] + i['localidade'] +
                          i['recebimento']).encode('utf-8'))
```

Fonte: Autoria própria (2023)

- HASH S2ID (Figura 20):

Figura 20 - String utilizada para criação do hash S2ID

```
hash = utils.set_md5_hash(str(row['UF'] + str(row['Município']) +
                          str(row['Registro']) + str(row['COBRADE'])+ str(row['Status']) +
                          str(row['População'])).encode('utf-8'))
```

Fonte: Autoria própria (2023)

Grande parte dos dados utilizados na geração do HASH de controle são autoexplicativos, porém ressalta-se alguns deles:

- ‘Status’: faz referência ao registro da observação, na plataforma da S2ID existe um controle de observação que pode assumir alguns valores, como ‘registro’ ou ‘reconhecido’.

- ‘COBRADE’: Pode ser observado nos dados da coleção S2ID, Quadro 7, e se trata do código do evento meteorol´ogico, seguido da legenda, por exemplo “13215 – tempestade local/convectiva – vendaval”.

- HASH IPMET (Figura 21):

Figura 21 - String utilizada para criação do hash IPMET

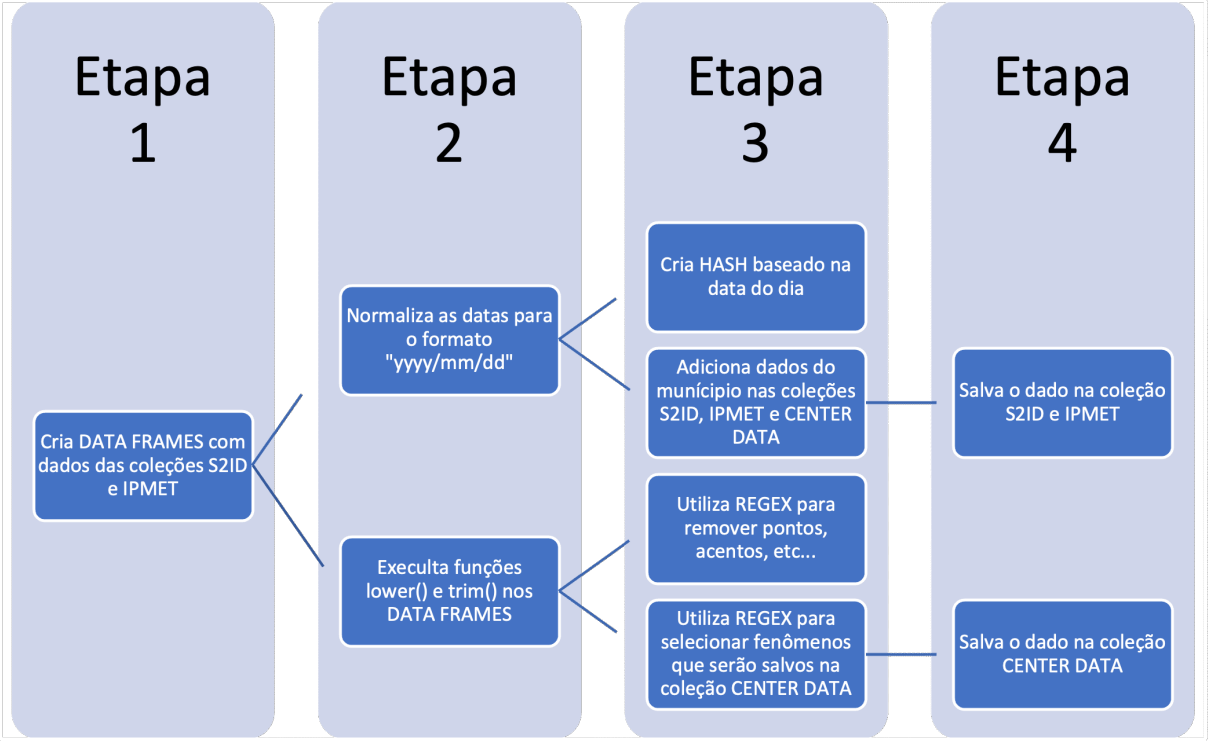
```
hash = utils.set_md5_hash(str(row['Data'] + str(row['Hora']) +
                          str(row['Cidade']) + str(row['Estado'])+ str(row['Endereco']) +
                          str(row['Fenomeno'])).encode('utf-8'))
```

Fonte - Autoria própria (2023)

É importante ressaltar que o processo geração e comparação dos HASHs gerados, é mesmo para as três plataformas, havendo distinção apenas dos dados utilizados para geração da string. Após a geração do HASH, ele é armazenado em uma lista que contém os HASHs já inseridos nas coleções, então a cada nova tentativa de inserção de um HASH, ele é comparado com os HASHs já presentes na lista. Se algum HASH que foi gerado a partir de um dado coletado, tiver o mesmo número de

algum outro HASH contido na lista de HASHs, entende-se que possivelmente são dados duplicados e então o dado é descartado. A Figura 22 apresenta o fluxo de processamento de dados das coleções IPMET e S2ID.

Figura 22 - Fluxo da etapa processamento de dados S2ID e IPMET



Fonte: Autoria própria (2023)

A etapa final consiste em salvar os dados já processados nas respectivas coleções e na coleção CENTER DATA. Antes do registro ser salvo na coleção CENTER DATA, ele passa pelo REGEX para filtragem e seleção apenas dos eventos climáticos de interesse contido na Figura 23.

4.2.7 Armazenamento dos dados na coleção Center Data

Após a etapa da coleta e processamento dos dados das plataformas, é realizada a centralização em uma única coleção de dados que foi chamada de “CENTER DATA”.

Nessa coleção, os dados de cada plataforma passam por um filtro REGEX, que é uma sequência de caracteres que especifica um padrão de busca, a fim de filtrar apenas os fenômenos de interesse para serem centralizados. O Quadro 11 apresenta o código REGEX utilizado:

Quadro 11 - REGEX utilizado para filtrar os eventos para serem centralizados

grãos de neve
cristais de gelo
pelotas de gelo
pelotas de neve
granizo
chuva
neve
tempestade
raios
vendaval
tornado
ventos
chuvisco
ciclone

Fonte: Autoria própria

O Quadro 12 apresenta os dados que são centralizados na coleção Center Data.

Quadro 12 - Dados da coleção Center Data

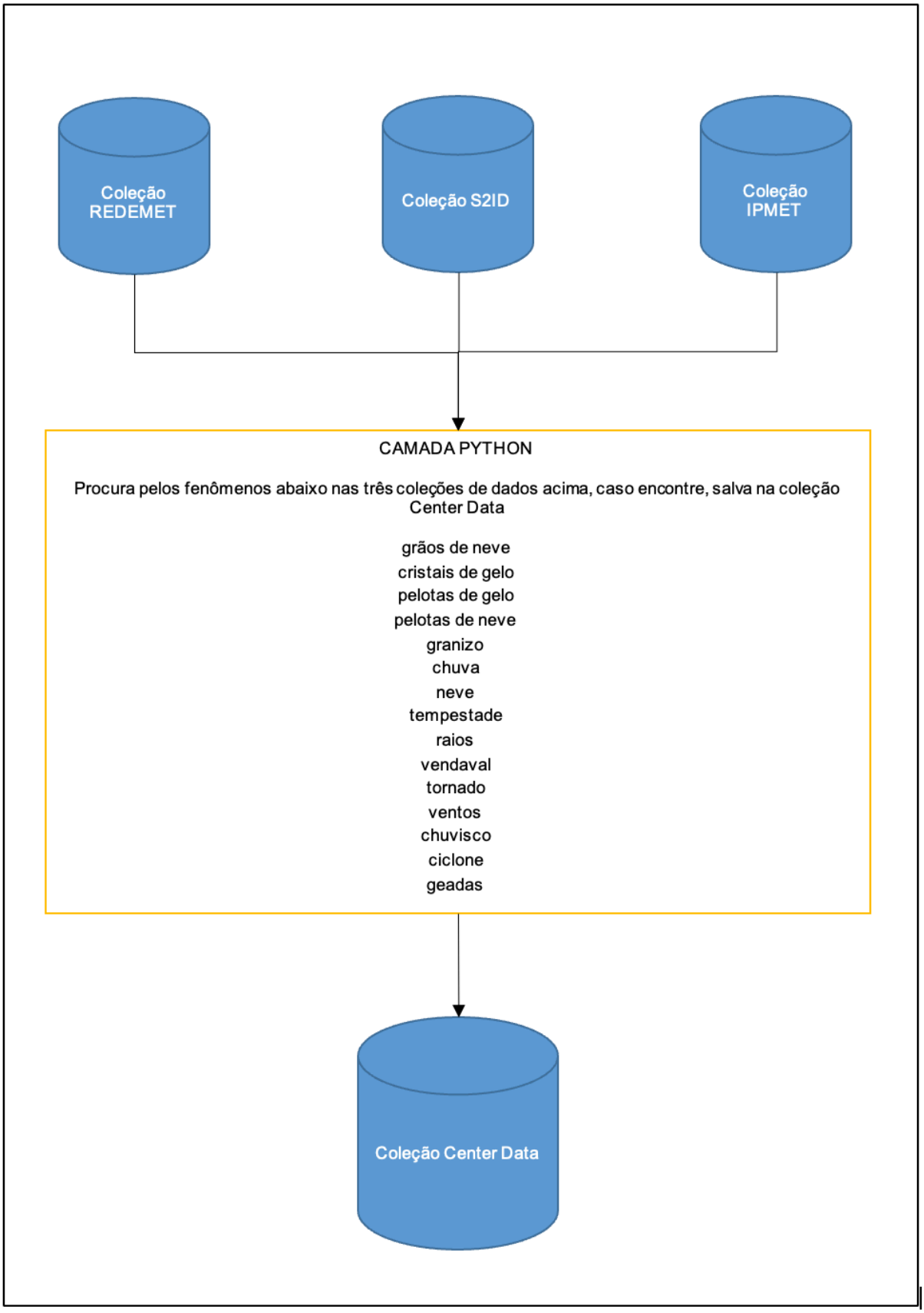
Nome do dado	Tipo do dado	Descrição do dado
Id	String	Idenficador único do registro, Um hash
Data	String	Data do recebimento no formato “YYYY-MM-dd”
Cidade	String	Cidade referente a observação do registro
Uf	String	Estado referente a observação do registro
Longitude	String	Longitude do local referente a observação do registro
Latitude	String	Latitude do local referente a observação do registro
Altitude	String	Altitude do local referente a observação do registro
Fonte	String	Referência a plataforma de onde o registro foi coletado, ex: REDEMET, S2ID ou IPMET
Fk_id	String	Faz referência ao Id, é utilizado para referenciar o registro original, podendo ser de uma das três coleções bases, REDEMET, S2ID ou IPMET

Fenomeno	Array	Lista de fenômenos em ordem de ocorrência ex: chuva, tempestade, granizo

Fonte: Autoria própria (2023)

A Figura 23, ilustra o fluxo de centralização de dados, inicialmente o algoritmo de coleta, salva os dados nas coleções individualmente, conforme foi descrito nos tópicos anteriores. Ao final do ciclo de processamento do algoritmo de coleta, ele executa a função de centralização, que submete cada registro coletado ao filtro REGEX (**Erro! Fonte de referência não encontrada.**) que tenta validar os eventos climáticos descritos na camada Python, e caso sejam validados, são armazenados na coleção Center Data.

Figura 23 - Método de centralização dos dados



Fonte: Autoria própria (2023)

4.3 Aplicação do algoritmo FP-Growth

O fp-growth é um algoritmo utilizado para gerar modelos de regras de associação proposto por Han et al. (2000) a partir de estrutura de dados da árvore de prefixos denominada FPTree. Ele é considerado uma melhoria do algoritmo apriori (HEGLAND, 2007) que identifica os itens frequentes no banco de dados.

O algoritmo fp-growth, faz a classificação dos itens de cada transação em ordem decrescente, sendo que cada transação é uma lista de itens, então os itens mais frequentes são colocados no topo da árvore de transações, enquanto os menos frequentes são colocados mais abaixo. Através dessa estruturação, o algoritmo é capaz de eliminar conjuntos de itens pouco frequentes e se tornar mais eficiente na análise de associações, sendo um algoritmo amplamente utilizado em problemas que envolvem mineração de dados.

A preparação dos dados é essencial para que o algoritmo fp-growth consiga realizar as análises dos padrões de forma correta, o processo de preparação dos dados envolve as seguintes etapas:

- Limpeza dos dados: É importante remover dados inconsistentes, duplicados, incompletos ou irrelevantes. Também pode ser necessário tratar valores ausentes, preenchendo-os com valores padrão ou removendo-os, dependendo do caso.

- Transformação dos dados: Em alguns casos, os dados podem estar em um formato inadequado para serem usados diretamente pelo algoritmo. É necessário transformá-los em um formato adequado. Por exemplo, dados em formato de texto podem precisar de conversão para dados numéricos ou binários.

- Agregação dos dados: Em alguns casos, pode ser necessário agrupar os dados em uma única transação ou linha.

- Seleção de dados: É importante selecionar apenas os dados relevantes para o problema a ser resolvido. Isso pode envolver a escolha de um subconjunto de atributos ou a filtragem de dados com base em um critério específico, por exemplo, selecionar apenas dados meteorológicos envolvendo o granizo.

É importante ressaltar que o algoritmo fp-growth foi utilizado em sua

implementação em Python e será abordado como uma função fpgrowth().

Após a etapa de processamento dos dados, foi utilizada a função fpgrowth() aos dados processados. Essa função é importada da biblioteca mlxtend.frequent_patterns, que recebe três parâmetros de entrada:

- res = **fpgrowth**(dataset, min_support=0.001, use_colnames=True)
    - Primeiro parâmetro: Se refere ao dataset, um conjunto de dados dos eventos climáticos.
    - Segundo parâmetro: É o min_support, se refere ao suporte mínimo que um conjunto de itens deve ter para ser considerado um conjunto frequente e valido, esse número varia de 0 a 1. Esse parâmetro pode ser definido pelo usuário e afeta o número e a qualidade dos conjuntos frequentes encontrados. Nesse estudo foi utilizado como min_support padrão o valor de 0,1%. Por exemplo, o fenômeno de chuva com granizo deverá aparecer em pelo menos em 0,1% dos registros de fenômenos totais, caso isso não aconteça, a função fpgrowth() não irá considerá-lo como um conjunto frequente.
    - Terceiro parâmetro: Indica se as colunas do conjunto de dados de entrada devem ser identificadas pelos nomes das colunas (ou seja, pelos cabeçalhos das colunas) em vez de seus índices numéricos.

A função fpgrowth() retorna um objeto ‘res’ que contém informações sobre os itemsets mais frequentes encontrados, então o objeto res é utilizado como parâmetro para a função association_rules() que será usada após esse retorno da função fpgrowth().

A função association_rules() também faz parte da biblioteca mlxtend.frequent_patterns que recebe mais dois outros parâmetros de entrada, conforme o exemplo abaixo:

- res = **association_rules**(res, metric=str(metric),min_threshold=float(min_threshold))
    - Primeiro parâmetro: É o objeto retornado anteriormente pela função fpgrowth().
    - Segundo parâmetro: O metric pode ser definido pelo usuário e assume valores como ‘*support', 'confidence', 'lift', 'leverage' e*

'conviction'. Nesse estudo foi utilizado como metric padrão o valor de *'lift'.*

- Terceiro parâmetro: É o min_thershold, recebe um valor mínimo para as associações que serão geradas. Nesse estudo foi utilizado como padrão o valor de min_thershold de 1,2.

As funções fpgrowth() e association_rules(), são utilizadas em uma outra função, essa de autoria própria, chamada "create_fpgrowth()".

A função create_fpgrowth(), recebe um conjunto de parâmetros para personalizar a análise, como caminho para o arquivo CSV de dados a serem analisados, limiares de suporte e confiança, métricas, etc.

A função tem como objetivo retornar uma tabela no formato csv, que representa as associações frequentes encontradas e gráficos no formato png utilizando a biblioteca Plotly.

As tabelas e gráficos gerados pela função, foram posteriormente utilizados para criação da dashboard.

4.3.1 Métricas utilizadas para a análise dos dados

As métricas utilizadas para as análises dos dados, consistem em cinco tipos, suporte (*support*), confiança (*confidence*), elevação (*lift*), alavancagem (*leverage*) e convicção (conviction).

- *Support*: Considere um evento climático A e B ocorrem juntamente e seu suporte é dado pela soma de suas ocorrências em conjunto, dividido pelo total de eventos climáticos no conjunto de dados, por exemplo, o "granizo e tempestades" representam um conjunto de eventos ou itens que ocorrem juntos, divididos por todos os eventos climáticos salvos nas coleções de dados.

$$support(A \rightarrow B) = \frac{Eventos\ que\ contêm\ ambos\ A\ e\ B}{Total\ de\ eventos\ climáticos}$$

(1)

$$suporte(tempestade \rightarrow granizo)$$

$$= \frac{número\ de\ vezes\ que\ as\ tempestades\ ocorreram\ junto\ com\ granizo}{número\ total\ de\ eventos\ climáticos\ registrados}$$

(2)

- *Lift:* Utilizado como parâmetro de métrica, indica o suporte de um fenômeno consequente ocorrer, enquanto calcula a probabilidade de um fenômeno B conter um fenômeno A. O *lift* é uma medida da força da associação entre dois itens que mede o quão mais provável é que os itens ocorram juntos em uma transação em comparação com a frequência esperada de sua co-ocorrência se eles fossem independentes. Um "*lift*" maior do que 1 indica uma associação positiva entre os dois itens, enquanto um "*lift*" menor do que 1 indica uma associação negativa ou irrelevante, abaixo a equação do *lift.*

$$lift(A \rightarrow B) = \frac{support(A \rightarrow B)}{(support(A)\ *\ support(B))}$$

(3)

$$lift(granizo\ \rightarrow\ tempestades) = \frac{support(granizo\ \rightarrow\ tempestades)}{\big(support(granizo) * support(tempestades)\big)}$$

(4)

- *Confidence:* É uma medida de quão confiante o algoritmo está na associação entre um conjunto de fenômenos A e um conjunto de fenômenos B. Suponha que o FP-Growth tenha identificado o padrão frequente {A, B, C}. Com base nesse padrão, podemos calcular a confiança da regra de associação {A, B} -> {C}, que representa a probabilidade de que o item C ocorra, dado que os itens A e B também ocorreram. A equação dessa regra pode ser calculada da seguinte forma:

$$confidence(\{A, B\} \rightarrow \{C\}) = \frac{support(\{A, B, C\})}{support(\{A, B\})}$$

(5)

$$confidence(\{tempestade, raios\} -> \{granizo\})$$

$$= \frac{support(\{\ tempestade, raios, granizo\})}{support(\{tempestade, raios\})}$$

(6)

- *Leverage:* É uma medida da diferença entre a frequência observada de co-ocorrência de A e B e a frequência esperada se eles fossem independentes. O valor do *leverage* varia de -1 a 1. Se o valor do *leverage* for 0, isso significa que a ocorrência de A e B juntos é completamente aleatória e não há nenhuma associação entre eles.

$$leverage(A \rightarrow B) = suporte(A -> B) - suporte(A) * suporte(Y)$$
(7)

$$leverage(granizo \rightarrow tempestades) = suporte(granizo -> tempestades) - suporte(granizo) * suporte(tempestades)$$
(8)

- *Convicction*: A convicção é uma métrica que avalia a força da implicação entre dois itens em um conjunto de transações, variando de 0 a infinito. Valores maiores indicam uma implicação mais forte, enquanto valores abaixo de 1 indicam uma diminuição na probabilidade de A ocorrer quando B ocorre, a equação abaixo exemplifica a fórmula.

$$conviction(A \rightarrow B) = \frac{(1 - support(A)))}{(1 - confidence(A \rightarrow B))}$$

(9)

$$conviction(granizo \rightarrow tempestades) = \frac{(1 - support(granizo)))}{(1 - confidence(granizo \rightarrow tempestades))}$$

(10)

4.4 Desenvolvimento da Dashboard de monitoramento

Após a coleta, processamento e centralização dos dados, iniciou-se o desenvolvimento da interface gráfica, também chamada de Dashboard de monitoramento.

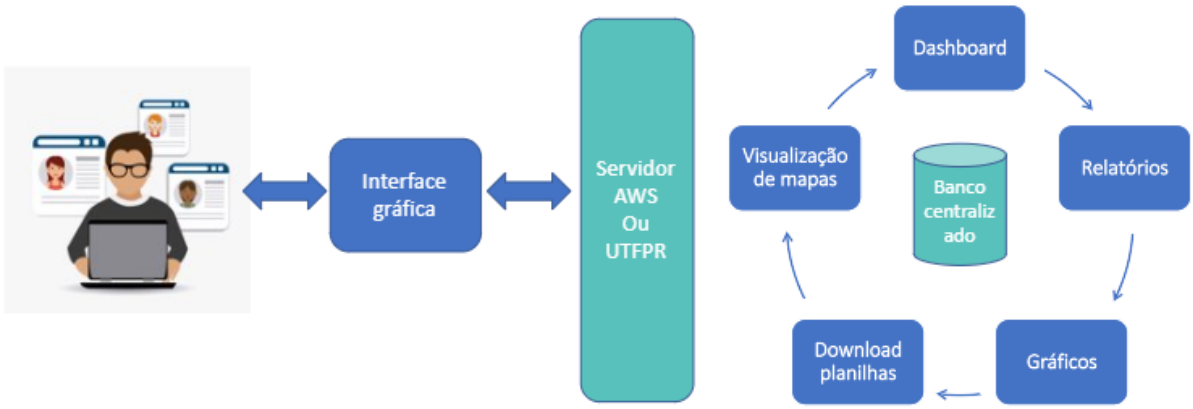
A Dashboard de monitoramento, foi desenvolvida utilizando principalmente os frameworks, Django e Bootstrap, que possuem recursos para acelerar o desenvolvimento. Além, dos frameworks citados, também foi utilizado o Django material admin, sendo ele um template de administração para controle de logins,

autenticação de usuários, e pode ser baixado publicamente na página do GitHub.

O Django material admin, já possui mecanismos de login e senha para acesso, além de painel de navegação já funcionais, sendo necessário apenas algumas modificações para adequá-lo a dashboard de monitoramento.

A dashboard foi dividida em algumas páginas, que exibem as informações como dados, tabelas e gráficos de cada uma das plataformas selecionadas nesse estudo. O ciclo apresentado na Figura 24, ilustra as funcionalidades encontradas na dashboard e como o usuário acessa tais funcionalidades.

Figura 24 - Fluxo de integração usuário com a dashboard



Fonte: Autoria própria (2023)

Os gráficos da dashboard, foram construídos utilizando as bibliotecas Matplotlib e Chart.js, sendo que o Matplotlib constrói gráficos e os salva em formato png, e o Chart.js exibe os gráficos em tempo real na tela.

Os relatórios exibidos na dashboard, são elementos <tables>, e são utilizadas algumas classes de CSS apenas para customização.

O download de planilhas utilizou principalmente a biblioteca Pandas para transformar os Dataframes dos eventos meteorológicos em arquivos csv que podem ser baixados pelo usuário e já estão formatados para possível uso com outras ferramentas.

Por último, a visualização de mapas foi desenvolvida utilizando recursos da Google Clouds, baseado no Google Maps. Após a realização de um cadastro e liberação de chaves de acesso, foi possível utilizar a API da google para apontar locais de interesse nos mapas, utilizando a latitude e longitude presente nos dados coletados e salvos nas coleções. Esses locais no mapa marcam informações sobre eventos de

granizo ocorridos pela região Sul e estado de São Paulo.

Outro ponto a destacar é que o dashboard se encontra online para acesso. Para implantação online, utilizou-se a estrutura da *Amazon Web Services* (AWS), e foi feito um cadastro de um domínio no registro BR, sendo endereço marcospgg.com.br.

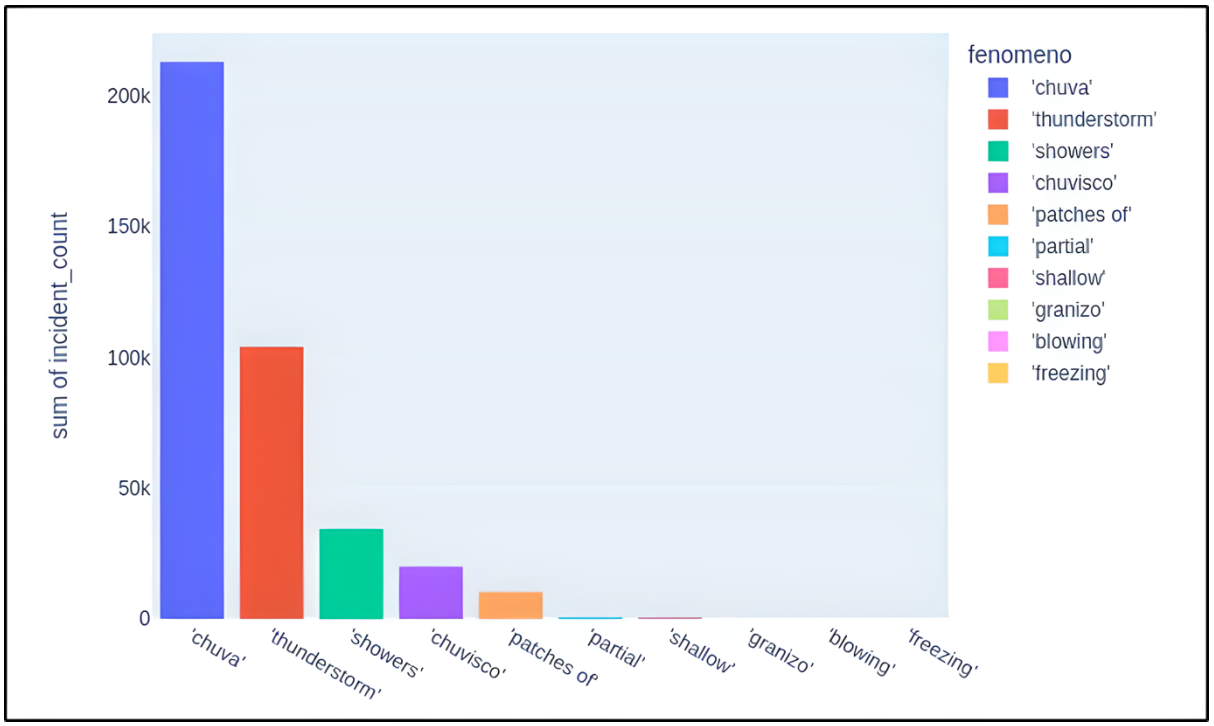
Em resumo, na AWS foram utilizados os recursos do Route 53 para resolução do endereço de DNS, o Virtual Private Server ou também chamado de instancia EC2. Localmente na instancia EC2 foram instalados os bancos de dados MongoDB e PostegreSQL.

5 RESULTADOS E DISCUSSÃO

5.1 Plataforma de dados da REDEMET

Considerando toda a região de estudo, foram encontrados 320.622 registros de eventos atmosféricos entre o período de 1 de janeiro de 2003 até 23 de fevereiro de 2023. Destes, apenas 52 registros contêm dados de tempestade de granizo. A Figura 25 apresenta um gráfico de barras com o número de ocorrências de fenômenos atmosféricos durante o período da coleta. Destaca-se o elevado número de eventos de chuva, tempestades e chuvas leves, que são fenômenos mais comuns em climas tropicais e subtropicais, como é o caso do Brasil. Já o granizo, corresponde à aproximadamente 0,02% do total de eventos meteorológicos. Isso sugere que o granizo é um fenômeno meteorológico raro em comparação com outros eventos mais frequentes, como chuvas e tempestades. No entanto, é importante notar que a ocorrência do granizo pode ser localizada em áreas específicas e pode ocorrer mais frequentemente em algumas regiões do que em outras (MARTINS et al., 2017).

Figura 25 - Gráfico das ocorrências de fenômenos REDEMET



Fonte: Autoria própria (2023)

A Tabela 1 apresenta informações em destaque sobre a coleção REDEMET, contendo o total de ocorrências (eventos meteorólogicos), dados únicos, os valores mais frequentes e o número de vezes que o valor mais frequente foi encontrado. Essa

tabela é muito útil para observar detalhes nos dados.

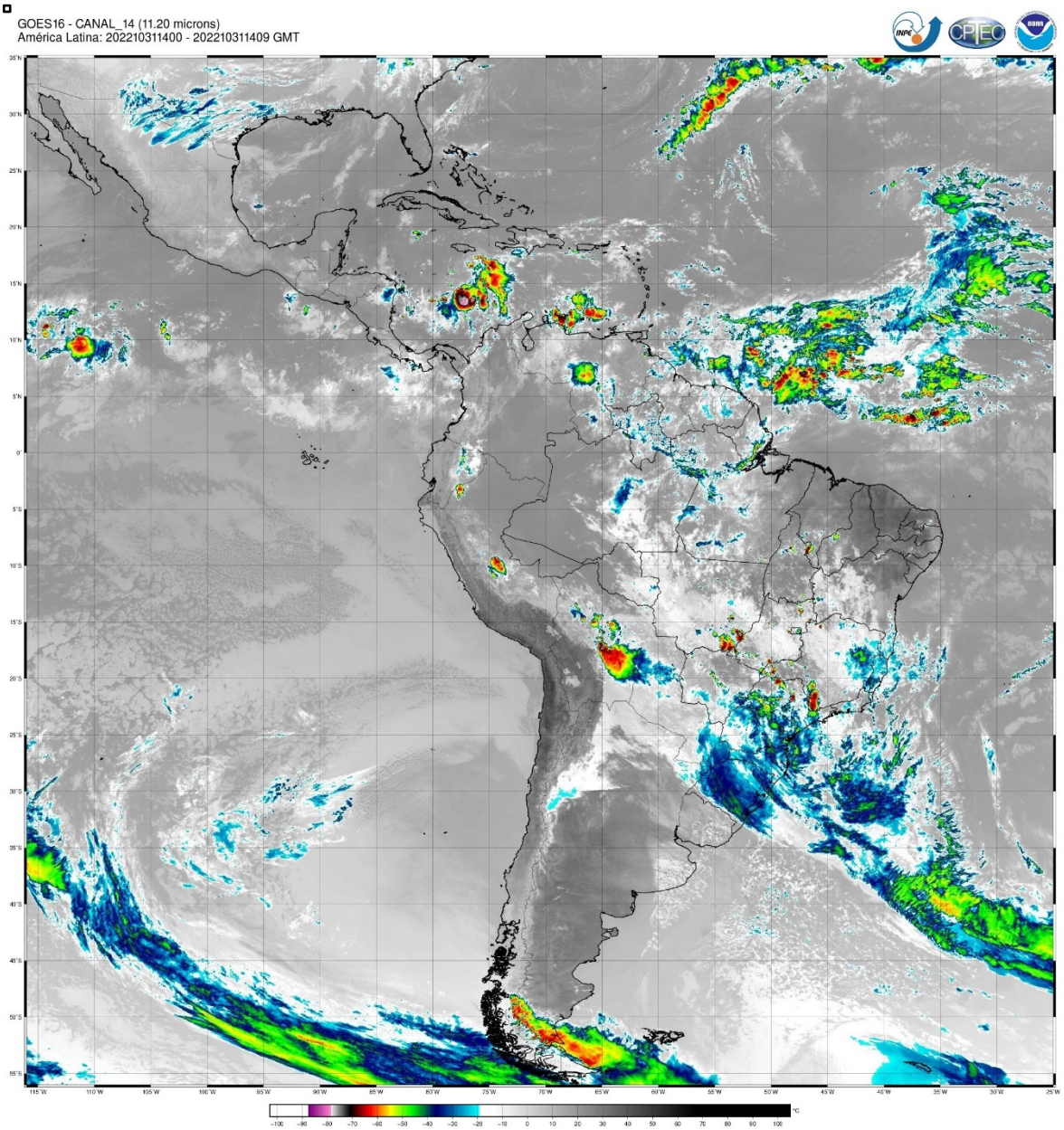
Tabela 1 – Análise exploratória dos dados da coleção REDEMET

Atributo	Total	Dados únicos	Maior Frequência	Número de ocorrências
Data	320622	6945	2022-10-31	284
Aeroporto	320622	47	aeroporto regional de maringá	19793
Cidade	320622	44	são paulo	40532
Uf	320622	4	sp	147886
Pais	320622	1	brasil	320622
Direcao_vento	320622	56	0 degrees	15199
Velocidade_vento	320622	6474	calm	15180
Rajada_vento	320622	67	none	303997
Temperatura	320622	45	21.0 c	35524
Fenomeno_legenda	320622	19	['chuva']	150465
Fenomeno_numero	320622	10	[15]	216967

Fonte: Autoria própria (2023)

Na Tabela 1, destaca-se a data mais frequente de eventos extremos no dia 31 de outubro de 2022 com 284 eventos detectado no METAR. Neste dia, houve a chegada de uma frente fria resultando em diversos temporais com vento forte na região sul do Brasil e no estado de São Paulo (Figura 26). A capital paulista, por exemplo, registrou um acumulado de 30 mm de precipitação, sendo o maior valor do mês (INMET, 2022).

Figura 26 - Imagem de satélite do canal 14 (11.2 μ) do GOES 16 para as 14:00 UTC do dia 31 de outubro de 2022



Fonte: CPTEC-INPE (http://satelite.cptec.inpe.br/acervo/)

A Tabela 2 apresenta informações sobre a frequência de associações entre os principais eventos meteorológicos registrados na coleção de dados "REDEMET" que foi gerada com as funções fpgrowth() e association_rules(). Em geral, os resultados sugerem que a ocorrência de chuva de forma isolada é o fenômeno mais frequente registrado na coleção de dados (aproximadamente 66,4%), seguido de tempestades isoladas (cerca de 32,5%). A associação entre chuva e tempestade é relativamente frequente, com quase 19,5% de ocorrências, sugerindo que esses dois fenômenos muitas vezes ocorrem simultaneamente. Outros fenômenos, como chuva leve, chuvisco também são relativamente comuns.

Um ponto muito imporatante a se destacar, são as somas percentuais da Tabela 2 e outras tabelas de associações que serão apresentadas posteriormente, essa somas em geral não fecham em 100%, pois eventos combinados, se tratam da interssecção de ambos os eventos. Devido a isso, soma total das frequências pode ser maior do que 100%. Por exemplo, a frequência da "chuva" é de 66,45%, e a frequência da "tempestade" é de 32,50%, mas quando consideramos a associação "chuva | tempestade", que inclui os casos em que ambos os fenômenos ocorreram ao mesmo tempo, a frequência é de 19,50%. Nesse caso, a soma das frequências dos três fenômenos é de 118,45%, que é maior do que 100%.

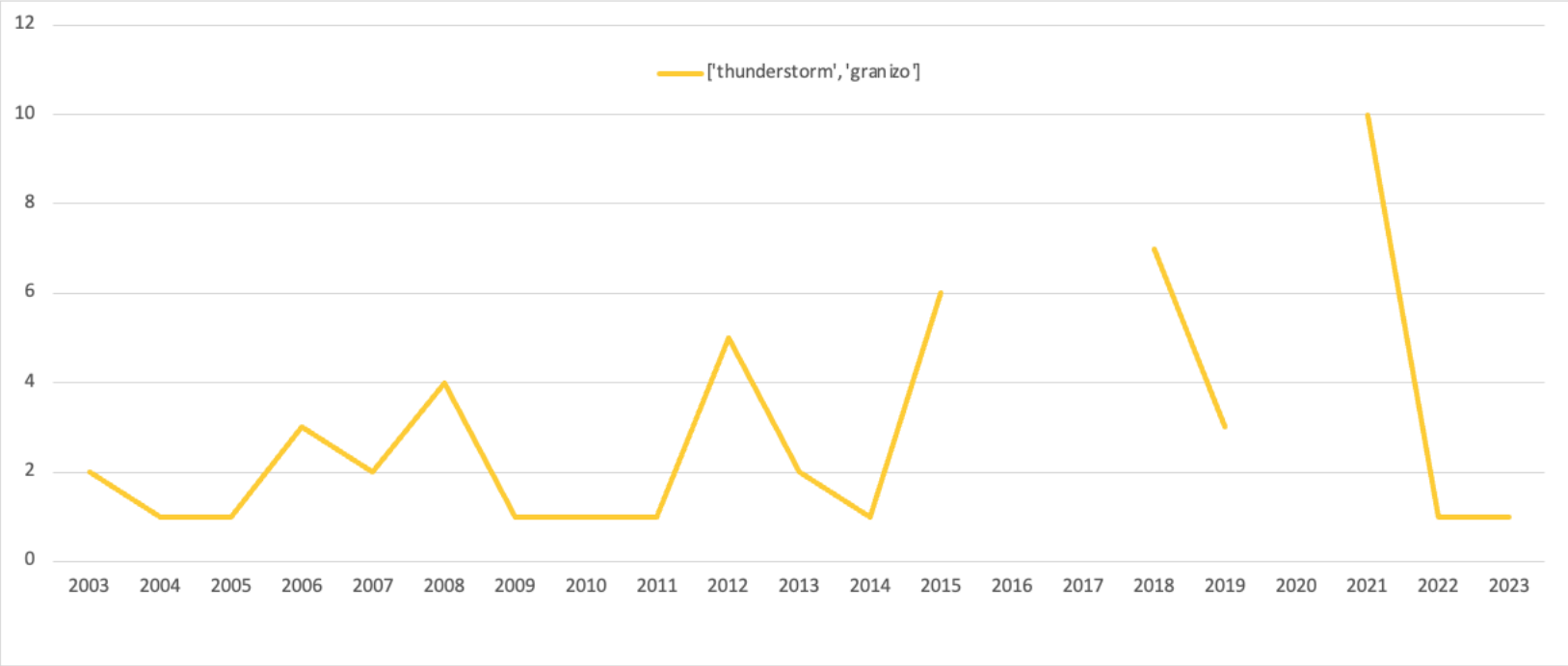
Tabela 2 - Frequência dos fenômenos da coleção REDEMET baseada em itens mais frequentes

Frequência De Fenômenos Total e min_support 0,1%	
Fenômeno	**Ocorrências em %**
chuva	66,45%
tempestade	32,50%
chuva | tempestade	19,50%
chuva leves	10,72%
chuvisco	6,22%
condições parcialmente nubladas	0,22%

Fonte: Autoria própria (2023)

A Figura 27 apresenta o número de ocorrência dos eventos associados por ano entre 2003 e 2023. É interessante notar que a ocorrência de granizo varia bastante de ano para ano. Por exemplo, em 2004, não houve registro de nenhum caso de granizo, enquanto em 2014 houve seis casos. No primeiro caso, houve muitos períodos de estiagem, em especial no estado do Paraná, enquanto no último houve muitos eventos de precipitação na região de estudo (DE BRITO et al., 2022). O **Erro! Fonte de referência não encontrada.** apresenta a associação do granizo com tempestade ao longo dos anos de 2003 e início de 2023.

Figura 27 - Evolução anual das ocorrências do granizo a partir dos dados do banco do REDEMET



Fonte: Autoria própria (2023)

Quadro 13 - Contagem do fenômeno granizo ao longo dos anos REDEMET

Evento Atmosférico	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	Total Geral
['blowing']	1										2	5	3	2	4	2	1	3	4	1	4	32
['chuva']	5633	4906	5609	4351	4906	5784	6530	7092	7355	6857	7976	7292	10593	8848	9341	8789	8589	8097	7901	11604	2412	150465
['chuvisco']	1459	832	1012	742	797	828	857	712	947	902	1133	1162	1387	1088	1033	945	892	1059	828	1147	146	19908
['freezing']																			2	3		5
['low drifting']												1										1
['partial']	25	29	15	12	13	18	14	28	28	22	22	29	40	58	57	22	48	40	50	100	24	694
['patches of']	104	82	184	205	206	312	302	311	328	488	690	685	823	562	605	607	710	432	909	1531	168	10244
['shallow']	17	26	40	49	30	10	34	12	10	17	27	18	17	50	51	30	39	36	42	117	10	682
['showers']	649	662	834	618	645	657	785	953	1157	1203	1600	2081	2511	1999	2192	2184	2472	2252	3018	4111	1798	34381
['thunderstorm']	907	588	1236	1425	1480	1521	2041	1897	2026	2105	2100	2848	3581	2629	2305	2407	2749	1836	2156	2701	1018	41556
Total Geral	10913	8465	11120	9322	10090	11343	13732	13952	14461	14491	16679	18128	24090	18871	19281	18513	19101	17054	18102	24655	6689	319646

Fonte: Autoria própria (2023)

Quadro 14 - Contagem do fenômeno granizo ao longo dos anos REDEMET

Evento Atmosférico	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	Total Geral
['blowing']	1										2	5	3	2	4	2	1	3	4	1	4	32
['chuva']	5633	4906	5609	4351	4906	5784	6530	7092	7355	6857	7976	7292	10593	8848	9341	8789	8589	8097	7901	11604	2412	150465
['chuvisco']	1459	832	1012	742	797	828	857	712	947	902	1133	1162	1387	1088	1033	945	892	1059	828	1147	146	19908
['freezing']																			2	3		5
['low drifting']												1										1
['partial']	25	29	15	12	13	18	14	28	28	22	22	29	40	58	57	22	48	40	50	100	24	694
['patches of']	104	82	184	205	206	312	302	311	328	488	690	685	823	562	605	607	710	432	909	1531	168	10244
['shallow']	17	26	40	49	30	10	34	12	10	17	27	18	17	50	51	30	39	36	42	117	10	682
['showers']	649	662	834	618	645	657	785	953	1157	1203	1600	2081	2511	1999	2192	2184	2472	2252	3018	4111	1798	34381
['thunderstorm']	907	588	1236	1425	1480	1521	2041	1897	2026	2105	2100	2848	3581	2629	2305	2407	2749	1836	2156	2701	1018	41556
Total Geral	10913	8465	11110	9322	10090	11343	13732	13952	14461	14491	16679	18128	24090	18871	19281	18513	19101	17054	18102	24655	6689	319646

Fonte: Autoria própria (2023)

Quadro 15 - Lista com os eventos atmosféricos coletados e seus totais

Evento Atmosférico	Total Geral
['freezing', 'chuva']	4
['freezing', 'chuvisco']	6
['thunderstorm', 'chuva']	64
['thunderstorm', 'chuvisco']	1388
['thunderstorm', 'granizo']	104
['thunderstorm', 'showers', 'chuva']	12
['thunderstorm', 'showers']	88
['blowing']	64
['chuva']	350892
['chuvisco']	42234
['low drifting']	2
['partial']	1388
['patches of']	20488
['shallow']	1364
['showers', 'thunderstorm']	6
['showers']	68762
['thunderstorm', 'chuva', 'showers']	6
['thunderstorm', 'granizo', 'showers']	20
['thunderstorm', 'showers', 'chuva']	12
['thunderstorm', 'showers']	88
['thunderstorm', 'chuva', 'showers', 'showers']	2
['thunderstorm', 'showers', 'showers', 'showers']	2
['thunderstorm', 'showers', 'showers']	2
['thunderstorm']	83112
Total Geral	641244

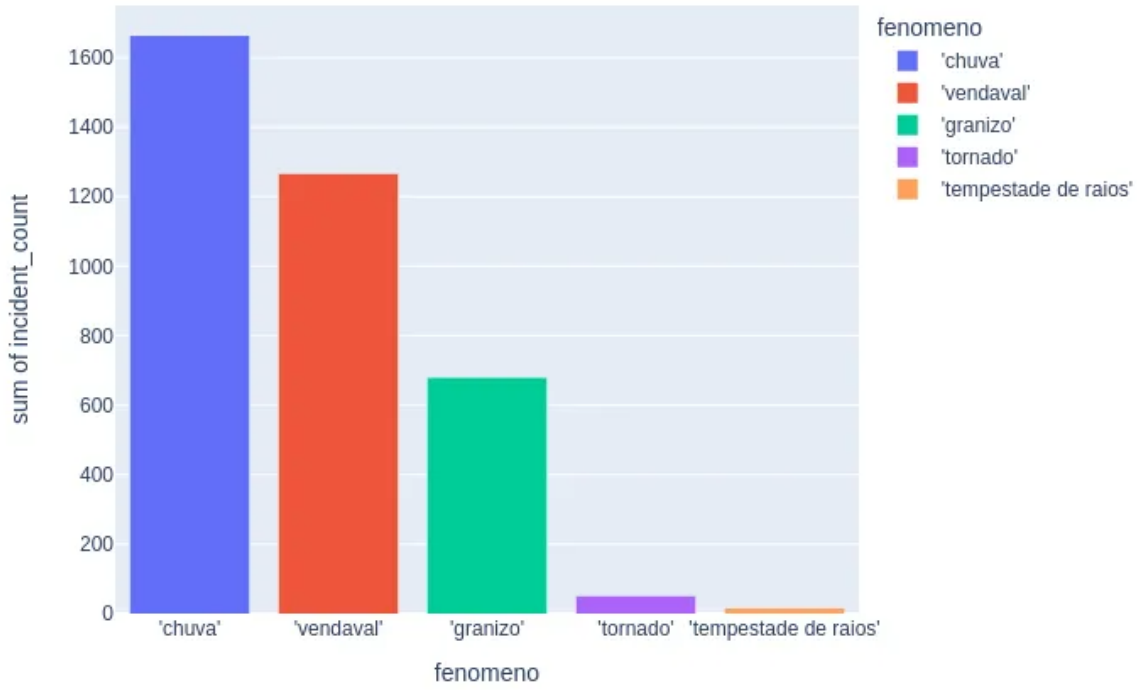
Fonte: Autoria própria (2023)

Ao analisar as associações de fenômenos da tabela, pode-se observar que a maioria das ocorrências de fenômenos climáticos são em dias de chuva, seguidos por dias de chuvisco e tempestades. Além disso, há uma associação significativa entre tempestades e raios (listados como "Tempestade"). Em geral, os dados da tabela mostram que a chuva é a forma de precipitação mais comum na região analisada, seguida pelo chuvisco e outras formas de precipitação menos frequentes, como granizo.

5.2 Plataforma de dados S2ID

Para os dados da plataforma de dados S2ID foram encontrados 3670 registros de eventos atmosféricos entre o período de 1 de janeiro de 2013 até 23 de fevereiro de 2023, com 678 registros contendo alguma informação de granizo. A Figura 28 resume as ocorrências dos eventos durante o período de dados disponível.

Figura 28 - Gráfico das ocorrências de fenômenos S2ID



Fonte: Autoria própria (2023)

A análise da figura mostra que a chuva é o fenômeno mais comum, com 1663 ocorrências registradas. Em seguida, o vendaval, com 1265 ocorrências, enquanto o tornado e a tempestade de raios são os fenômenos menos comuns, com apenas 50

e 14 ocorrências registradas na coleção S2ID, respectivamente. Diferentemente da coleção de dados da REDEMET, no S2ID, o granizo é o terceiro fenômeno mais comum registrado. Uma justificativa para este elevado número de registro comparado as outras coleções, é que o granizo é um fenômeno potencialmente perigoso capaz de causar danos materiais e pessoais, e a coleção de dados da S2ID está associada as ocorrências registradas na Defesa Civil dos estados brasileiros.

A Tabela 3 apresenta informações em destaque dos atributos extraídos de ocorrência de eventos meteorológicos da coleção de dados S2ID. Assim como no caso da Tabela 1, apresenta-se o total de ocorrências, dados únicos, o valor mais frequente e o número de vezes que o valor mais frequente foi encontrado. Com relação aos atributos disponíveis na base de dados S2ID, estes são:

- Status: É referente ao status do dado, com 3 valores possíveis: registro, atualização e exclusão. A maioria dos dados (1916) está com o status de "registro".
- Data: contém informações sobre a data em que o evento meteorológico foi registrado. Existem 1392 datas únicas na coleção de dados.
- Uf: contém informações sobre o estado em que o evento meteorológico foi registrado. Existem 4 estados únicos no conjunto de dados: SC, PR, RS e SP, sendo SC o estado mais comum, com 1862 ocorrências.
- Cidade: contém informações sobre a cidade em que o evento meteorológico foi registrado. Existem 1062 cidades únicas no conjunto de dados, com Santa Cruz do Sul sendo a cidade mais comum, com 49 ocorrências.
- Cobrade: É classificação dos diferentes tipos de eventos meteorológicos registrados. Existem 5 categorias únicas, sendo a categoria "tempestade local/convectiva - chuvas intensas" a mais comum, com 1663 ocorrências.
- Fenomeno_legenda: É a descrição textual do evento meteorológico registrado.
- Fenomeno_numero: É o código numérico atribuído ao evento meteorológico registrado.
- Fk_id: parece ser um identificador único para cada registro, com o mesmo valor do _id em todos os registros.

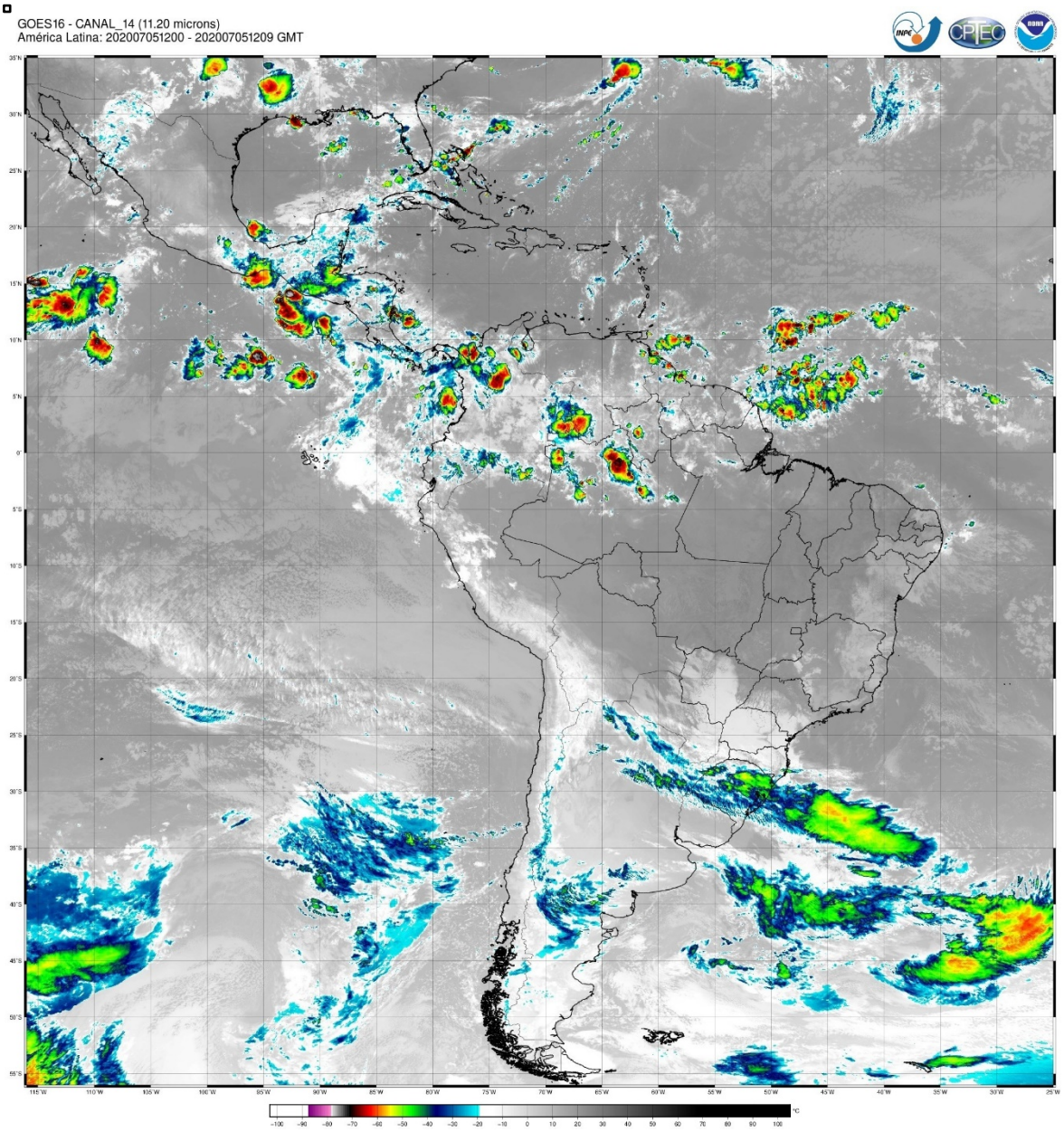
Tabela 3 - Detalhes técnicos obtidos dos dados da coleção S2ID

Atributo	Total	Dados únicos	Maior Frequência	Número de ocorrências
Status	3670	3	registro	1916
Data	3670	1392	2020-07-06	60
Uf	3670	4	sc	1862
Cidade	3670	1062	santa cruz do sul	49
Cobrade	3670	5	tempestade local/convectiva - chuvas intensas	1663
Fenomeno_legenda	3670	5	['chuva']	1663
Fenomeno_numero	3670	5	[2]	1663

Fonte: Autoria própria (2023)

Assim como no caso do REDEMET, a Tabela 3 apresenta algumas informações destacadas, como por exemplo, a data mais frequente entre os dados registrados, que foi o dia 06 de julho de 2020. Este dia foi após a chegada de um sistema frontal que trouxe chuvas e muito frio para a região do Rio Grande do Sul (Figura 29). Nesta primeira semana de julho de 2020, a Secretaria da Agricultura, Pecuária, Produção Sustentável e Irrigação do estado do Rio Grande do Sul emitiu um alerta de muita chuva, frio e geada no estado (SAPPI-RS, 2020). Além disso, a cidade gaúcha de Santa Cruz do Sul, outro registro frequente da Tabela 3, teve recorde de precipitação em um único dia nesta primeira semana de julho de 2020, com 116 mm (JORNAL CIDADES, 2020).

Figura 29 - Imagem de satélite do canal 14 (11.2 μ) do GOES 16 para as 09:00 UTC do dia 05 de julho de 2020



Fonte: CPTEC-INPE (http://satelite.cptec.inpe.br/acervo/)

A Tabela 4 apresenta a frequência de diferentes fenômenos meteorológicos registrados na coleção S2ID. O suporte mínimo de 0,1% significa que apenas os fenômenos que aparecem em pelo menos 0,1% das ocorrências foram considerados. Isso pode garantir que algumas associações mais raras, sejam consideradas, como por exemplo o granizo que geralmente é consideravelmente raro. A chuva é o fenômeno mais frequente registrado, representando 45% de todas as ocorrências com associações. Em segundo lugar, vem o vendaval, com 34% das ocorrências, seguido pelo granizo, com 0,18% das ocorrências. Tornado e tempestade de raios foram registrados em proporções significativamente menores, representando apenas 1% e

0,3%, respectivamente. Nota-se que nenhuma associação entre os eventos foi observada com o valor de suporte estipulado.

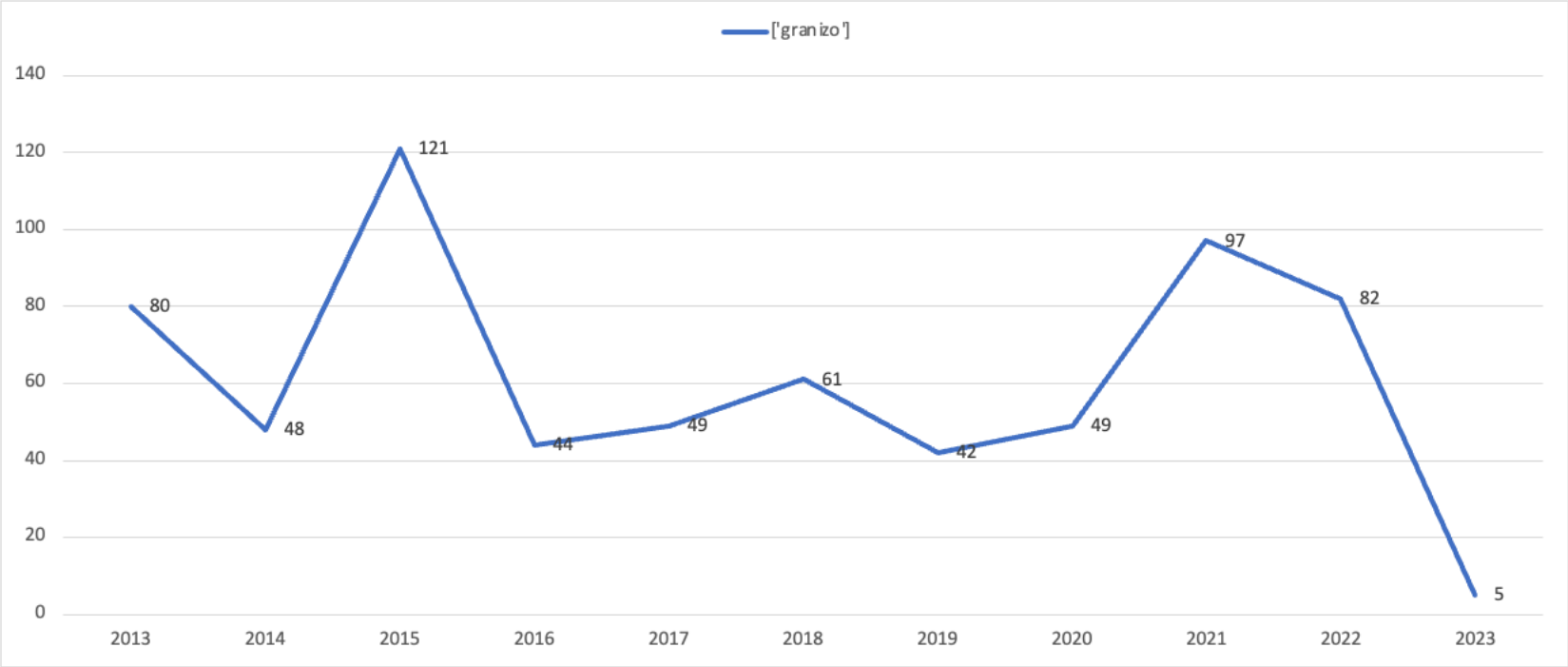
Tabela 4 - Frequência dos fenômenos da coleção S2ID

Frequência de fenômenos total e min. Suporte: 0,001	
Fenômenos	**Ocorrências em %**
chuva	45,31%
vendaval	34,47%
granizo	18,47%
tornado	1,36%
tempestade de raios	0,38%

Fonte: Autoria própria (2023)

Com relação ao granizo, esses dados sugerem que é um fenômeno meteorológico relativamente comum especificamente nessa coleção de dados, porém menos frequente do que a chuva e o vendaval. A Figura 30 e a Quadro 16 apresentam os dados de ocorrências de granizo em diferentes anos a partir da base de dados do S2ID. Nota-se que entre 2013 e 2023 (até a data atual), o número de ocorrências de granizo tem se mantido relativamente estável, com a maior quantidade de ocorrências em 2015 (121) e um total geral de 678 ocorrências no período de 10 anos. Nota-se que as ocorrências de granizo são um fenômeno variável anualmente. É importante ressaltar que a análise é baseada em dados limitados e específicos, e não pode ser generalizada para outras regiões ou períodos.

Figura 30 - Gráfico de evolução temporal de ocorrências de granizo por ano baseado na base de dados do S2ID



Fonte: Autoria própria (2023)

Quadro 16 - Número de ocorrências por ano de eventos atmosféricos extremos obtidos da base de dados do S2ID

	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	Total Geral
['chuva']	106	225	136	95	200	92	104	80	147	443	35	**1663**
['granizo']	80	48	121	44	49	61	42	49	97	82	5	**678**
['tempestade de raios']		2	3		2		2	1	1	1	2	**14**

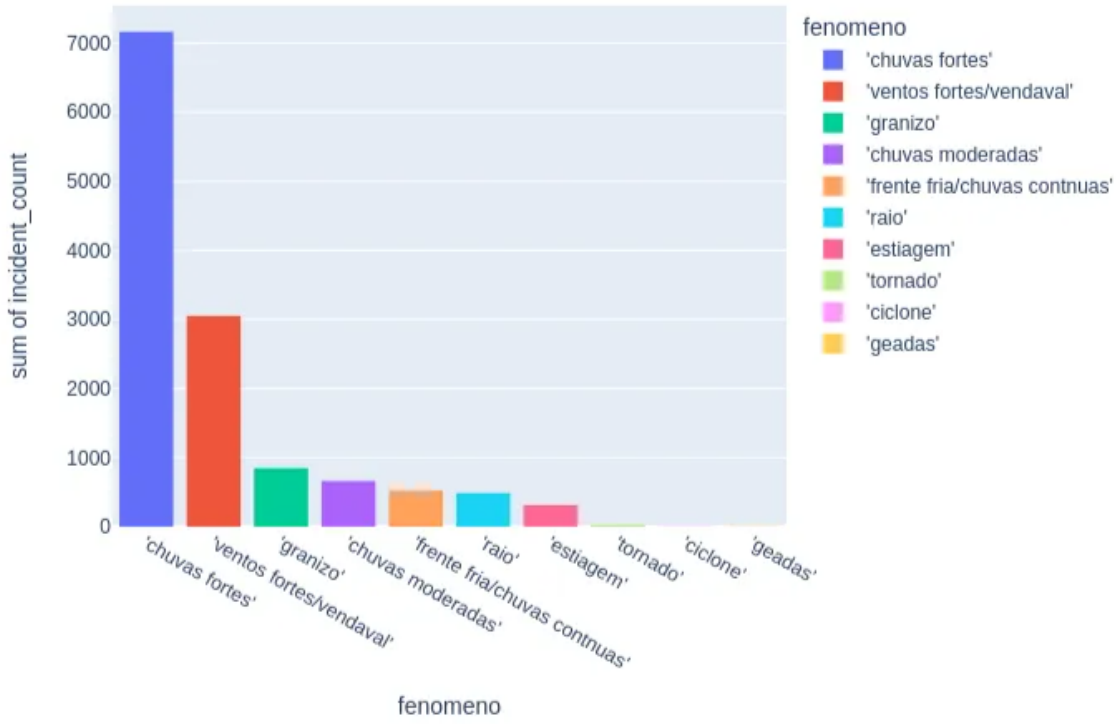
['tornado']	2	6	9	3	6	8		11	3	2		**50**
['vendaval']	85	73	99	49	177	90	67	374	111	132	8	**1265**
Total Geral	**273**	**354**	**368**	**191**	**434**	**251**	**215**	**515**	**359**	**660**	**50**	**3670**

Fonte: Autoria própria (2023)

5.3 Plataformas de dados do IPMET

Com relação a plataforma de dados do IPMet foram encontrados 8988 registros de todos os eventos atmosféricos extremos entre o período de 1 de janeiro de 1967 até 23 de janeiro de 2023. Durante este período, observou-se 847 registros contendo granizo, correspondendo a 9,42%. A Figura 31 resume as ocorrências de fenômenos atmosféricos durante o período de dados disponível. Observa-se que o fenômeno mais frequente é a ocorrência de chuvas fortes, com um total de 7.165 ocorrências registradas. Em seguida, temos ventos fortes/vendaval com 3.052 ocorrências e granizo com 847 ocorrências. As ocorrências de chuvas moderadas, frente fria/chuvas contínuas, raio, estiagem, tornado e ciclone apresentam números menores, com 658, 521, 486, 313, 27 e 9 ocorrências, respectivamente. Destaca-se que esta base tem uma particularidade, pois só abrange informações dos estados de São Paulo e do Paraná.

Figura 31 - Gráfico das ocorrências de fenômenos extremos da base de dados do IPMET



Fonte: Autoria própria (2023)

A Tabela 5 apresenta informações em destaque dos atributos extraídos de ocorrência de eventos meteorológicos da coleção de dados do IPMET. Neste caso, a

data mais frequente foi o dia 01 de dezembro de 1981, com várias ocorrências de vento forte/vendaval - e em alguns lugares chuva forte - ao longo de todo estado do Paraná. Estas chuvas resultaram em centenas de desabrigados no estado, conforme noticiado nos jornais da época Figura 32.

Tabela 5 - Descrição dos dados da coleção IPMET

Atributo	Total	Dados únicos	Maior Frequência	Número de ocorrências
Data	8988	2693	1981-12-01	83
Hora	8988	685	sem informação	2286
Uf	8988	102	sp	6407
Endereço	8988	2785	-1	4970
Cidade	8986	1354	bauru	1230
Estado	8988	102	SP	6407
Latitude	8988	2885	-1	4763
Longitude	8988	2974	-1	4763
Uf	8988	102	sp	6407
Fenomeno_número	8988	48	[3]	3946
Fenomeno_legenda	8988	48	['chuvas fortes']	3946

Fonte: Autoria própria (2023)

Figura 32 - Notícias do dia 02 de dezembro de 1981 resultado dos efeitos das chuvas e vento forte no estado do Paraná. (a) Diário do Paraná e (b) Diário da Tarde

Prefeituras apontadas como responsáveis por enchentes

A Prefeitura Municipal de Curitiba e as prefeituras da Região Metropolitana estão sendo apontadas como principais responsáveis pelas enchentes ocorridas na periferia de Curitiba que causaram graves prejuízos e desabrigaram centenas de famílias. A falta de infra-estrutura urbana e a autorização das prefeituras para o loteamento de áreas impróprias para a construção de residências, segundo uma grande parcela das vítimas, são as causas das tragédias.

Somente no Centro de Triagem da Fundação de Promoção Social do Paraná, da Secretaria de Saúde, foram abrigadas 400 pessoas, entre as seis horas e as 23 horas de segunda-feira última. Eram moradores de casas que estavam submersas nas águas do Rio Atuba e seus tributários, nas regiões de Santa Cândida, Rio Verde e Campo Alto, recolhidos pelo pessoal do Corpo de Bombeiros e da Polícia Militar.

dos, a responsabilidade pela autorização do loteamento dessas áreas, sem verificar a necessidade de implantação de uma infra-estrutura, principalmente no tocante à altura do terreno, em relação aos níveis atingidos, anualmente, pelas águas dos rios da periferia da cidade.

Ontem, em Santa Cândida, um grupo de moradores de casas que foram atingidas pelas águas, estavam tentando reunir um maior número possível de vítimas, para verem a possibilidade de uma ação judicial contra a Prefeitura de Curitiba e de Colombo. A reportagem do Diário do Paraná não conseguiu fazer contato com esse grupo, mas algumas pessoas se prontificaram a procurá-lo para fazer contato.

A Secretaria de Saúde distribuiu informações ontem, dando conta do atendimento que vêm sendo dado às vítimas da cheia, homens, mulheres e crianças, resgatados de suas casas submersas. Segundo as informações, através do Cetas, foram distribuídos 400 cobertores, lençóis, fronhas, colchões e 800 refeições até a terde de ontem, com 300 quilos de leite em pó, 1800 pães e 15 quilos de café (LBA—Promopar). Foram distribuídos 120 quilos de roupas usadas.

O diretor do Centro de Triagem, José Maria Tank, fez um novo apelo à comunidade, na tarde de ontem, para que sejam enviados àquele centro, mantimentos, principalmente, já que tinha em suas residências.

Estão sendo pedidos gêneros alimentícios para 180 crianças e 220 adultos. O telefone do Cetas é 262-6565.

(a)

ENCARTE

Mau tempo continuou, mas sem as enchentes

Depois da enchente de anteontem, que atingiu mais profundamente o Jardim Maracanã, no bairro Santa Cândida, as chuvas do amanhecer de ontem voltaram a preocupar os moradores das zonas baixas da cidade, porque ninguém sabia o que poderia acontecer nas horas seguintes. É aquela história de que "gato escaldado tem medo de água fria". Principalmente se é água que transborda e vai invadindo tudo, destruindo o que está pela frente, causando prejuízos de todos os tipos e ameaçando vidas. Felizmente, o que caiu de água, ontem, foi apenas um susto e nada mais, porque nenhum alagamento mais problemático atingiu a cidade. Mesmo assim, quem mora nos bairros mais distantes, sem vias pavimentadas, continuou sofrendo as conseqüências da falta de uma infraestrutura básica que lhes garanta um pouco mais de conforto. Os lamaçais causados pelo maior temporal dos últimos dias, o de anteontem, não desapareceram e atravessar determinadas ruas, só com o pé no chão e calças arregaçadas, como nas picadas do interior. Por outro lado, em outros locais onde as ruas se apresentavam um pouco melhor, as chuvas lavaram o revestimento de saibro e muitas crateras começaram a aparecer. Se o tempo não melhorar, a situação ficará ainda pior, pois cada pancada abrirá mais os buracos, tornando as vias insuportáveis para os veículos que, evidentemente, sofrem com isso em sua estrutura. Valetas entupidas são outro problema que agrava mais a situação de milhares de moradores periféricos, sempre que chove.

(b)

Fonte: Memorial da Biblioteca Nacional.

A Tabela 6 apresenta a frequência de ocorrência de eventos climáticos individuais e a Tabela 7 a ocorrência das associações dos eventos climáticos, pelos algoritmos fpgrowth() e association_rules(), sendo que a associação mínima de suporte é de 0,1%. O fenômeno mais frequente é o de chuvas fortes, com uma ocorrência de 79,71%, seguida de vento forte/vendaval, com uma ocorrência de 33,96%, e granizo, com uma ocorrência de 9,42%. O granizo é o quarto fenômeno mais frequente. Diferentemente das outras coleções de dados, as associações mais fortes com o granizo são com chuvas fortes, ventos fortes/vendaval e raio. A associação entre ventos fortes/vendaval e chuvas fortes ocorreu 28,22%, e a associação entre ventos fortes/vendaval, chuvas fortes e raio aconteceu com uma ocorrência de 3,57%.

Tabela 6 - Frequência dos fenômenos individuais da coleção IPMET

Frequência de fenômenos total com associações e min_support: 0,1%

Fenômenos	Ocorrências em %
chuvas fortes	79.72%
granizo	9.42%
chuvas moderadas	7.32%
frente fria/chuvas contínuas	5.80%
raio	5.41%
estiagem	3.48%
tornado	0.30%
ciclone	0.10%

Fonte: Autoria própria (2023)

Tabela 7 - Frequência das associações dos fenômenos da coleção IPMET

Frequência de fenômenos total com associações e min_support: 0,1%	
Fenômenos	**Ocorrências em %**
ventos fortes/vendaval	33.96%
ventos fortes/vendaval | chuvas fortes	28.22%
chuvas fortes | chuvas moderadas	4.45%
chuvas fortes | raio	4.34%
ventos fortes/vendaval | raio	3.65%
chuvas fortes | granizo	3.63%
ventos fortes/vendaval | chuvas fortes | raio	3.57%
ventos fortes/vendaval | granizo	2.74%
ventos fortes/vendaval | chuvas fortes | granizo	2.61%
ventos fortes/vendaval | chuvas moderadas	1.25%
ventos fortes/vendaval | chuvas fortes | chuvas moderadas	0.66%
granizo | raio	0.61%
chuvas fortes | chuvas moderadas | frente fria/chuvas contínuas	0.56%
chuvas fortes | granizo | raio	0.52%
ventos fortes/vendaval | granizo | raio	0.40%
ventos fortes/vendaval | chuvas fortes | granizo | raio	0.39%
tornado | ventos fortes/vendaval	0.18%
tornado | chuvas fortes	0.18%
tornado | chuvas fortes | ventos fortes/vendaval	0.17%
ventos fortes/vendaval | frente fria/chuvas contínuas	0.17%

chuvas moderadas | granizo	0.14%
ventos fortes/vendaval | chuvas moderadas | raio	0.13%
granizo | chuvas fortes | chuvas moderadas	0.12%
ventos fortes/vendaval | chuvas fortes | frente fria/chuvas contínuas	0.11%

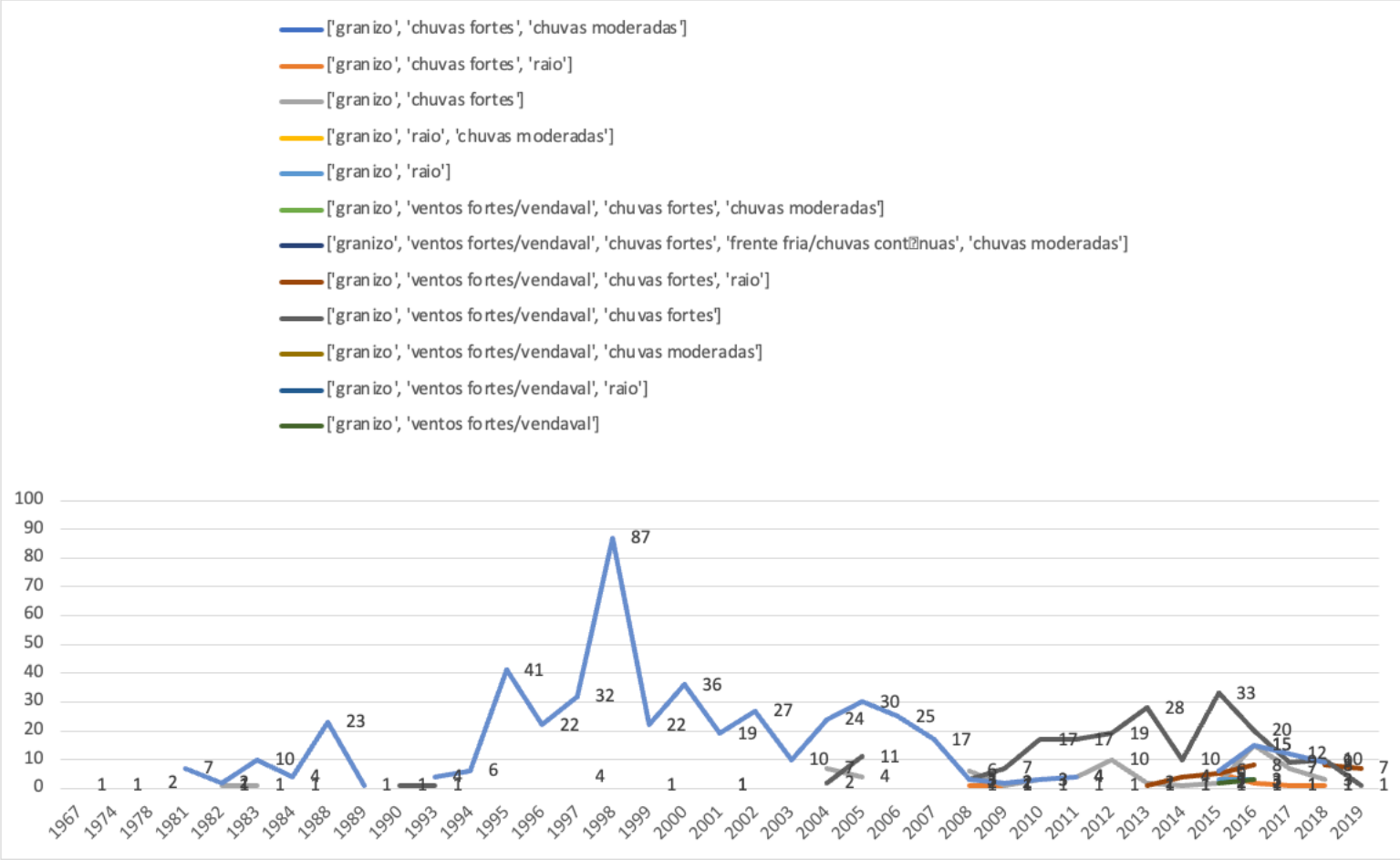
Fonte: Autoria própria (2023)

É importante lembrar que essas associações não implicam necessariamente em uma relação de causalidade entre os fenômenos. No caso do granizo, a associação com chuvas fortes e ventos fortes/vendaval sugere que sua ocorrência pode estar relacionada a situações climáticas intensas.

A Figura 33 e os Quadro 17, Quadro 18 e Quadro 19 apresentam as associação do fenômeno granizo ao longo dos anos e em diferentes combinações de eventos climáticos. Pode-se observar que há uma grande variabilidade na quantidade de ocorrências de granizo. Destaca-se o período entre 1991 e 2010. O CEPED UFSC (2013) destaca que entre este período houve um total de 6.224 construções danificadas e 22 destruídas por granizos somente no estado de São Paulo.

Ao analisar as combinações de eventos climáticos que foram registrados com granizo, pode-se observar que a maioria das ocorrências envolveu chuvas fortes e ventos fortes/vendaval. O raio também foi um evento climático que esteve presente em muitas ocorrências. Pode-se notar ainda que há uma tendência crescente no número de ocorrências de granizo registradas. No entanto, essa tendência não é linear e apresenta variações de ano para ano.

Figura 33 - Evolução anual das ocorrências de granizo e suas associações a partir dos dados do banco do IPMET



Fonte: Autoria própria (2023)

Quadro 17 - Contagem do fenômeno granizo ao longo dos anos e em diferentes combinações de condições climáticas IPMET parte 1

	['granizo', 'chuvas fortes', 'chuvas moderadas']	['granizo', 'chuvas fortes', 'raio']	['granizo', 'chuvas fortes']	['granizo', 'raio', 'chuvas moderadas']
1967				
1974			1	
1978				
1981				
1982			1	
1983			1	
1984				
1988				
1989				
1990				
1993				
1994				
1995				
1996				
1997			4	
1998				
1999				
2000				
2001			1	
2002				
2003				
2004			7	
2005			4	
2006				
2007				
2008		1	6	
2009		1	1	1

2010	1		3	
2011			4	
2012			10	
2013		1	2	
2014	1		1	
2015	2	5	2	
2016		2	15	
2017	1	1	7	
2018	1	1	3	
2019				
Total Geral	**6**	**12**	**73**	**1**

Fonte: Autoria própria (2023)

Quadro 18 - Contagem do fenômeno granizo ao longo dos anos e em diferentes combinações de condições climáticas IPMET parte 2

	['granizo', 'raio']	['granizo', 'ventos fortes/vendaval', 'chuvas fortes', 'chuvas moderadas']	['granizo', 'ventos fortes/vendaval', 'chuvas fortes', 'frente fria/chuvas cont&#65533;nuas', 'chuvas moderadas']	['granizo', 'ventos fortes/vendaval', 'chuvas fortes', 'raio']
1967				
1978				
1981				
1982				
1983				
1984				
1988				
1989				
1990				
1993				
1994				
1995				
1996				
1997				
1998				
1999				
2000				
2001				
2002				
2003				
2004				
2005				
2006				
2007				
2008				

2009		2	1	2
2010				
2011		1		
2012				
2013				1
2014				4
2015	3	1		5
2016	3			8
2017				
2018				8
2019				7
Total Geral	**6**	**4**	**1**	**35**

Fonte: Autoria própria (2023)

Quadro 19 - Contagem do fenômeno granizo ao longo dos anos e em diferentes combinações de condições climáticas IPMET parte 3

	['granizo', 'ventos fortes/vendaval', 'chuvas fortes']	['granizo', 'ventos fortes/vendaval', 'chuvas moderadas']	['granizo', 'ventos fortes/venda-val', 'raio']	['granizo', 'ventos for-tes/vendaval']	['grani-zo']	Total Geral
1967	1					1
1978	2					2
1981					7	7
1982	1				2	3
1983					10	10
1984				1	4	5
1988					23	23
1989					1	1
1990	1					1
1993	1				4	5
1994					6	6
1995					41	41
1996					22	22
1997					32	32
1998					87	87
1999	1				22	23
2000					36	36
2001	1				19	20
2002					27	27
2003					10	10
2004	2				24	26
2005	11				30	41
2006					25	25
2007					17	17
2008	3			2	3	8
2009	7				2	9
2010	17				3	20

2011	17				4	21
2012	19			1		20
2013	28					28
2014	10					10
2015	33			2	6	41
2016	20	1	1	3	15	40
2017	9				12	21
2018	10				9	19
2019	1					1
Total Geral	**195**	**1**	**1**	**9**	**503**	**709**

Fonte: Autoria própria (2023

5.4 Coleção de dados centralizada (CENTER DATA)

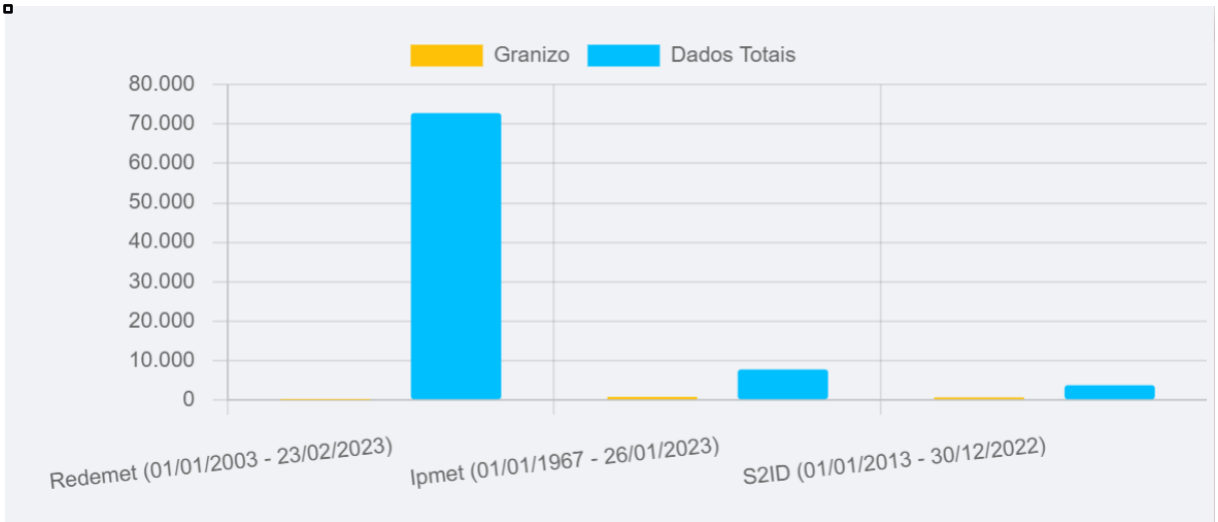
A partir de todas as bases de dados descritas nas subseções anteriores, foi desenvolvida uma base de dados centralizada, com um total de 84298 registros entre os períodos de 01 de janeiro de 1967 a 23 de fevereiro de 2023. Nota-se que a quantidade de registros centralizados, é inferior a quantidade de registros da plataforma da REDEMET, pois antes de centralizar os dados para a coleção CENTER DATA, os registros passam por um filtro que seleciona apenas os fenômenos contidos no filtro regex do Quadro 11, pois a plataforma da REDEMET, possui vários registros de fenômenos que não estão contidos no mencionado filtro, assim sendo, a fim de padronizar os tipos de fenômenos climáticos que serão centralizados das três plataformas, REDEMET, IPMET e S2ID.

Para este período, foram encontrados 1520 registros de granizo. Com base nos dados da Figura 34, observa-se que a incidência de granizo é relativamente baixa em comparação ao número total de registros. Na REDEMET, por exemplo, apenas 41 dos 72755 registros (ou cerca de 0,06%) contêm relatos de granizo, é importante ressaltar que esses 72755 registros da REDEMET, são os registros que passaram pelo filtro de eventos (**Erro! Fonte de referência não encontrada.**), em resumo a REDEMET possui em sua coleção original (coleção REDEMET) cerca de 320.622 registros, desses apenas 72755 passaram pelo filtro. No entanto, na S2ID, observa-se uma proporção muito maior, com 680 dos 3776 registros (ou cerca de 18%) contendo granizo.

Em relação ao IPMET, a proporção de registros contendo granizo é de aproximadamente 10%, com 799 dos 7767 registros apresentando relatos desse fenômeno meteorológico. Essas diferenças entre as bases de dados podem ser explicadas por uma variedade de fatores, como diferenças geográficas, no período, nas metodologias de coleta e processamento de dados entre as diferentes redes. No entanto, mesmo com as diferenças nas proporções.

A Figura 34 apresenta a proporção de dados entre as três coleções de dados.

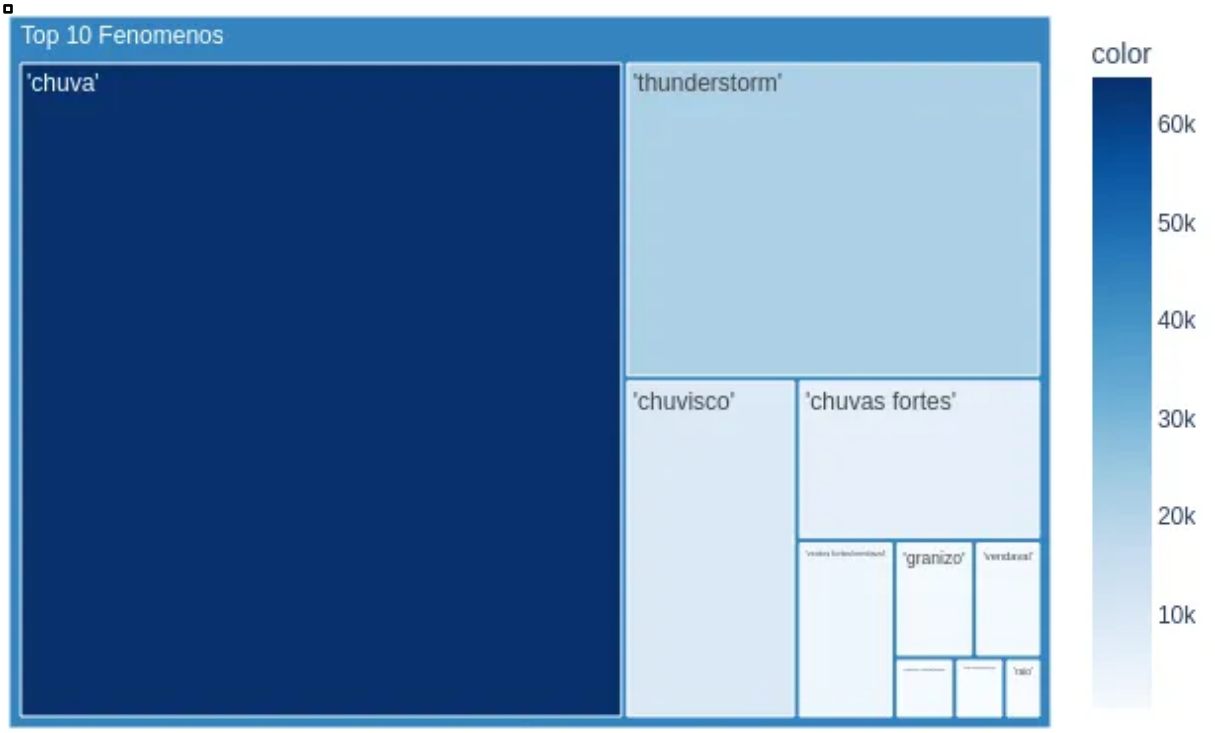
Figura 34 - Dados da coleção Center Data, Totais x Granizo



Fonte: Autoria própria (2023)

A Figura 35 apresenta um *Tree Map* contendo todos os eventos da coleção CENTER DATA. Um *Tree Map* é um gráfico que representa a proporção de cada evento, onde a área e cor de cada evento representam sua proporção, quanto maior área e mais escura a cor, maior é a contagem de incidências registradas. Assim como nas coleções de dados individuais, o evento atmosférico ‘'chuva’ é o evento mais frequente durante o período, com 64905 ocorrências. Já o granizo é o 7º evento mais frequente, logo após vento/vendaval.

Figura 35 - Tree Map top fenômenos da coleção CENTER DATA



Fonte: Autoria própria (2023)Tabela 8 apresenta informações em destaque dos atributos extraídos de ocorrência de fenômenos meteorológicos da coleção de dados centralizada CENTER DATA. Os significados de cada atributo da Tabela são:

- Data: Coluna que contém a data em que o evento ocorreu. Existem 7538 datas únicas, com algumas datas ocorrendo mais vezes que outras.
- Uf: Coluna que contém o estado onde o evento ocorreu. Existem 4 estados diferentes, com São Paulo sendo o mais frequente, aparecendo em 40.519 registros.
- Cidade: Colunas que contêm a cidade onde o evento ocorreu. Existem 2000 cidades diferentes, com São Paulo novamente sendo a mais frequente, aparecendo em 6342 registros.
- Latitude e Longitude: Colunas que contêm as coordenadas geográficas do local do evento. Existem 2621 valores únicos na coluna de latitude e 2700 valores únicos na coluna de longitude. Há muitos registros com valor -1 nessas colunas, o que pode indicar informações ausentes ou imprecisas.

- Fenômeno: Coluna que contém o tipo de evento climático registrado. Existem 56 valores únicos, com "chuva" sendo o mais frequente, aparecendo em 43.436 registros.

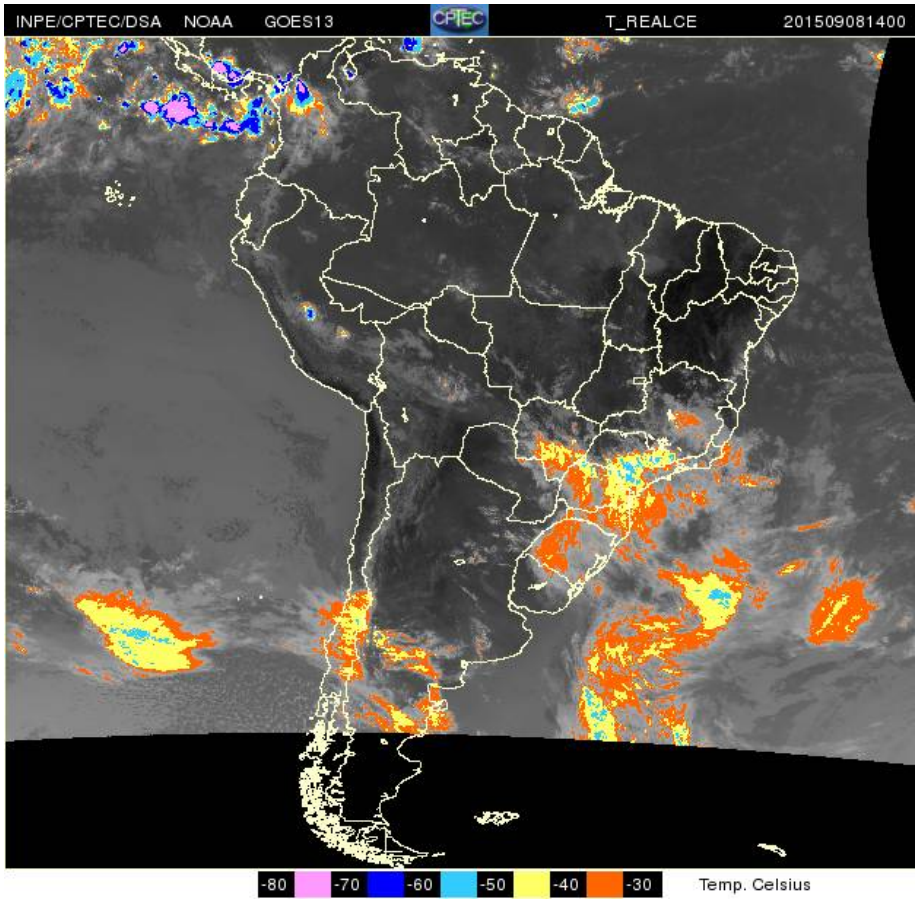
Tabela 8 - Descrição dos dados da coleção CENTER DATA

Atributo	Total	Dados únicos	Maior Frequência	Número de ocorrências
Data	84363	7538	2015-09-08	95
Uf	84363	4	sp	40519
Cidade	84361	2000	são paulo	6342
Latitude	80597	2621	-1	4390
Longitude	80597	2700	-1	4390
Fenomeno	84363	56	['chuva']	43436

Fonte: Autoria própria (2023)

Neste caso, a data mais frequente foi o dia 08 de setembro de 2015, com 95 ocorrências. Neste dia, uma linha de instabilidade gerou muita precipitação, causando várias ocorrências (Figura 36) com temporais isolados em toda a região Sul do Brasil, e tempestades severas no estado de São Paulo. Só neste dia, na capital paulista, choveu 80 mm, sendo a normal no mês é de 71 mm, com fortes ventos em toda a cidade, e granizo no bairro do Paraíso (PEGORIM, 2015).

Figura 36 - Imagem de satélite de Temperatura Realçada do GOES-13 das 14:00 UTC do dia 08 de setembro de 2015.



Fonte: CPTEC-INPE (http://satelite.cptec.inpe.br/acervo/)

A Tabela 9 apresenta o resultado da do processamento dos dados contidos na coleção CENTER DATA, apresentando a frequência de ocorrência de diferentes associações entre eventos climáticos e a Tabela 10 as ocorrências dos fenômenos individualmente, obtidos com os algoritmos fpgrowth() e association_rules(), atribuindo um suporte mínimo de 0,1%. A análise da tabela mostra que a ocorrência de chuva é o evento mais comum, com uma frequência de cerca de 77% das observações. O segundo evento mais comum é tempestade com trovões, com uma frequência aproximadamente de 26% das observações. A associação entre chuva e tempestade tem uma frequência 25,5%, o que indica que esses eventos geralmente ocorrem juntos. Outras associações significativas incluem ventos fortes/vendaval com chuvas fortes e/ou granizo, que aparecem em 2,8% e 2,4% das ocorrências, respectivamente. Outras associações com frequência acima de 0,1% incluem chuvisco, chuvas fortes, ventos fortes/vendaval, granizo, vendaval, chuvas moderadas e frente fria/chuvas contínuas. Algumas associações entre eventos climáticos também são observadas, como chuvas fortes e raio, chuvas fortes e

granizo, ventos fortes/vendaval e raio, e ventos fortes/vendaval, chuvas fortes e granizo. Essas associações indicam que esses eventos podem ocorrer simultaneamente numa frequência bem mais baixa.

Tabela 9 - Frequência das associações dos fenômenos da coleção CENTER DATA

Frequência das associações fenômenos e min_supporte: 0,1%	
Fenômenos	**Ocorrências em %**
chuva | tempestade	25,51%
ventos fortes/vendaval | chuvas fortes	2,85%
chuvas fortes | raio	0,42%
chuvas fortes | chuvas moderadas	0,42%
ventos fortes/vendaval | raio	0,36%
ventos fortes/vendaval | chuvas fortes | raio	0,36%
chuvas fortes | granizo	0,34%
chuvas fortes | frente fria/chuvas contínuas	0,29%
ventos fortes/vendaval | granizo	0,26%
ventos fortes/vendaval | chuvas fortes | granizo	0,24%
chuvas moderadas | frente fria/chuvas contínuas	0,13%
ventos fortes/vendaval | chuvas moderadas	0,12%

Fonte: Autoria própria (2023)

Tabela 10 - Frequência dos fenômenos individuais da coleção CENTER DATA

Frequência de fenômenos individuais e min_supporte: 0,1%	
Fenômenos	**Ocorrências em %**
chuva	76,99%
tempestade	25,59%
chuvisco	11,32%
chuvas fortes	7,65%
granizo	1,80%
vendaval	1,53%
chuvas moderadas	0,71%
frente fria/chuvas contínuas	0,59%
raio	0,45%

Fonte: Autoria própria (2023)

A Tabela 11 mostra o suporte, ou seja, a frequência relativa de ocorrência dos eventos antecedentes e consequentes, bem como a associação entre eles medida

pela confiança, considerando um suporte mínimo de 0,1%. O suporte indica a frequência com que uma associação ocorre na coleção de dados, enquanto a confiança indica a probabilidade de que a consequência ocorra dado que o antecedente ocorreu. Na primeira coluna destacam-se os antecedentes, ou seja, os eventos que antecedem a ocorrência de outro evento. Na segunda coluna se tem os consequentes, que são os fenômenos que ocorrem em consequência dos antecedentes. As colunas "Antecedente Suporte" e "Consequente Suporte" indicam a frequência de ocorrência dos eventos antecedentes e consequentes, respectivamente. Já a coluna "Suporte da associação" indica a frequência com que os antecedentes e consequentes ocorrem juntos, ou seja, a frequência da associação entre os eventos.

Tabela 11 – Suporte dos Fenômenos Consequentes e Antecedentes min_support: 0,1%

Associações dos fenômenos Consequentes e Antecedentes				
Antecedentes	**Consequentes**	**Antecedente Suporte**	**Consequente Suporte**	**Suporte da associação**
raio	ventos fortes/vendaval	0,45%	3,36%	0,36%
chuvas fortes | raio	ventos fortes/vendaval	0,42%	3,36%	0,36%
ventos fortes/vendaval	chuvas fortes	3,36%	7,64%	2,84%
ventos fortes/vendaval | granizo	chuvas fortes	**0,25%**	**7,64%**	**0,24%**
raio	chuvas fortes	0,45%	7,64%	0,42%
ventos fortes/vendaval | raio	chuvas fortes	0,36%	7,64%	0,36%
tempestade	chuva	25,60%	77,01%	25,52%

Fonte: Autoria própria (2023)

Observa-se que a associação entre chuvas fortes e ventos fortes/vendaval é bastante comum, com um suporte de associação de 2,84%, o que indica que esses eventos ocorrem juntos em cerca de 2,8% das ocorrências, totais. Outra associação comum é entre raio e ventos fortes/vendaval, com um suporte de associação de 0,36%, indicando que esses eventos ocorrem juntos em cerca de 0,36% das ocorrências totais.

É importante notar que a associação entre ventos fortes/vendaval e chuvas fortes é bidirecional, ou seja, ambos os eventos podem ser tanto antecedentes quanto consequentes em relação ao outro. Além disso, a associação entre tempestades e chuva tem o maior suporte entre todas as associações apresentadas na tabela, indicando uma relação forte entre eles.

O suporte refere-se à frequência com que uma associação específica ocorre em relação ao total de registros da base de dados. A confiança, por sua vez, representa a proporção de vezes que o fenômeno consequente ocorre quando o fenômeno antecedente também acontece.

Em geral, podemos observar que as associações apresentam valores de suporte e *confidence* relativamente baixos, o que sugere que essas associações não ocorrem com frequência na base de dados centralizada, ou seja, as combinações de eventos específicos não são tão comuns ou frequentes ao longo do período analisado. Apesar dos valores baixos de suporte e confiança, os dados obtidos ainda têm sua utilidade. Eles podem auxiliar na identificação de possíveis correlações entre os fenômenos atmosféricos e na orientação da análise de riscos associados a eventos climáticos extremos.

Com relação ao granizo, dentre essas associações, podemos observar que a sua associação com chuvas fortes apresenta um suporte de ocorrência relativamente baixo (0,24%), quando comparada a outras associações, como a associação entre tempestade e chuva (25,52%). Além disso, é possível notar que o granizo está associado principalmente aos ventos fortes/vendavais como antecedente, e às chuvas fortes como consequente.

Essa associação entre granizo e (ventos fortes/vendavais) apresenta um suporte de ocorrência ainda mais baixo do que a associação entre granizo e chuvas fortes (0,25%). Por outro lado, as associações entre raio com (ventos fortes/vendavais) e entre raio com chuvas fortes, apresentam suportes de ocorrência mais altos do que a associação entre granizo com chuvas fortes. Assim, pode-se dizer que o granizo apresenta uma associação relativamente fraca com chuvas fortes quando comparado a outros fenômenos, como raio e ventos fortes/vendavais.

Já a Tabela 12 apresenta as medidas de *Confidence, Lift, Leverage* e *Conviction* dos eventos climáticos consequentes e antecedentes, com um suporte mínimo de 0,1.%. Essas metricas são utilizadas em análises de associação entre variáveis e fornecem informações sobre o grau de dependência entre os eventos. O *confidence* representa a proporção de transações que contêm tanto o antecedente quanto o consequente em relação ao número de transações que contêm apenas o antecedente e variam de 0 a 1. Nesta tabela, é possível observar que as associações com maior *confidence* são: "raio -> ventos fortes/vendaval", com 81,65% de *confidence*, e "chuvas fortes | raio -> ventos fortes/vendaval", com 83,80% de

confidence. O *lift* mede a força da relação entre o antecedente e o consequente em relação à ocorrência do consequente na população geral. Valores acima de 1 indicam que a associação é mais forte do que o esperado ao acaso. Na Tabela 12, as associações com maior *lift* são: "ventos fortes/vendaval | granizo -> chuvas fortes", com um valor de 12,42, e "ventos fortes/vendaval | raio -> chuvas fortes", com um valor de 12,79. Já o *leverage* mede a diferença entre a frequência observada da ocorrência conjunta do antecedente e do consequente e a frequência esperada se eles fossem independentes, com a associação com maior *leverage* é "raio -> chuvas fortes", com um valor de 0,00390. Valores positivos indicam que a associação é mais frequente do que o esperado ao acaso. Por fim, o *conviction* mede a dependência do consequente em relação ao antecedente, levando em consideração a proporção de transações que não contêm o consequente. Quanto maior o valor de *conviction*, mais forte é a dependência entre as variáveis. Na Tabela 12, a associação com maior *conviction* é "ventos fortes/vendaval | raio -> chuvas fortes", com um valor de 40,51.

Tabela 12 – *Confidence, Lift, Leverage* e *Conviction* dos Fenômenos Consequentes e Antecedentes baseados em *min_support* 0,1%

Associações dos fenômenos Consequentes e Antecedentes					
Antecedentes	**Consequentes**	**Confi-dence**	**Lift**	**Leve-rage**	**Convic-tion**
raio	ventos fortes/vendaval	0,81648	24,32256	0,00348	5,26634
chuvas fortes | raio	ventos fortes/vendaval	0,83798	24,96301	0,00341	5,96521
ventos fortes/vendaval	chuvas fortes	0,84745	11,09295	0,02588	6,05473
ventos fortes/vendaval | granizo	chuvas fortes	0,94883	12,41997	0,00222	18,05225
raio	chuvas fortes	0,95212	12,46304	0,00390	19,29306
ventos fortes/vendaval | raio	chuvas fortes	0,97719	12,79122	0,00327	40,50663
tempestade	chuva	0,99694	1,29458	0,05806	75,22276

Fonte: Autoria própria (2023)

Com relação ao granizo, observa-se na Tabela 12 que as associações entre granizo e chuvas fortes, assim como entre raio e chuvas fortes, apresentam altos valores de *Confidence* (0,94883 e 0,95212, respectivamente), o que indica que a ocorrência de granizo ou raio é altamente dependente da presença de chuvas fortes. Além disso, a associação entre granizo e chuvas fortes apresenta um alto valor de *Lift* (12,41997), o que indica que a ocorrência de granizo aumenta consideravelmente a probabilidade de ocorrer chuvas fortes. O valor de *Leverage* (0,00222) indica que essa

associação é baixa em relação à ocorrência geral de chuvas fortes, mas ainda assim é significativa.

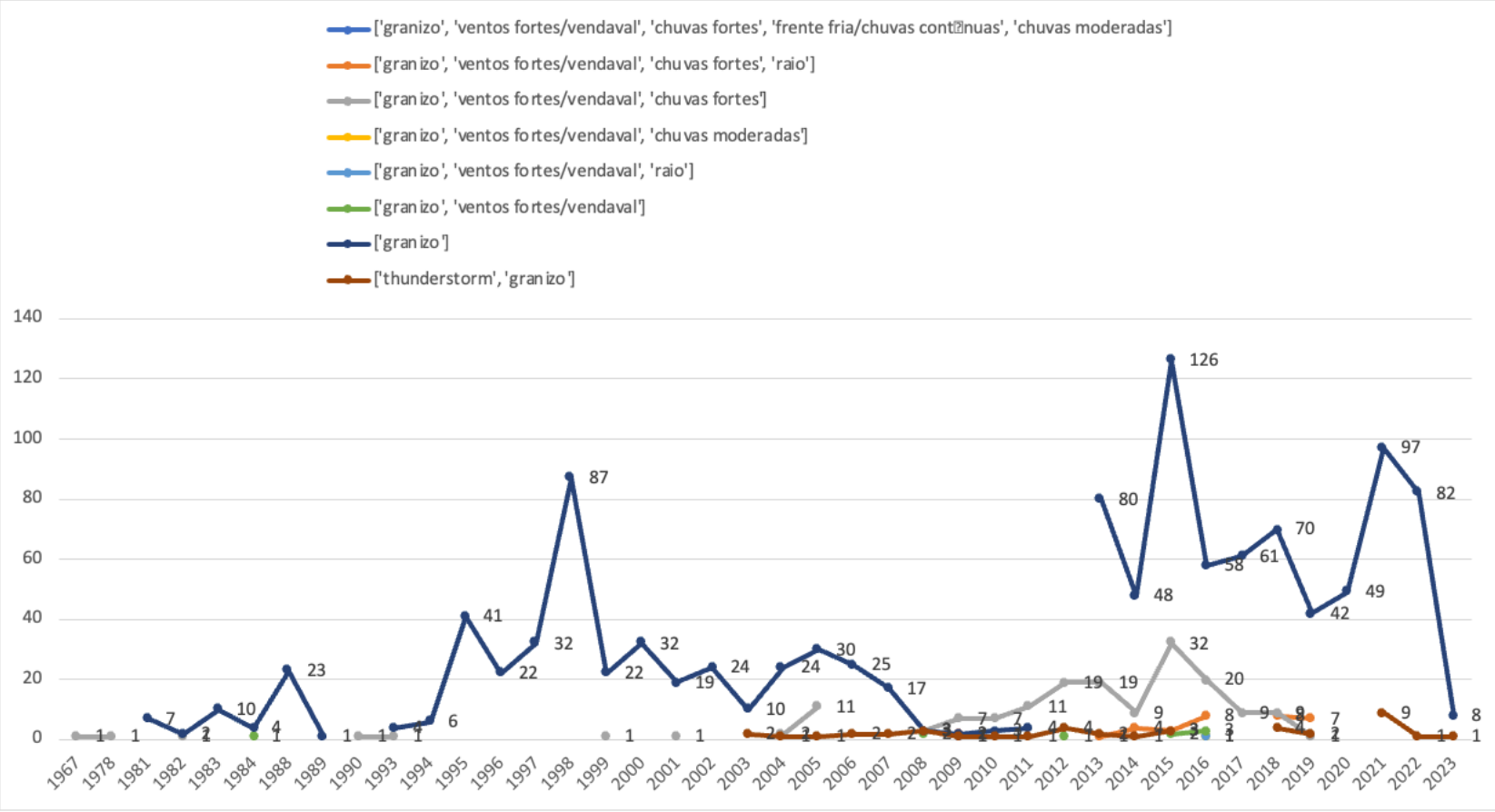
Em relação a associação "ventos fortes/vendaval | granizo" antecede "chuvas fortes" indica que quando ocorrem os fenômenos "ventos fortes/vendaval" e "granizo" juntos, há uma probabilidade de 94,88% de ocorrer também o fenômeno "chuvas fortes". O valor do Lift (12,41997) indica que essa associação é relevante, sendo mais de 12 vezes mais provável de ocorrer em relação à ocorrência geral do fenômeno "chuvas fortes". O Leverage (0,00222) mostra que essa associação é rara em relação à ocorrência geral de "chuvas fortes", enquanto o Conviction (18,05225) revela que a ocorrência de "chuvas fortes" é mais de 18 vezes mais provável dado o "ventos fortes/vendaval" e "granizo" juntos.

Por outro lado, a associação entre tempestade e chuva apresenta o maior valor de *Confidence* (0,99694) e *Conviction* (75,22276), indicando que a presença de tempestade aumenta significativamente a probabilidade de ocorrer chuva. No entanto, o valor de *Lift* (1,29458) sugere que a relação entre esses eventos é menos forte do que as associações envolvendo granizo e raio.

Analisando evolução anual dos eventos de granizo associados apresentados na Figura 37, Quadro 20 e Quadro 21, observa-se que a combinação de granizo, chuvas fortes e ventos fortes/vendaval é a mais comum, com um total de 166 ocorrências em todos os anos listados. A combinação de granizo e chuvas moderadas é a segunda mais comum, com um total de 33 ocorrências. Nota-se também que, nos últimos anos, houve um aumento significativo no número de ocorrências de granizo com outras condições meteorológicas, como chuvas fortes e trovões.

Em 2021, por exemplo, houve 106 ocorrências de granizo em combinação com trovões, tornando-se a combinação mais frequente neste ano. Assim compreende-se que, analisando todos os dados disponíveis para o estado de São Paulo e a região Sul do Brasil, o granizo é um evento meteorológico relativamente comum quando combinado com outros eventos, especialmente com chuvas fortes e ventos fortes/vendaval, e que o número de ocorrências pode ter aumentado nos anos mais recentes, possivelmente devido às mudanças climáticas ou devido a melhores técnicas de captação de dados e equipamentos de coleta, assim como observado em diversas partes do mundo (MAHONEY et al., 2012; PREIN; HOLLAND, 2018c; ALLEN, 2018; RÄDLER et al., 2019; RAUPACH et al., 2021).

Figura 37 - Evolução anual das ocorrências de granizo e suas associações a partir dos dados do banco do CENTER DATA



Fonte: Autoria própria (2023)

Quadro 20 - Contagem do fenômeno granizo ao longo dos anos e em diferentes combinações de condições climáticas CENTER DATA parte 1

	['granizo', 'chuvas fortes', 'chuvas moderadas']	['granizo', 'chuvas fortes', 'raio']	['granizo', 'chuvas fortes']	['granizo', 'raio', 'chuvas moderadas']	['granizo', 'raio']	['granizo', 'ventos fortes/vendaval', 'chuvas fortes', 'chuvas moderadas']
1967						
1974			1			
1978						
1981						
1982			1			
1983			1			
1984						
1988						
1989						
1990						
1993						
1994						
1995						
1996						
1997			1			
1998						
1999						
2000						
2001			1			
2002						
2003						
2004			7			
2005			4			
2006						
2007						
2008		1	6			
2009		1	1	1		2

2010	1		3			
2011			4			1
2012			4			
2013		1	2			
2014	1		1			
2015	2	5	2		3	1
2016		2	15		3	
2017	1	1	7			
2018	1	1	3			
2019						
2020						
2021						
2022						
2023						
Total Geral	**6**	**12**	**64**	**1**	**6**	**4**

Fonte: Autoria própria (2023)

Quadro 21 - Contagem do fenômeno granizo ao longo dos anos e em diferentes combinações de condições climáticas CENTER DATA parte 2

	['granizo', 'ventos fortes/vendaval', 'chuvas fortes', 'frente fria/chuvas cont&#65533;nuas', 'chuvas moderadas']	['granizo', 'ventos fortes/vendaval', 'chuvas fortes', 'raio']	['granizo', 'ventos fortes/vendaval', 'chuvas fortes']	['granizo', 'ventos fortes/vendaval', 'chuvas moderadas']	['granizo', 'ventos fortes/vendaval', 'raio']	['granizo', 'ventos fortes/vendaval']	['granizo']	['thunderstorm', 'granizo']	Total Geral
1967			1						1
1978			1						1
1981							7		7
1982			1				2		3
1983							10		10
1984						1	4		5
1988							23		23
1989							1		1
1990			1						1
1993			1				4		5
1994							6		6
1995							41		41
1996							22		22
1997							32		32
1998							87		87
1999			1				22		23
2000							32		32
2001			1				19		20
2002							24		24
2003							10	2	12
2004			2				24	1	27
2005			11				30	1	42
2006							25	2	27
2007							17	2	19
2008			3			2	3	3	11

2009	1	2	7				2	1	13
2010			7				3	1	11
2011			11				4	1	16
2012			19			1		4	24
2013		1	19				80	2	102
2014		4	9				48	1	62
2015		3	32			2	126	3	166
2016		8	20	1	1	3	58		91
2017			9				61		70
2018		8	9				70	4	91
2019		7	1				42	2	52
2020							49		49
2021							97	9	106
2022							82	1	83
2023							8	1	9
Total Geral	**1**	**33**	**166**	**1**	**1**	**9**	**1175**	**41**	**1427**

Fonte: Autoria própria (2023)

5.5 Dashboard

No Apêndice C encontram-se as imagens das telas disponíveis do sistema completo para consulta. Atualmente o sistema conta com 5 seções disponíveis:

- Dados Processados: referente ao conjunto completo através da coleção de dados centralizada.
- Seções com coleções de dados individuais (REDEMET, S2ID e IPMET): dados e informações de cada base de dados analisada individualmente.
- Configurações do FP-growth: permite ao usuário selecionar algumas variáveis com relação ao algoritmo fp-growth, como o índice confiança, suporte e métricas.

Na tela principal é apresentado todas as informações referentes a cada coleção de dados utilizados neste trabalho, além de um mapa com pontos de ocorrência de granizo, no qual ao clicado no determinado ponto, mais informações sobre esta ocorrência é apresentada na tela. Atualmente o sistema se encontra em um servidor cloud da Amazon (AWS) instância t3g.micro, com 2 vCPU e 2 GiB de RAM.

6 CONCLUSÃO

As tempestades de granizo se destacam como um evento destrutivo e danoso em várias regiões do planeta, incluindo a região Sul do Brasil e o Sudeste da América do Sul. Esses eventos podem causar impactos sociais, ecológicos e danos em estruturas tecnológicas, além de perdas e prejuízos em culturas agrícolas e áreas urbanas.

A previsão meteorológica pode auxiliar as pessoas e organismos a se prepararem para possíveis emergências, mas a complexidade em detectar tempestades de granizo ainda é um desafio. A criação de modelos estatísticos robustos é dificultada pela falta dados, pois são extremamente difíceis de se obter, principalmente em regiões de difícil acesso. Portanto, é necessário estudar as tempestades de granizo para possibilitar a transmissão de informações com antecedência à população para que tomem as medidas necessárias e reduzam os impactos.

Um dos desafios do presente estudo foi a criação dos algoritmos para coletar os dados de três plataformas distintas, pois cada plataforma registra os dados de acordo com o seu padrão. Após a coleta foi realizado um trabalho de limpeza, padronização e preparação dos dados. Isso é necessário para utilizá-los como dados de entrada para os algoritmos de aprendizagem de máquina ou algoritmos de extração de padrões.

Os resultados mostraram que a centralização dos dados é fundamental para avaliar a ocorrência de granizo associado com outros eventos atmosféricos. Quando analisado cada banco de dado individualmente, o granizo tinha pouca ou nenhuma ocorrência com outros fenômenos como tempestades e ventos fortes. Entretanto, a partir da centralização de todos os bancos de dados, foi viabilizada a visualização de algumas associações. Isso demonstra a necessidade de centralização dos dados para a aplicação de qualquer modelo de inteligência artificial que auxilie na previsão de tempestades de granizo. As associações foram observadas apenas quando as bases de dados estavam centralizadas, não sendo possível observar associações significativas com as bases individuais.

Para trabalhos futuros sugere-se aumentar o banco de dados com outras informações, como por exemplo dados não estruturados (dados jurídicos, redes sociais). Além disso, faz-se necessário a inclusão de dados de outros estados como por exemplo, Minas Gerais (MG), onde também há um elevando índice de ocorrência

de granizo. A partir disso pode-se testar modelos de aprendizagem de máquina.

REFERÊNCIAS

ALLEN, J. T. Climate Change and Severe Thunderstorms. Em: **Oxford Research Encyclopedia of Climate Science**. [s.l.] Oxford University Press, 2018.

ALLEN, J. T. et al. Understanding Hail in the Earth System. **Reviews of Geophysics**, v. 58, n. 1, 2020.

ALLEN, J. T.; TIPPETT, M. K.; SOBEL, A. H. An empirical model relating U.S. monthly hail occurrence to large-scale meteorological environment. **Journal of Advances in Modeling Earth Systems**, v. 7, n. 1, p. 226–243, 2015.

ALSHAMMARI, R. et al. Data Extraction Based on Web Scrapy. **Advances in Intelligent Systems and Computing**, v. 1372 AISC, p. 506–514, 2021.

BAGUI, S.; DEVULAPALLI, K.; COFFEY, J. A heuristic approach for load balancing the FP-growth algorithm on MapReduce. **Array**, v. 7, p. 100035, 1 set. 2020.

BATSAKIS, S.; PETRAKIS, E. G. M.; MILIOS, E. Improving the performance of focused web crawlers. **Data and Knowledge Engineering**, v. 68, n. 10, p. 1001–1013, 2009.

BEAL, A. et al. Climatology of hail in the triple border Paraná, Santa Catarina (Brazil) and Argentina. **Atmospheric Research**, v. 234, p. 104747, 1 abr. 2020a.

BEAL, A. et al. Climatology of hail in the triple border Paraná, Santa Catarina (Brazil) and Argentina. **Atmospheric Research**, v. 234, p. 104747, 1 abr. 2020b.

BEAL, A. et al. Evaluation of the chemical composition of hailstones from triple border Paraná, Santa Catarina (Brazil) and Argentina. **Atmospheric Pollution Research**, v. 12, n. 3, p. 184–192, 2021a.

BEAL, A. et al. Evaluation of the chemical composition of hailstones from triple border Paraná, Santa Catarina (Brazil) and Argentina. **Atmospheric Pollution Research**, v. 12, n. 3, p. 184–192, 1 mar. 2021b.

BIAN, Y. et al. Hail climatology and its possible attributions in Beijing, China: 1980-2021. **Frontiers in Environmental Science**, v. 10, 4 jan. 2023.

BORGELT, C. An implementation of the FP-growth algorithm. **Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, p. 1–5, 2005.

BOTZEN, W. J. W.; BOUWER, L. M.; VAN DEN BERGH, J. C. J. M. Climate change and hailstorm damage: Empirical evidence and implications for agriculture and insurance. **Resource and Energy Economics**, v. 32, n. 3, p. 341–362, 1 ago. 2010.

BREIMAN, L. Random Forests. **Machine Learning 2001 45:1**, v. 45, n. 1, p. 5–32, out. 2001.

CALDANA, N. F. DA S.; NITSCHE, P. R.; CARAMORI, P. H. Precipitações de granizo e os impactos na Mesorregião Sudoeste Paranaense, Brasil (Hail precipitations and impacts in the Southwest Mesoregion Of Parana State, Brazil). **Revista Brasileira de Geografia Física**, v. 12, n. 4, p. 1327–1339, 15 out. 2019.

CEPED UFSC. ATLAS BRASILEIRO DE DESASTRES NATURAIS1991 A 2012. v. Volume Amazonas, 2013.

CHAULAGAIN, R. S. et al. Cloud Based Web Scraping for Big Data Applications. **Proceedings - 2nd IEEE International Conference on Smart Cloud, SmartCloud 2017**, p. 138–143, 22 nov. 2017.

DE BRITO, A. P. et al. Analysis of the Rain Anomaly Index and Precipitation Trend for Pluviometric Stations in Central Amazonia. **Revista Brasileira de Meteorologia**, v. 37, n. 1, p. 19–30, 2022.

DE CARVALHO, A. L. et al. Impacts of extreme climate events on Brazilian agricultural production. **Sustentabilidade em Debate**, v. 11, n. 3, p. 197–210, 2020.

DEWI, L. C.; MEILIANA; CHANDRA, A. Social Media Web Scraping using Social Media Developers API and Regex. **Procedia Computer Science**, v. 157, p. 444–449, 1 jan. 2019.

DIELING, C.; SMITH, M. L.; BERUVIDES, M. Inaccuracy of the energy-matching calibration technique for hailpads. **Results in Engineering**, v. 12, p. 100277, 1 dez. 2021.

FATMASARI; KUNANG, Y. N.; PURNAMASARI, S. D. Web Scraping Techniques to Collect Weather Data in South Sumatera. **Proceedings of 2018 International Conference on Electrical Engineering and Computer Science, ICECOS 2018**, p. 385–390, 7 jan. 2019.

FELIPE, N. et al. Gênese, Impacto e a Variabilidade das Precipitações de Granizo na Mesorregião Centro-Sul Paranaense, Brasil / Genesis, Impact and Variability of Hail Precipitations in the Central South Mesoregion of the State of Paraná, Brazil. **Caderno de Geografia**, v. 29, n. 56, p. 61–61, 20 fev. 2019.

GALAZ, V. et al. **Can web crawlers revolutionize ecological monitoring? Frontiers in Ecology and the Environment**John Wiley & Sons, Ltd, , 1 mar. 2010. Disponível em: <http://wiki.resalliance.org>. Acesso em: 16 maio. 2021

GALLO, K. et al. Validation of Satellite Observations of Storm Damage to Cropland with Digital Photographs. **Weather and Forecasting**, v. 34, n. 2, p. 435–446, 1 abr. 2019.

GIAMBASTIANI, Y. et al. Web scraping technology for a dynamics analysis of tree crown streamlining, in relationships with wind and meteorological data. **2022 IEEE Workshop on Complexity in Engineering, COMPENG 2022**, 2022.

GRIESER, J.; HILL, M. How to Express Hail Intensity—Modeling the Hailstone Size Distribution. **Journal of Applied Meteorology and Climatology**, v. 58, n. 10, p.

2329–2345, 1 out. 2019.

HAN, J.; PEI, J.; YIN, Y. Mining Frequent Patterns without Candidate Generation. **SIGMOD 2000 - Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data**, p. 1–12, 2000.

HEGLAND, M. The apriori algorithm - A tutorial. **Mathematics And Computation In Imaging Science And Information Processing**, p. 209–262, 1 jan. 2007.

HERMIDA, L. et al. Climatic trends in hail precipitation in France: Spatial, altitudinal, and temporal variability. **The Scientific World Journal**, v. 2013, 2013.

HOHL, R.; SCHIESSER, H. H.; KNEPPER, I. The use of weather radars to estimate hail damage to automobiles: an exploratory study in Switzerland. **Atmospheric Research**, v. 61, n. 3, p. 215–238, 1 mar. 2002.

HSU, K.; GUPTA, H. V.; SOROOSHIAN, S. Artificial Neural Network Modeling of the Rainfall-Runoff Process. **Water Resources Research**, v. 31, n. 10, p. 2517–2530, 1 out. 1995.

HU, H.; GE, Y.; HOU, D. Using web crawler technology for geo-events analysis: A case study of the Huangyan Island incident. **Sustainability (Switzerland)**, v. 6, n. 4, p. 1896–1912, 9 abr. 2014.

INMET. **Balanço: São Paulo (SP) registrou, em outubro de 2022, a maior temperatura do ano até então: 34,0°C**. Disponível em: <https://portal.inmet.gov.br/noticias/balan%C3%A7o-s%C3%A3o-paulo-sp-teve-chuvas-dentro-da-m%C3%A9dia-e-temperaturas-acima-em-outubro-de-2022>. Acesso em: 12 abr. 2023.

JORNAL CIDADES. **Santa Cruz do Sul registra novo recorde chuva para um dia de julho**. Disponível em: <https://www.jornaldocomercio.com/_conteudo/jornal_cidades/2020/07/746809-santa-cruz-do-sul-registra-maior-chuva-em-um-dia-de-julho-desde-1914.html>. Acesso em: 12 abr. 2023.

KALAIVANI, G.; KAMALAKKANNAN, S. Web Scraping Technique for Prediction of Air Quality through Comparative Analysis of Machine Learning and Deep Learning Algorithm. p. 263–273, 16 jan. 2023.

KHDER, M. A. Web scraping or web crawling: State of art, techniques, approaches and application. **International Journal of Advances in Soft Computing and its Applications**, v. 13, n. 3, p. 144–168, 2021.

KIM, J. et al. Hail Climatology in the Mediterranean Basin Using the GPM Constellation (1999–2021). **Remote Sensing 2022, Vol. 14, Page 4320**, v. 14, n. 17, p. 4320, 1 set. 2022.

KOUKARAS, P.; TJORTJIS, C.; ROUSIDIS, D. Mining association rules from COVID-19 related twitter data to discover word patterns, topics and inferences. **Information Systems**, v. 109, p. 102054, 1 nov. 2022.

KUMAR, S.; ROY, U. B. A technique of data collection: web scraping with python. **Statistical Modeling in Machine Learning**, p. 23–36, 1 jan. 2023.

LI, G. Y.; CAO, D. Y.; GUO, J. W. Association Rules Mining with Multiple Constraints. **Procedia Engineering**, v. 15, p. 1678–1683, 1 jan. 2011.

MA, J. et al. Uptake and mobilization of organic chemicals with clouds: Evidence from a hail sample. **Environmental Science and Technology**, v. 47, n. 17, p. 9715–9721, 2 set. 2013.

MAHONEY, K. et al. Changes in hail and flood risk in high-resolution simulations over Colorado's mountains. **Nature Climate Change 2012 2:2**, v. 2, n. 2, p. 125–131, 10 jan. 2012.

MAIDMENT, DAVID R, C. V. T.; MAYS, L. W. Applied Hidrology. [s.d.].

MANFIO, Edio Roberto; MORAIS, Marcos Vinicius Bueno de; MORENO, Fabio Carlos; BARBOSA, Cinthyan Renata Sachs Camerlengo de; GUERRA, Marcos Paulo Guimarães. Monitoring System for Agrometeorological Application with Voice-Controlled Interface. International Journal of Advanced Engineering Research and Science, v.5, p.12-17, 2018.

MANFIO, E. R.; GUERRA, M. P. G.; MANCUZO, E. Inversor de baixo custo controlado por voz integrado a sistema fotovoltaico automatizado. Revista e-f@tec., v.12, 2022. Disponível em: <https://pesquisafatec.com.br/ojs/index.php/efatec/article/view/282/201>

MANFIO, E. R.; GUERRA, M. P. G.; MORENO, F. C.; MORAIS, M. V. B. IHCs dedicadas a energias sustentáveis. Revista e-f@tec., v.6, p.7 -, 2016. Disponível em: <https://pesquisafatec.com.br/ojs/index.php/efatec/article/view/38/36>

MANJARI, K. U. et al. Extractive Text Summarization from Web pages using Selenium and TF-IDF algorithm. **Proceedings of the 4th International Conference on Trends in Electronics and Informatics, ICOEI 2020**, p. 648–652, 1 jun. 2020.

MANZATO, A. et al. Hailstone Characteristics in Northeast Italy from 29 Years of Hailpad Data. **Journal of Applied Meteorology and Climatology**, v. 61, n. 11, p. 1779–1795, 4 nov. 2022a.

MANZATO, A. et al. Hailstone Characteristics in Northeast Italy from 29 Years of Hailpad Data. **Journal of Applied Meteorology and Climatology**, v. 61, n. 11, p. 1779–1795, 4 nov. 2022b.

MARCOS, J. L. et al. Spatial and temporal variability of hail falls and estimation of maximum diameter from meteorological variables. **Atmospheric Research**, v. 247, p. 105142, 1 jan. 2021.

MARSLAND, S. Machine Learning An Algorithmic Perspective Second Edition. p. 457, 2014.

MARTINS, J. A. et al. Climatology of destructive hailstorms in Brazil. **Atmospheric**

Research, v. 184, n. October 2016, p. 126–138, 2017.

MARTIUS, O. et al. Challenges and Recent Advances in Hail Research. **Bulletin of the American Meteorological Society**, v. 99, n. 3, p. ES51–ES54, 1 mar. 2018.

MCGOVERN, A. et al. Using artificial intelligence to improve real-time decision-making for high-impact weather. **Bulletin of the American Meteorological Society**, v. 98, n. 10, p. 2073–2090, 2017.

MCPHILLIPS, L. E. et al. Defining Extreme Events: A Cross-Disciplinary Review. **Earth's Future**, v. 6, n. 3, p. 441–455, 1 mar. 2018.

MICHAEL, R.; STUART, A. L. The fate of volatile chemicals during wet growth of a hailstone. **Environmental Research Letters**, v. 4, n. 1, p. 015001, 13 jan. 2009.

MORAIS, M. V. B. DE. Comparação bibliográfica sobre ensino de matemática para pessoas com Transtorno Autista utilizando técnica de Mineração de Texto. **REMAT: Revista Eletrônica da Matemática**, v. 8, n. 1, p. e2002–e2002, 31 jan. 2022.

NAGA CHANDRIKA, G. et al. Web Scraping for Unstructured Data Over Web. **Advances in Intelligent Systems and Computing**, v. 1076, p. 853–859, 2020.

PEGORIM. **Linha de instabilidade arrasa em SP - Notícias Climatempo**. Disponível em: <https://www.climatempo.com.br/noticia/2015/09/08/linha-de-instabilidade-arrasa-em-sp-6455>. Acesso em: 12 abr. 2023.

PREIN, A. F.; HOLLAND, G. J. Global estimates of damaging hail hazard. **Weather and Climate Extremes**, v. 22, p. 10–23, 1 dez. 2018a.

PREIN, A. F.; HOLLAND, G. J. Global estimates of damaging hail hazard. **Weather and Climate Extremes**, v. 22, p. 10–23, 1 dez. 2018b.

PREIN, A. F.; HOLLAND, G. J. Global estimates of damaging hail hazard. **Weather and Climate Extremes**, v. 22, p. 10–23, 1 dez. 2018c.

PULLMAN, M. et al. Applying Deep Learning to Hail Detection: A Case Study. **ITGRS**, v. 57, n. 12, p. 10218–10225, 1 dez. 2019.

PUSKEILER, M.; KUNZ, M.; SCHMIDBERGER, M. Hail statistics for Germany derived from single-polarization radar data. **Atmospheric Research**, v. 178–179, p. 459–470, 2016.

QI, R.; GUO, X. Analysis of Intelligent Energy Saving Strategy of 4G/5G Network Based on FP-Tree. **Procedia Computer Science**, v. 198, p. 486–492, 1 jan. 2022.

RÄDLER, A. T. et al. Frequency of severe thunderstorms across Europe expected to increase in the 21st century due to rising instability. **npj Climate and Atmospheric Science 2019 2:1**, v. 2, n. 1, p. 1–5, 27 ago. 2019.

RASCHKA, S. **Python Machine learning, deep learning, scikit-learn, and tensorflow**. [s.l: s.n.]. v. 31

RAUPACH, T. H. et al. The effects of climate change on hailstorms. **Nature Reviews Earth & Environment 2021 2:3**, v. 2, n. 3, p. 213–226, 9 fev. 2021.

RENCHER, A. C.; SCHAALJE, G. BRUCE. Linear models in statistics. p. 672, 2008.

SAPPI-RS. **Mês de julho começa com chuva, frio e geadas no Rio Grande do Sul - Secretaria da Agricultura, Pecuária, Produção Sustentável e Irrigação**. Disponível em: <https://www.agricultura.rs.gov.br/mes-de-julho-comeca-com-chuva-frio-e-geadas-no-rio-grande-do-sul>. Acesso em: 12 abr. 2023.

SELAN, M. et al. Hail net cover, cultivar and pod size influence the chemical composition of dwarf French bean. **Scientia Horticulturae**, v. 175, p. 95–104, 15 ago. 2014.

SHAH, H.; HELLEGERS, P.; SIDERIUS, C. Climate risk to agriculture: A synthesis to define different types of critical moments. **Climate Risk Management**, v. 34, p. 100378, 1 jan. 2021.

ULFAH, A.; NAJIAH, I. IMPLEMENTASI WEB SCRAPING PADA SITUS JURNAL SINTA MENGGUNAKAN FRAMEWORK SELENIUM WEBDRIVER PYTHON. **JIKA (Jurnal Informatika)**, v. 7, n. 1, p. 29–36, 16 fev. 2023.

WANG, J.; CHENG, Z. FP-Growth based Regular Behaviors Auditing in Electric Management Information System. **Procedia Computer Science**, v. 139, p. 275–279, 1 jan. 2018.

WANG, K. et al. Top down FP-growth for association rule mining. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 2336, p. 334–340, 2002.

APÊNDICE A – Configurações restantes CRON do ambiente linux utilizado para automatização da coleta de dados

CONFIGURAÇÕES RESTANTES DO AMBIENTE LINUX UTILIZADO PARA AUTOMATIZAÇÃO DA COLETA DE DADOS

40 23 * * * python3 ~/aplicacao/dashboard_granizo/apps/redemet/views.py

A configuração acima, executa o arquivo views.py da pasta redemet, uma vez por dia no horário das 23:40 minutos, todos os dias do mês. O arquivo views.py contém o algoritmo de coleta de dados referente a base da REDEMET, responsável por diariamente coletar os dados e os salvar coleção TEMP.

30 23 * * * python3 ~/aplicacao/dashboard_granizo/apps/ipmet/views.py

A configuração acima, executa o arquivo views.py da pasta ipmet, uma vez por dia no horário das 23:30 minutos, todos os dias do mês. O arquivo views.py contém o algoritmo de coleta de dados referente a base da IPMET, responsável por diariamente coletar os dados e os salvar nas coleções IPMET e CENTER DATA.

50 23 * * * python3 ~/aplicacao/dashboard_granizo/apps/s2id/views.py

A configuração acima, executa o arquivo views.py da pasta s2id, uma vez por dia no horário das 23:50 minutos, todos os dias do mês. O arquivo views.py contém o algoritmo de coleta de dados referente a base da S2ID, responsável por diariamente coletar os dados e os salvar nas coleções.

APÊNDICE B – Aeródromos utilizados para coleta de dados METAR da REDEMET

AERÓDROMOS QUE FORAM SELECIONADOS COMO PARÂMETROS DE COLETA RESTRITA AOS ESTADOS LIMITADOS NO ESTUDO, ESTADOS DA REGIÃO SUL DO BRASIL E ESTADO DE SÃO PAULO.

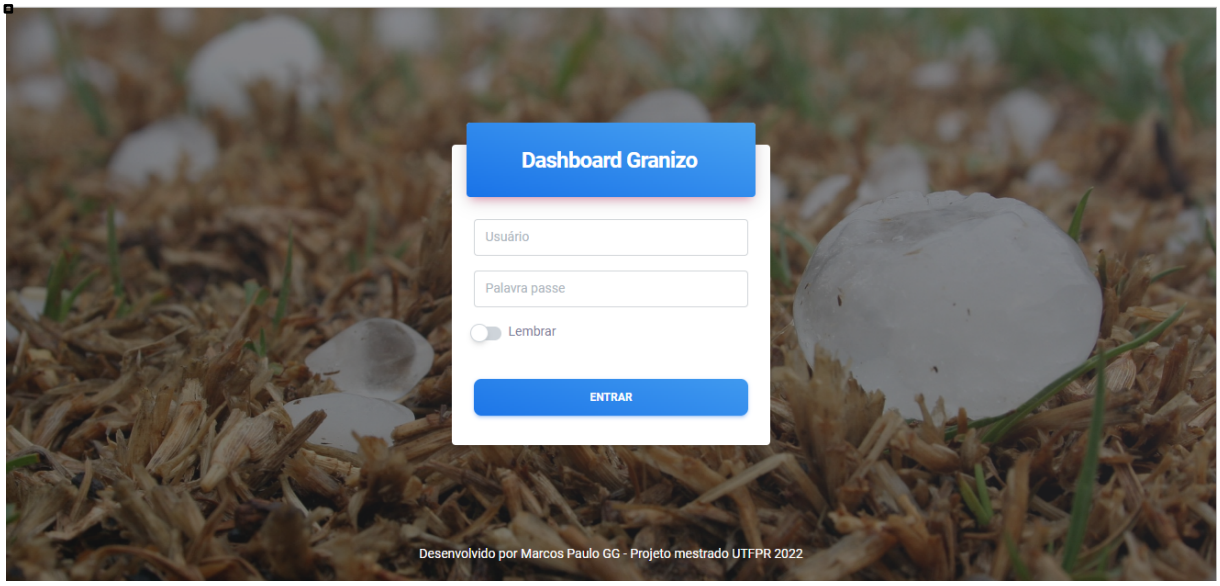
cidade	cod	id_ext	latitude	longitude	nome	pais	uf
BAURU	SBAE	3228	-22.1585	-49.0735	Aeroporto Internacional Bauru-Arealva / Moussa Nak	BRASIL	SP
ARARACUARA	SBAQ	3231	-21.8119	-48.1328	Aeroporto de Araraquara	BRASIL	SP
ARACATUBA	SBAU	3234	-21.1411	-50.4247	Aeroporto de Aracatuba	BRASIL	SP
BRAGANÇA PAULISTA	SBBP	3240	-22.9792	-46.5375	Aeroporto Estadual Arthur Siqueira	BRASIL	SP
BAURU	SBBU	3243	-22.345	-49.0536	Aeródromo de Bauru	BRASIL	SP
PRESIDENTE PRUDENTE	SBDN	3264	-22.175	-51.4244	Aeroporto de Presidente Prudente	BRASIL	SP
GAVIÃO PEIXOTO	SBGP	3278	-21.7644	-48.4047	Aeródromo de Gavião Peixoto	BRASIL	SP
SÃO PAULO	SBGR	3279	-23.4322	-46.4692	Aeroporto Internacional de São Paulo/Guarulhos	BRASIL	SP
GUARATINGUETÁ	SBGW	3282	-22.7914	-45.2047	Aeroporto de Guaratinguetá Edu Chaves	BRASIL	SP
JUNDIAÍ	SBJD	3292	-23.1817	-46.9436	Aeroporto Estadual de Jundiaí / Comandante Rolim Adolfo Amaro	BRASIL	SP
SÃO ROQUE	SBJH	4236	-23.4269	-47.1658	SAO PAULO CATARINA AEROPORTO EXECUTIVO	BRASIL	SP
CAMPINAS	SBKP	3301	-23.0081	-47.1344	Aeroporto Internacional de Viracopos / Campinas	BRASIL	SP
MARÍLIA	SBML	3314	-22.1967	-49.9264	Aeroporto Estadual de Marília - Frank Miloye Milenkovich	BRASIL	SP
SÃO PAULO	SBMT	3320	-23.5089	-46.6375	Aeroporto Campo de Marte	BRASIL	SP
RIBEIRÃO PRETO	SBRP	3345	-21.1342	-47.7742	Aeroporto de Ribeirão Preto / Dr. Leite Lopes	BRASIL	SP
SÃO JOSÉ DOS CAMPOS	SBSJ	3348	-23.2292	-45.8614	Aeroporto de São José dos Campos	BRASIL	SP
SÃO PAULO	SBSP	3352	-23.6267	-46.6553	Aeroporto de Congonhas / São Paulo	BRASIL	SP
SÃO JOSÉ DO RIO PRETO	SBSR	3353	-20.8164	-49.4064	Aeroporto Estadual Prof. Eribelto Manoel Reino	BRASIL	SP
GUARUJÁ	SBST	3354	-23.925	-46.2875	Base Aérea de Santos / BAST	BRASIL	SP
TAUBATÉ	SBTA	3356	-23.04	-45.5158	Base de Aviação de Taubaté / BAvT	BRASIL	SP
PIRASSUNUNGA	SBYS	3379	-21.9853	-47.3381	Academia da Força Aérea	BRASIL	SP
CAMPINAS,	SDAM	4241	-22.8591	-47.1080	Estadual de Campos dos Amarais	BRASIL	SP
SOROCABA,	SDCO	4238	-23.4830	-47.4863	Sorocaba	BRASIL	SP
CHAPECÓ	SBCH	3251	-27.1342	-52.6564	Aeroporto de Chapecó / Serafin Enoss Bertaso	BRASIL	SC
CRICIÚMA	SBCM	3254	-28.7256	-49.4247	Aeroporto Diomício Freitas	BRASIL	SC

FLORIANÓPOLIS	SBFL	3271	-27.6725	-48.5478	Aeroporto Internacional de Florianópolis - Hercílio Luz	BRASIL	SC
JAGUARUNA	SBJA	3290	-28.675	-49.0606	Aeroporto Regional Sul Humberto Ghizzo Bortoluzzi	BRASIL	SC
JOINVILLE	SBJV	3299	-26.2247	-48.7972	Aeroporto de Joinville- Lauro Carneiro de Loyola	BRASIL	SC
LAJES	SBLJ	3304	-27.7819	-50.2814	Aeroporto de Lages	BRASIL	SC
NAVEGANTES	SBNF	3322	-26.88	-48.6514	Aeroporto Internacional de Navegantes	BRASIL	SC
CORREIA PINTO,	SNCP	4248	-27.6341	-50.3583	Aeroporto Regional do Planalto Serrano	BRASIL	SC
BAGÉ	SBBG	3237	-31.3903	-54.1122	Aeroporto Comandante Gustavo Kraemer	BRASIL	RS
PORTO ALEGRE	SBCO	3256	-29.9458	-51.1444	Base Aérea de Canoas /BACO	BRASIL	RS
CAXIAS DO SUL	SBCX	3261	-29.1969	-51.1875	Aeroporto Regional de Caxias do Sul / Hugo Canterg	BRASIL	RS
PORTO ALEGRE	SBPA	3327	-29.9942	-51.1714	Aeroporto Internacional Salgado Filho	BRASIL	RS
PASSO FUNDO	SBPF	3330	-28.2439	-52.3264	Aeroporto Lauro Kurtz	BRASIL	RS
PELOTAS	SBPK	3333	-31.7183	-52.3275	Aeroporto de Pelotas	BRASIL	RS
SANTA MARIA	SBSM	3350	-29.7111	-53.6881	Aeroporto de Santa Maria	BRASIL	RS
TORRES	SBTR	3364	-29.41	-49.8103	Aeroporto de Torres	BRASIL	RS
URUGUAIANA	SBUG	3371	-29.7819	-57.0381	Aeroporto Internacional de Uruguaiana/Rubem Berta	BRASIL	RS
CURITIBA	SBBI	3239	-25.405	-49.2319	Aeroporto de Bacacheri	BRASIL	PR
CASCAVEL	SBCA	3246	-25.0003	-53.5006	Aeroporto Municipal de Cascavel	BRASIL	PR
SÃO JOSÉ DOS PINHAIS	SBCT	3259	-25.5283	-49.1756	Aeroporto Internacional Afonso Pena	BRASIL	PR
FOZ DO IGUAÇU	SBFI	3270	-25.5961	-54.4869	Aeroporto Internacional Cataratas	BRASIL	PR
LONDRINA	SBLO	3305	-23.3336	-51.13	Aeroporto de Londrina / Governador José Richa	BRASIL	PR
MARINGÁ	SBMG	3311	-23.4794	-52.0122	Aeroporto Regional de Maringá / Silvio Name Junior	BRASIL	PR
PONTA GROSSA -	SBPG	3331	-25.1877	-50.1444	Aeroporto Municipal de Ponta Grossa -	BRASIL	PR
PATO BRANCO /	SBPO	3336	-26.2167	-52.6867	Aeroporto de Pato Branco	BRASIL	PR
TOLEDO	SBTD	3359	-24.6853	-53.6967	Aeroporto de Toledo / Luiz Dal Canalle Filho	BRASIL	PR
GUARAPUAVA	SSGG	3280	-25.3883	-51.5236	Aeroporto Regional de Guarapuava - Tancredo Thomas de Faria (SSGG)	BRASIL	PR
UMUARAMA	SSUM	4233	-23.7999	-53.3138	Orlando de Carvalho	BRASIL	PR

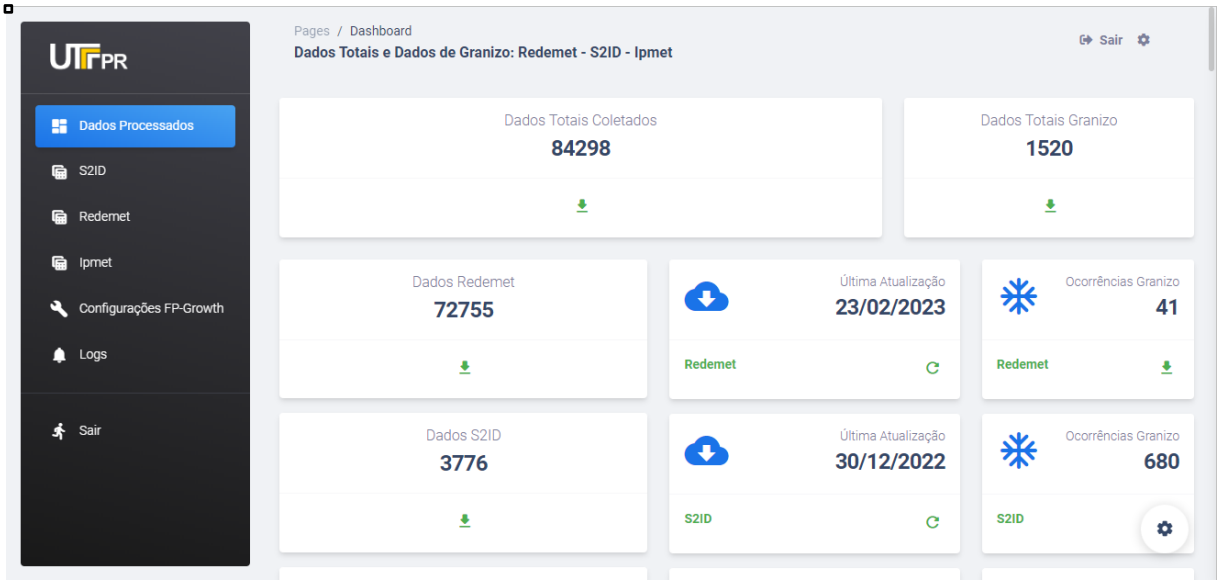
APÊNDICE C – Telas de ilustração Dashboard de monitoramento granizo

EXEMPLO DAS TELAS QUE COMPÕES A DASHBOARD DE MONITORAMENTO DESENVOLVIDA

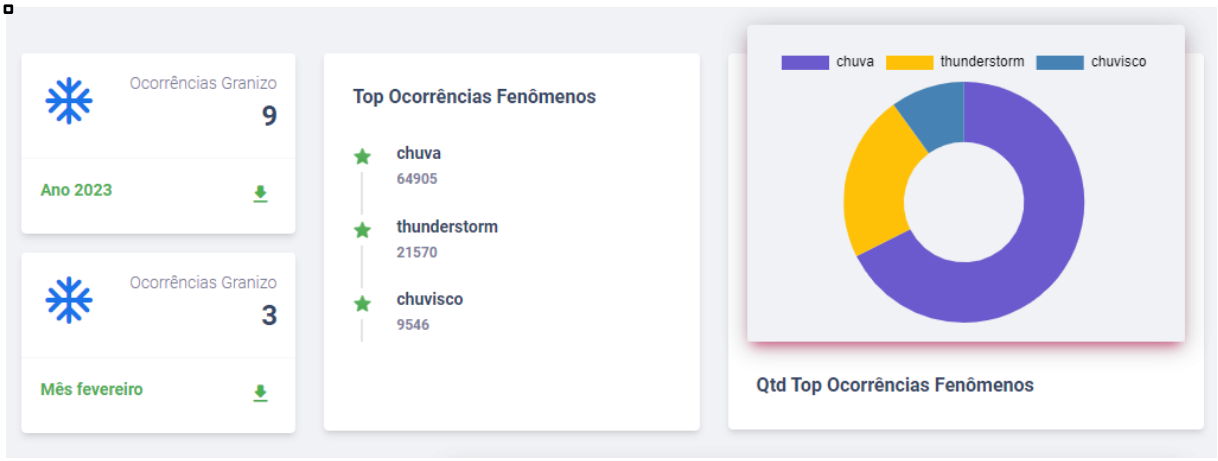
1. Tela inicial



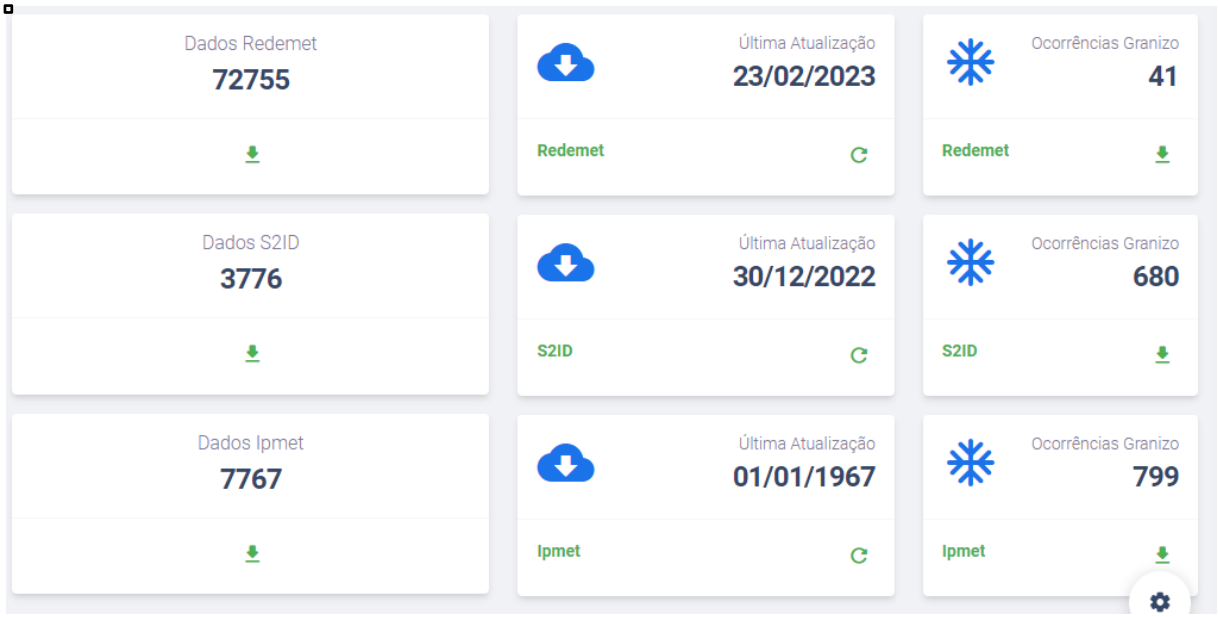
2. Tela principal da dashboard parte 1



3. Tela dos dados processados parte 2

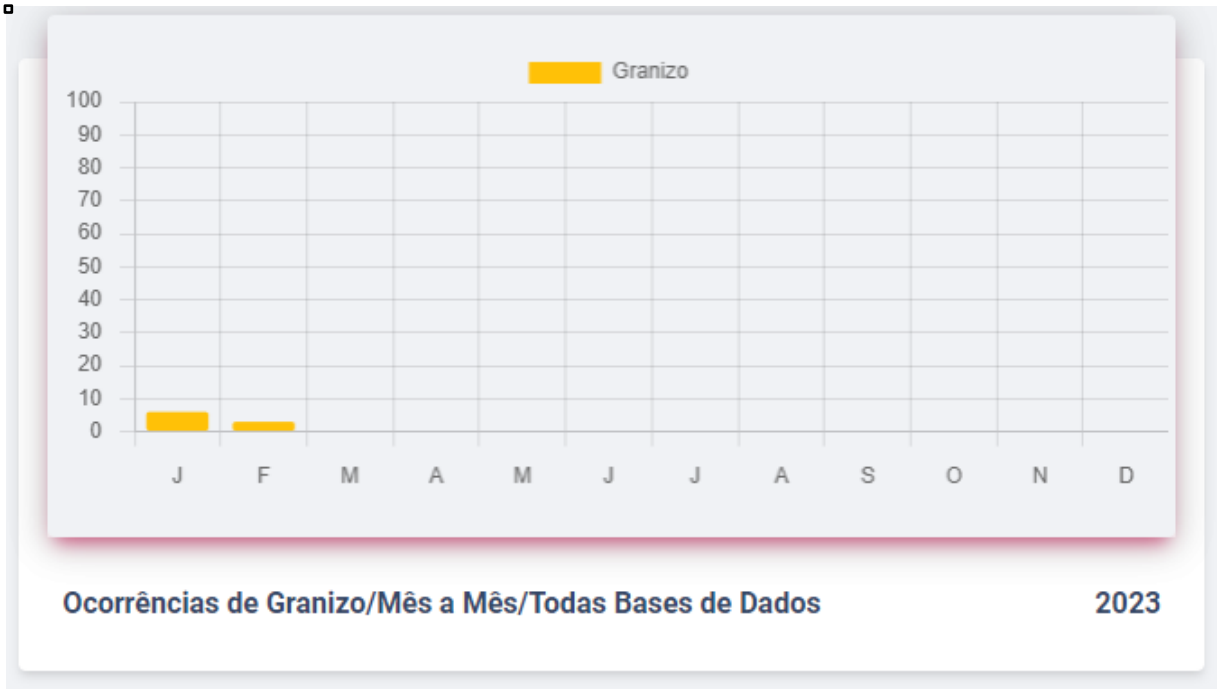


4. Tela principal da dashboard parte 3

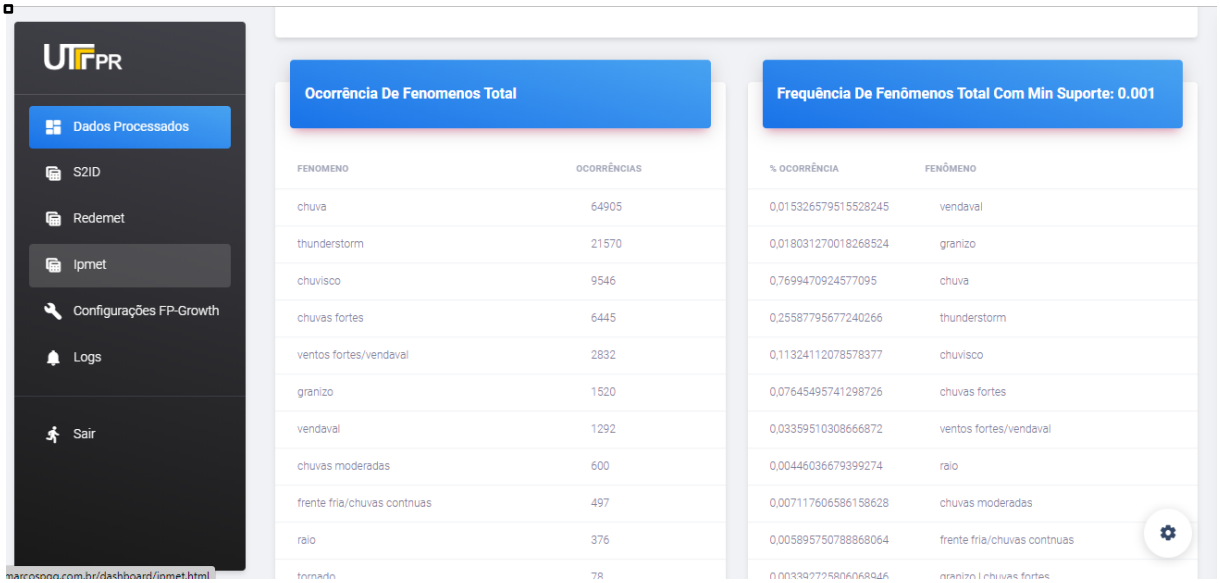


5. Tela principal da dashboard parte 4

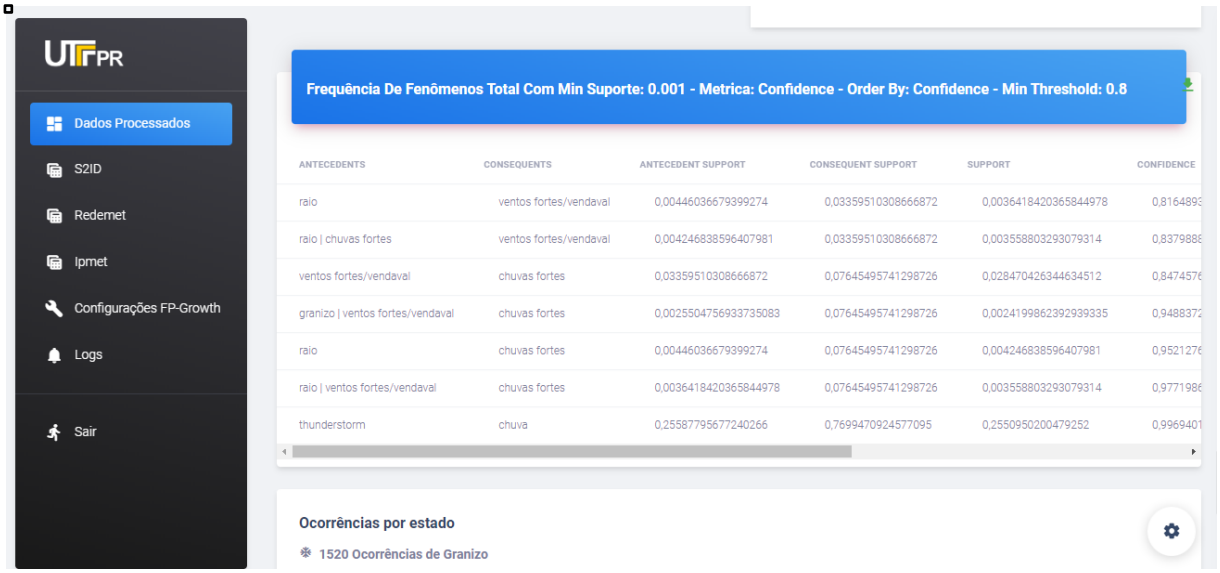
6. Tela principal da dashboard parte 5



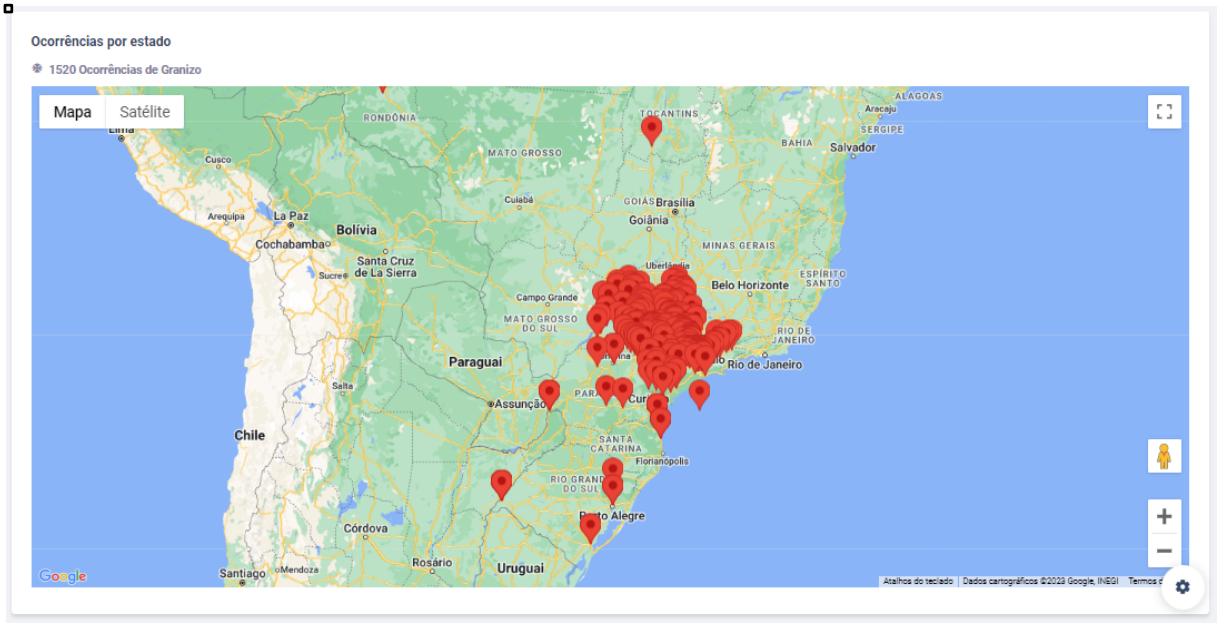
7. Tela principal da dashboard parte 6



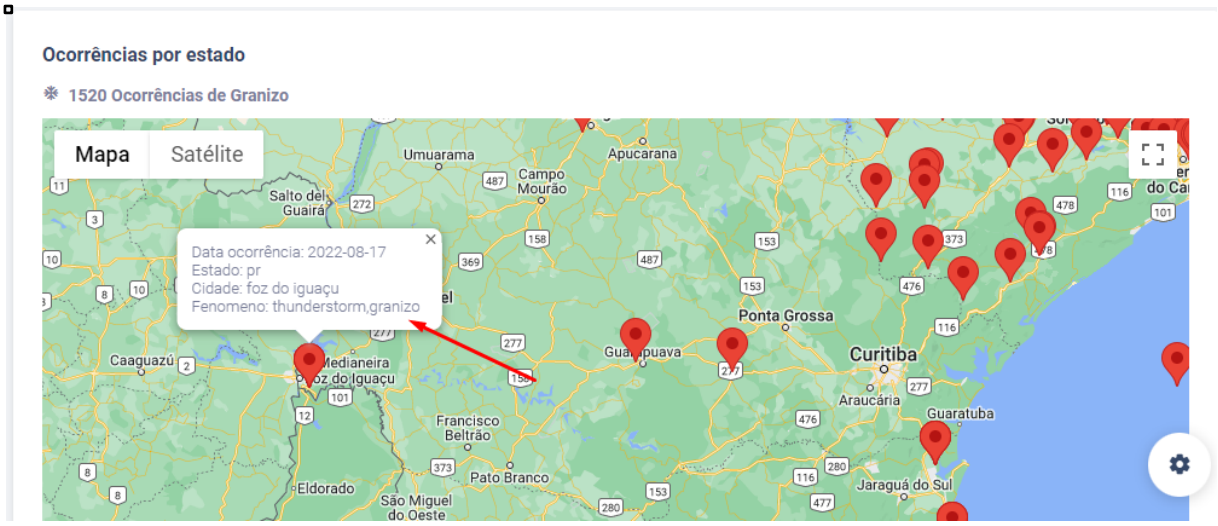
8. Tela principal da dashboard parte 7



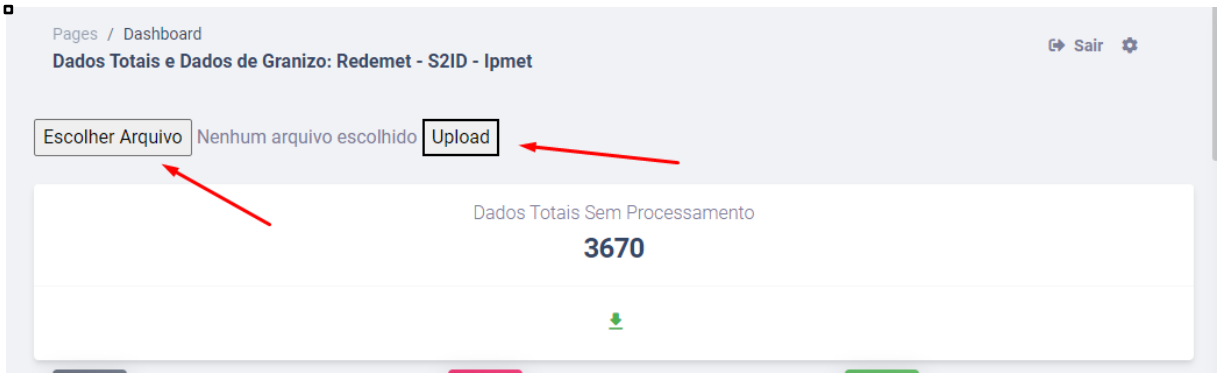
9. Tela principal da dashboard parte 8



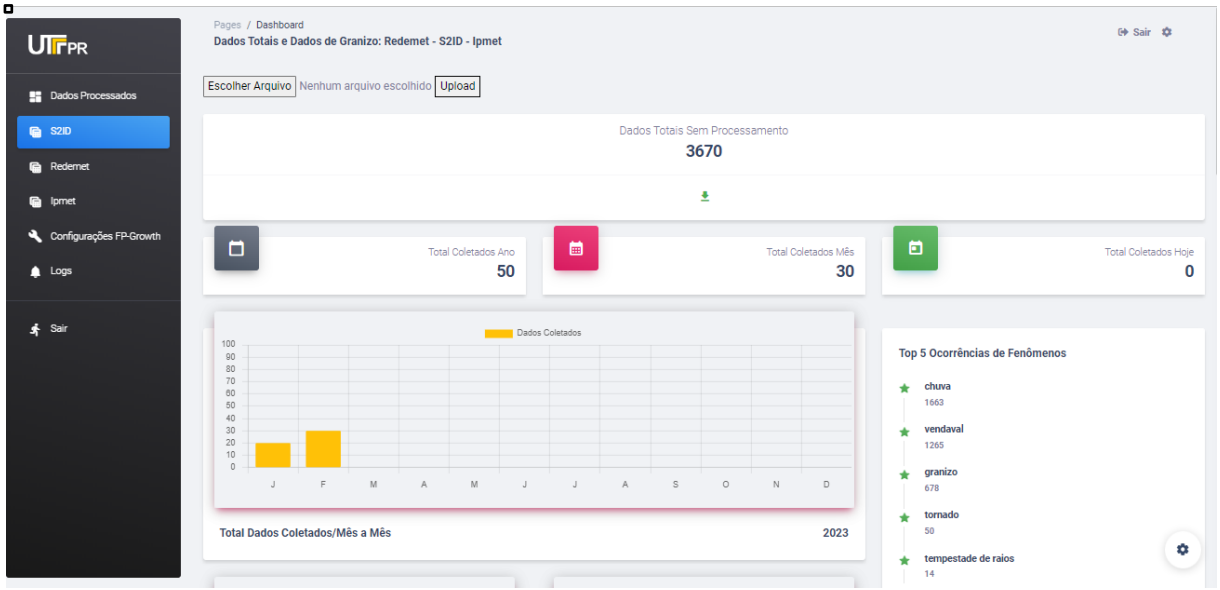
10. Tela principal da dashboard parte 9



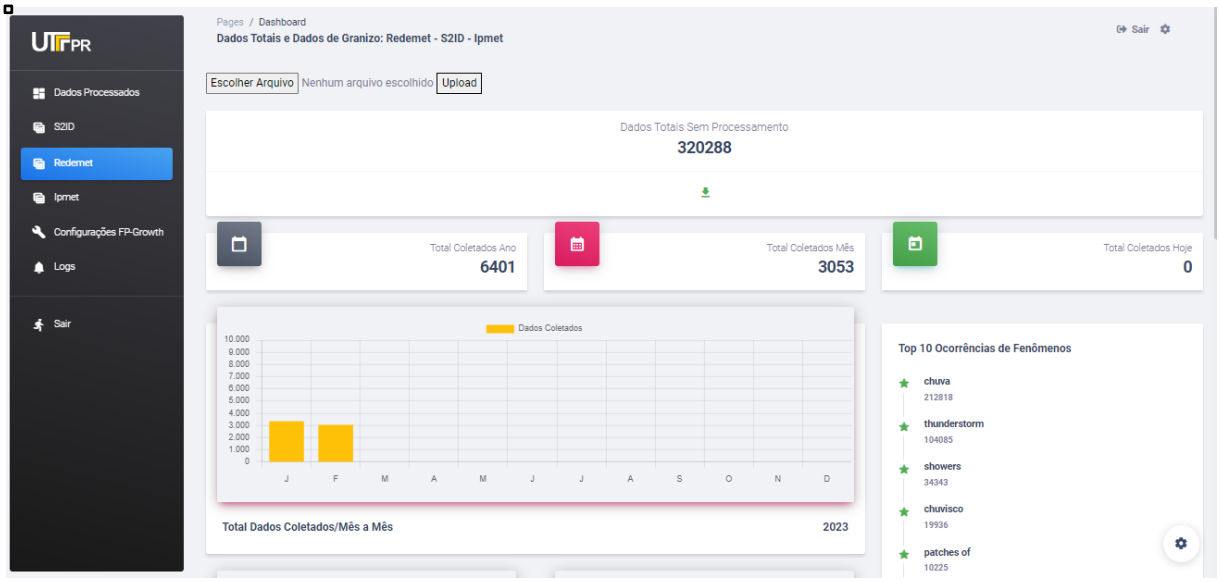
11. Tela demostrando que é possível fazer upload de arquivos .csv para agredar mais dados as bases de dados.



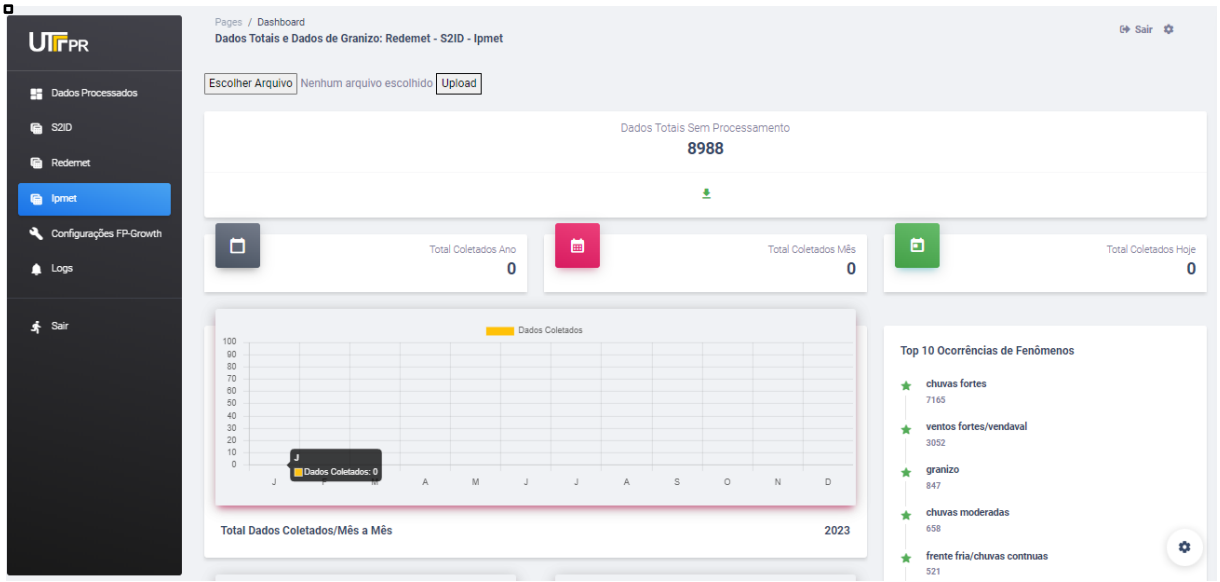
12. Tela da coleção de dados S2ID



13. Tela da coleção de dados REDEMET

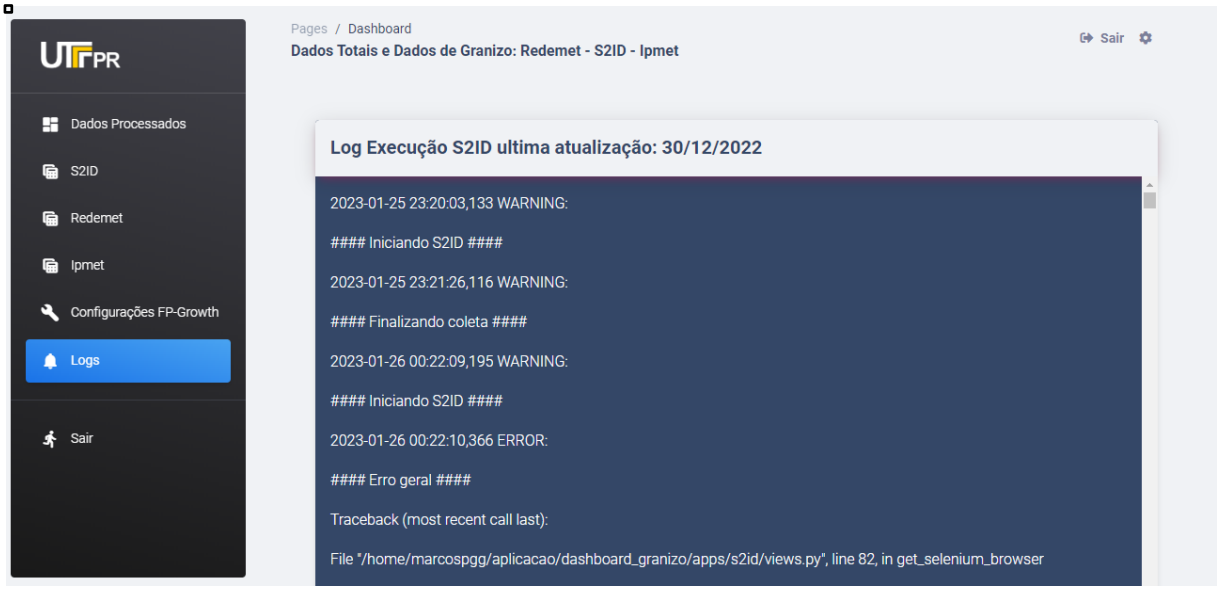


14. Tela da coleção de dados IPMET



15. Tela de configurações de parâmetros

16. Tela de logs dos buscadores automáticos de dados



17. **Tela com algumas configurações de personalização e seleção do ano para visualização dos dados.**