UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO E SISTEMAS
CURSO DE GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO CIVIL

Henrique Furini Lobato

MODELO PREDITIVO PARA AVALIAR A DEMANDA DE ITENS HUMANITÁRIOS EM CASOS DE DESASTRES ASSOCIADOS A EVENTOS CLIMÁTICOS EXTREMOS

Florianópolis
2025

Henrique Furini Lobato

MODELO PREDITIVO PARA AVALIAR A DEMANDA DE ITENS HUMANITÁRIOS EM CASOS DE DESASTRES ASSOCIADOS A EVENTOS CLIMÁTICOS EXTREMOS

Trabalho de Conclusão de Curso submetido ao curso de Engenharia de Produção Civil do Centro Tecnológico da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do título de Engenheiro Civil habilitado em Produção

Orientador(a): Prof.Ricardo Villarroel Dávalos

Florianópolis

2025

Lobato, Henrique Furini
Modelo preditivo para avaliar a demanda de itens humanitários em casos de desastres associados a eventos climáticos extremos / Henrique Furini Lobato ; orientadora, Ricardo Villarroel Dávalos, 2025.
106 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Engenharia de Produção Civil, Florianópolis, 2025.

Inclui referências.

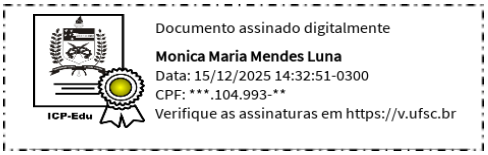
1. Engenharia de Produção Civil. 2. Machine Learning. 3. Desastres Naturais. 4. Logística Humanitária. 5. Modelos de Previsão. I. Dávalos, Ricardo Villarroel. II. Universidade Federal de Santa Catarina. Graduação em Engenharia de Produção Civil. III. Título.

Henrique Furini Lobato

MODELO PREDITIVO PARA AVALIAR A DEMANDA DE ITENS HUMANITÁRIOS EM CASOS DE DESASTRES ASSOCIADOS A EVENTOS CLIMÁTICOS EXTREMOS

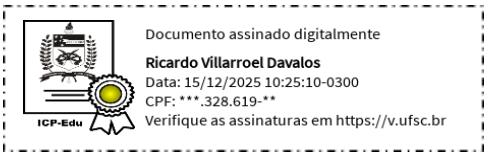
Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do título de Engenheiro Civil habilitado em Produção e aprovado em sua forma final pelo Curso de Engenharia de Produção Civil.

Florianópolis, 4 de dezembro de 2025.



Coordenação do Curso

Banca examinadora



Prof. Ricardo Villarroel Dávalos, Dr

Orientador

Insira neste espaço a assinatura

Prof(a). Analucia Schiaffino Morales, Dr(a)

Universidade Federal de Santa Catarina

Insira neste espaço a assinatura

Prof(a). Fabiana Lima, Dr(a)

Universidade Federal de Santa Catarina

Florianópolis, 2025

Este trabalho é dedicado especialmente aos meus pais e meu irmão, que me apoiaram e me inspiraram por toda minha vida.

AGRADECIMENTOS

Agradeço primeiramente a meus pais, Maurício e Márcia, e meu irmão Daniel, que fizeram de tudo por mim. Desde me apoiar nos estudos do vestibular para ingressar em engenharia, até me dar todo suporte para que eu pudesse vir à Florianópolis viver meu sonho de fazer faculdade em outra cidade. Esse trabalho é apenas um resumo de tudo que vocês me permitiram ser.

Agradeço também aos meus tios e tias e minhas avós, que também me apoiaram e sempre mostraram seu carinho quando eu retornava para São Paulo para visitar, estando todos dispostos a me ajudar com o que eu precisasse.

Em seguida agradeço aos amigos que fiz durante meu período na faculdade, que sempre estiveram ao meu lado para sorrir, chorar e brindar ao longo dos anos aqui na UFSC.

Por fim agradeço meu orientador, Ricardo, por todo o suporte e direcionamentos que me deu ao longo dos anos. Foi quem acreditou em mim e me concedeu uma bolsa de iniciação científica, que me levou ao trabalho com dados que faço hoje em dia, e quem aceitou me orientar na elaboração deste trabalho.

RESUMO

O estado de Santa Catarina (SC) apresenta uma vulnerabilidade significativa de eventos climáticos extremos, como enchentes, deslizamentos de terra e tempestades severas. Neste contexto, a Logística Humanitária (LH) assume um papel imprescindível na mitigação das perdas humanas e materiais, visando proporcionar um atendimento rápido e eficiente às áreas atingidas. A falta da aplicação de sistemas de previsão precisos e metodologias de resposta eficazes exacerba a vulnerabilidade dessas regiões, tornando-as mais suscetíveis a danos catastróficos. O objetivo deste trabalho é propor um modelo preditivo baseado em Machine Learning (ML) para estimar a demanda de itens de assistência humanitária solicitados à Secretaria de Defesa Civil de Santa Catarina (SEDEC/SC) em casos de desastres naturais. A partir desta proposta pretende-se o planejamento da LH frente ao aumento previsto de desastres hidroclimatológicos no estado. A metodologia adotada segue uma adaptação do processo de Descoberta de Conhecimento em Bases de Dados (KDD) , utilizando dados históricos de solicitações da SEDEC/SC , registros de eventos de desastre (CEMADEN) e dados meteorológicos (INMET). Os dados foram modelados dimensionalmente para integrar os indicadores climáticos com os registros de desastres e as solicitações de itens. A análise exploratória indicou predominância de eventos climatológicos e meteorológicos, como Chuvas Intensas e Estiagem, levando à exclusão dos eventos hidrológicos no treinamento. O modelo final é um *pipeline* preditivo de quatro etapas — previsão da ocorrência, tipo de evento, itens e quantidade — executado pelo algoritmo *XGBoost* em todas as fases. O *XGBoost* foi selecionado por apresentar métricas superiores de acurácia e *F1-Score*, além da melhor capacidade de captar a distribuição das quantidades solicitadas na previsão das variáveis contínuas.

Palavras-chave: Machine Learning; Logística Humanitária; Desastres Naturais; Modelos de Previsão.

ABSTRACT

The state of Santa Catarina (SC) presents significant vulnerability to extreme weather events, such as floods, landslides, and severe storms. In this context, Humanitarian Logistics (HL) assumes an indispensable role in mitigating human and material losses, aiming to provide a rapid and efficient response to the affected areas. The lack of application of accurate forecasting systems and effective response methodologies exacerbates the vulnerability of these regions, making them more susceptible to catastrophic damage. The objective of this work is to propose a predictive model based on Machine Learning (ML) to estimate the demand for humanitarian assistance items requested from the Santa Catarina Civil Defense Secretariat (SEDEC/SC) in cases of natural disasters. This proposal is intended to enable the planning of HL in face of the predicted increase in hydro-climatological disasters in the state. The adopted methodology follows an adaptation of the Knowledge Discovery in Databases (KDD) process, using historical data of requests from SEDEC/SC, disaster event records (CEMADEN - National Center for Monitoring and Early Warning of Natural Disasters), and meteorological data (INMET - National Institute of Meteorology). The data was dimensionally modeled to integrate the climate indicators with the disaster records and item requests. The exploratory analysis indicated a predominance of climatological and meteorological events, such as Intense Rainfall and Drought, leading to the exclusion of hydrological events from the training. The final model is a four-step predictive pipeline — forecasting occurrence, event type, items, and quantity — executed by the XGBoost algorithm in all phases. XGBoost was selected for presenting superior accuracy and F1-Score metrics, in addition to a better capacity to capture the distribution of the requested quantities in the prediction of continuous variables.

Keywords: Machine Learning; Humanitarian Logistics; Natural Disasters; Forecasting Methods.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

Em países como o Brasil, desastres naturais muitas vezes se mostram comuns devido a diversos fatores, mas principalmente por causa de mudanças climáticas e de ações humanas indevidas para com a natureza. No período de 2013 a 2022 foram registrados 58.469 decretos de emergência e estados de calamidade pública no país devido a desastres naturais (CNM, 2023).

Esses valores tendem a aumentar, como mostra o Relatório Síntese do Sexto Ciclo de Avaliação (AR6) do IPCC (2023). Ele aponta que o Brasil enfrentará um aumento na frequência e intensidade de desastres naturais de origem hidroclimatológica nos próximos anos. Com isso, o país sofrerá maior incidência de eventos extremos como inundações, chuvas intensas e secas severas em diferentes regiões.

Destacando o estado de Santa Catarina (SC) na série histórica do CNM, ele é o 4° estado em número de emergências decretadas, sendo um total de 4.732 decretos emergenciais no período analisado, representando 342.854 pessoas afetadas (CNM, 2023). Já quanto ao futuro previsto para SC, o estado será muito afetado pelas mudanças climáticas, com efeitos já visíveis, como o aumento de graves enchentes no estado apontado pela *World Weather Attribution* (WWA) (2025).

Em eventos de desastre, a Secretária de Defesa Civil de Santa Catarina (SEDEC/SC) é acionada para prestar assistência humanitária ao local e às pessoas afetadas. Para garantir essa assistência, ela faz a logística de recursos humanos e materiais para essa assistência seguindo métodos de Logística Humanitária (LH), a qual caracteriza os processos de mobilização desses recursos para assistir comunidades afetadas por desastres, entre outros tipos de emergência, fazendo isso no menor tempo possível e com um orçamento limitado, mas visando a prestar ajuda para o maior número de pessoas possível (Meirim, 2006).

Entre os recursos materiais, a SEDEC/SC fornece itens de assistência humanitária destinados às pessoas afetadas por diferentes tipos de desastres. Entre esses recursos, destacam-se os itens de nutrição e hidratação, como cestas básicas e reservatórios de água potável. Também são distribuídos kits de higiene contendo sabonetes, escovas de dentes e outros produtos essenciais. Nos casos em que há

pessoas desabrigadas, a SEDEC/SC presta apoio adicional, disponibilizando colchões, roupas e edredons às famílias necessitadas.

Em seu Plano Nacional de Defesa Civil (2003) a SEDEC define quatro grandes etapas (ou planos) para a atuação frente a desastres: Prevenção, Preparação, Resposta e Reconstrução. Destacando o segundo plano, ele visa otimizar as ações preventivas, de resposta aos desastres e de reconstrução, por meio de projetos voltados ao desenvolvimento institucional, capacitação de recursos humanos, planejamento operacional e de contingência, apoio logístico, entre outros.

Dentro do planejamento operacional, estimar as quantidades de itens necessários para atender às pessoas afetadas por algum desastre pode ser caracterizado como um problema de previsão de demanda, já que há a necessidade de estimar os recursos necessários para a execução dos processos de LH em meio à incerteza do local, tipo e momento do desastre. E para a previsão da demanda de recursos, Rushton, Croucher e Baker (2010) afirmam que a análise da série histórica desta deve ser feita aliada a um levantamento dos fatores que a influenciam.

Neste contexto, o presente trabalho visa propor um modelo preditivo baseado em *Machine Learning* que utilize as séries de dados históricos do clima em Santa Catarina, de eventos de desastres naturais no estado, e das solicitações de itens humanitários à SEDEC/SC, para prever os futuros desastres e itens que serão solicitados à secretaria.

1.1. PROBLEMA

Modelos preditivos baseados em *Machine Learning* (ML) são usados amplamente em empresas de diversos setores. Seu amplo uso aponta que os resultados das aplicações desses modelos são surpreendentes, podendo reduzir desperdícios e auxiliar no planejamento operacional de diferentes estruturas produtivas (Davenport, 2018).

Com a crescente de desastres naturais apontada pelo AR6 (IPCC, 2023), e os benefícios do uso de modelos preditivos baseados em ML para estimar demandas, surge o problema da pesquisa: Qual o melhor modelo de previsão baseado em ML para prever a demanda de itens de assistência humanitária solicitados em eventos de desastres naturais futuros?

1.2. OBJETIVOS

1.2.1. Objetivo Geral

Propor um modelo de previsão baseado em ML para estimar a demanda de itens de assistência humanitária em situações de desastres naturais no estado de Santa Catarina.

1.2.2. Objetivos Específicos

a) Organizar os dados de desastres ocorridos e itens solicitados para análise exploratória e aplicação de modelos de ML;
b) Analisar a distribuição de desastres naturais pelo estado de Santa Catarina (SC);
c) Testar diferentes modelos de ML frente aos dados organizados;
d) Criar um modelo de previsão baseado em ML capaz de prever possíveis eventos de desastres naturais no estado;
e) Identificar qual o melhor algoritmo de ML para ser aplicado no modelo proposto.

1.3. JUSTIFICATIVA

Desastres naturais afetam o acesso de populações inteiras à comida, água, produtos de higiene e até itens de vestuário. Com a previsão do IPCC (2023) do aumento de desastres naturais dada às mudanças climáticas, mais pessoas serão afetadas por desastres e precisarão da ajuda de órgãos humanitários para o envio de itens de assistência.

Nesse contexto, a SEDEC tem protagonismo e grandes responsabilidades para atingir seu objetivo geral de reduzir as ocorrências e intensidades dos desastres (SEDEC, 2003). Essa intensidade de desastres está intrinsecamente ligada à quantidade de pessoas afetadas por eles, que podem ser afetadas pela falta de recursos prontamente disponíveis para os esforços de LH.

Uma falta desses recursos pode vir da falta de itens específicos ou da demora de processos burocráticos para sua aquisição. Souza (2012) aponta que para um evento de enchente no Vale do Itajaí em 2012, foram necessários 53 dias para que houvesse a liberação de recursos e posterior aquisição de itens de assistência.

Logo, o planejamento antecipado e analítico dos itens que compõem a ajuda prestada pela SEDEC é de extrema importância para auxiliar a secretária a trabalhar alinhada com seu objetivo geral. Isso porque com a visão da demanda de itens a secretaria pode antecipar as burocracias necessárias para compra e regular os estoques de seus centros de distribuição, reduzindo assim a intensidade dos desastres futuros no estado.

E para garantir uma assertiva visão da demanda futura, os modelos de ML são os que entregam um melhor resultado quando comparados com métodos estatísticos de previsão. Grobman e Cugnasca (2025) apontam isso, demonstrando que uma previsão baseada em um modelo estatístico teve um desempenho pior do que todos os 5 modelos de ML testados contra ele para a previsão de demanda aplicada ao setor varejista.

1.4. ESTRUTURA DO TRABALHO

Este trabalho está estruturado partindo de uma contextualização sobre o tema de logística humanitária em desastres naturais em SC, seguido pelos objetivos e justificativa da relevância do trabalho. Na segunda seção do documento é realizada uma revisão bibliográfica acerca do tema de Logística Humanitária (LH); sobre as definições usadas pela SEDEC sobre desastres, defesa civil e tipos de desastres naturais; tratando também da explicação detalhada sobre os algoritmos de ML utilizados no trabalho, e suas formas de avaliação; terminando com a explicação da metodologia usada para a organização dos dados do trabalho.

Em seguida, na seção 3, é apresentada a metodologia usada: uma adaptação do *Knowledge Discovery in Databases* (KDD). Na seção 4, é apresentada a proposta central do trabalho: o pipeline preditivo de quatro estágios (previsão de ocorrência, tipo de evento, itens e quantidade), e justifica o uso do algoritmo *XGBoost*, que demonstrou as melhores métricas avaliativas. Por fim, a Conclusão apresenta a síntese dos principais resultados alcançados, verifica o cumprimento do objetivo geral e propõe caminhos para o desenvolvimento de pesquisas futuras.

2. REVISÃO TEÓRICA

2.1. LOGÍSTICA HUMANITÁRIA

A Logística Humanitária (LH) é um campo de estudo e prática que, embora tenha ganhado maior tração acadêmica no início dos anos 2000, é fundamental para a resposta a desastres. Ela é parte essencial da Cadeia de Suprimentos Humanitária (CSH) e foca em garantir o fluxo de bens, serviços e informações (Cislaghi; Fernandes; Doriscat, 2024).

Uma das primeiras definições para o campo foi proposta por Thomas e Kopczack (2005), que a descreveram como:

> O processo de planejar, implementar e controlar o fluxo e armazenamento eficiente e de custo-benefício de bens e materiais, bem como informações relacionadas, desde o ponto de origem até o ponto de consumo, com o propósito de aliviar o sofrimento de pessoas vulneráveis (Thomas; Kopczack, 2005).

A partir dessa definição e estudos realizados por Nogueira, Gonçalves e Novaes (2008), pode-se definir que o objetivo principal da LH é auxiliar as vítimas, focando no alívio do sofrimento e na preservação da vida, mobilizando pessoas, recursos e conhecimento para atender comunidades vulneráveis no menor tempo possível.

As operações de LH atuam dentro do ciclo de gerenciamento de desastres, para o qual a SEDEC define 4 planos de ação (ver seção 2.2.2). No segundo plano, a Preparação, que foca em ações anteriores à resposta do desastre, a literatura aponta uma série de processos necessários para garantir que o objetivo da LH seja cumprido.

O primeiro é o planejamento de resposta. Nesta etapa, realiza-se a avaliação de riscos sob os locais de interesse, a avaliação de necessidades de recursos – desde itens até pessoas – e o planejamento das ações de resposta para possíveis eventos (Altay et al., 2024). Um componente vital deste planejamento é a formação de parcerias estratégicas, que são identificadas como um determinante significativo para o desempenho da cadeia de suprimentos (Brian; Shale, 2017).

Após o planejamento, a aquisição de itens humanitários e seu armazenamento são atividades centrais (Thomas; Kopczack, 2005). A prática de pré-posicionamento de estoque, em particular, demonstrou afetar positiva e significativamente o desempenho da cadeia de suprimentos humanitária. O

posicionamento estratégico de suprimentos em centros de distribuição próximos às áreas de maior risco não só permite que as entidades humanitárias antecipem desastres de forma mais eficaz, mas também possibilita a entrega rápida durante emergências. Além disso, essa prática contribui para a redução de custos logísticos, muitas vezes através de acordos pré-negociados com fornecedores (Brian; Shale, 2017).

Por fim, a gestão de transporte é o processo que conecta os estoques pré-posicionados às áreas afetadas. Este processo é complexo e vai além da simples definição de rotas, especialmente ao considerar que a infraestrutura local, como estradas, pontes e aeroportos, pode estar danificada ou destruída (Thomas; Kopczak, 2005). Para uma gestão eficaz, é necessário que sejam definidos modos de transporte possíveis para cada região de interesse, levantando alternativas de rotas em casos de dano às vias inicialmente selecionadas. Outro ponto de definição é a possibilidade de terceirização deste transporte, uma prática que se mostrou benéfica no estudo sobre uma ONG de resposta à desastres no Quênia feito por Brian e Shale (2017).

Toda a operação das entidades devem também estar preparadas para conseguir monitorar seu plano de resposta nos casos de desastres. Uma revisão literária feita por Rejeb et al. (2021) mostra que o uso de drones é uma ótima alternativa para isto, já que eles conseguem prover imagens panorâmicas da área afetada, além de conseguirem chegar ao local de forma muito mais rápida do que transportes terrestres comuns.

Apesar desta complexidade elevada LH, que se soma ao tratamento da vida humana, ela ainda carece de mais estudos sobre inovação e aplicação de novas tecnologias quando comparada à iniciativa privada. Essa comparação, apresentada no estudo de Altay et al. (2023) conclui que, quando comparada à logística empresarial, a pesquisa de inovação na CSH é um tópico subdesenvolvido.

Essa lacuna justifica a busca por métodos inovadores, como os de métodos de previsão com *machine learning* aplicados na logística empresarial. Um famoso caso de métodos inovadores, e adequado para inspirar a LH, é o uso de modelos de otimização e previsão de demandas para otimizar o tempo e quantidade de entregas feitas pelos caminhões da UPS, uma gigante de distribuição e logística nos Estados Unidos. Essa aplicação possibilitou que consumidores pudessem realizar um pedido

e ter sua entrega em poucas horas, além de diminuir a redundância no percurso dos entregadores e, consequentemente, o tempo de entrega (Davenport, 2017).

Inovações como essa, se aplicada em contextos humanitários, poderia salvar muitas vidas. A aplicação de modelos preditivos, como os de *Machine Learning*, surge como uma ferramenta essencial para a fase de Preparação, permitindo estimar o que será necessário e onde, otimizando o pré-posicionamento de estoque e o planejamento da resposta (Brian; Shale, 2017).

2.2. DEFINIÇÕES DA SEDEC

Essa seção apresenta inicialmente as definições sobre os termos trabalhados neste estudo, dadas pela própria SDEC em seu glossário, seguindo por apresentar cada um dos eventos de desastre estudados para a formulação do modelo preditivo, com suas definições oficiais e causas.

2.2.1. Desastres

O Glossário de Defesa Civil (s.d., p. 57) define desastre como: “resultado de eventos adversos, naturais ou provocados pelo homem, sobre um ecossistema (vulnerável), causando danos humanos, materiais e/ou ambientais e consequentes prejuízos econômicos e sociais.”

A intensidade de um desastre é o resultado da interação entre a magnitude do evento e a vulnerabilidade do ecossistema afetado, quantificado em função dos danos (humanos, materiais e ambientais) e dos prejuízos econômicos e sociais resultantes.

De acordo com a SEDEC, os desastres são classificados em quatro níveis baseando-se na relação entre a necessidade de recursos para o restabelecimento da normalidade e a disponibilidade desses recursos na área afetada.

Os desastres de Nível I (Pequena Intensidade ou Acidentes) são caracterizados por poucos danos e prejuízos baixos, sendo facilmente suportáveis e superáveis pelas comunidades afetadas. A normalidade é restabelecida com os recursos existentes e disponíveis na área (município), sem necessidade de grandes mobilizações.

Já os desastres de Nível II (Média Intensidade) apresentam danos de alguma importância e prejuízos significativos. Podem ser suportáveis e superáveis por

comunidades bem informadas, preparadas e participativas, utilizando os recursos existentes no município, desde que racionalmente mobilizados.

Desastres de Nível III (Grande Intensidade) são marcados por danos importantes e prejuízos vultosos. Apesar disso, são suportáveis e superáveis por comunidades bem informadas e preparadas, mas exigem que os recursos mobilizados na área afetada sejam reforçados com o aporte de recursos estaduais e federais já disponíveis.

Enfim, desastres de Nível IV (Muito Grande Intensidade) caracterizam-se por danos muito importantes e prejuízos muito vultosos e consideráveis. Nessas condições, não são superáveis e suportáveis pelas comunidades afetadas por si só, mesmo quando bem preparadas, e o restabelecimento da normalidade depende da mobilização e ação coordenada dos três níveis do Sistema Nacional de Defesa Civil (SINDEC) e, em alguns casos, de ajuda internacional.

2.2.2. Defesa Civil

O Glossário de Defesa Civil (s.d., p. 54) define defesa civil como: “conjunto de ações preventivas, de socorro, assistenciais e reconstrutivas destinadas a evitar ou minimizar os desastres, preservar a moral da população e restabelecer a normalidade social”.

No Plano Nacional de Defesa Civil (2007), a SEDEC define 4 diferentes tipos de planos para realizar a defesa civil: prevenção, preparação, resposta e reconstrução.

Planos de Prevenção de desastres focam em evitar ou minimizar a ocorrência e a intensidade dos desastres. Isso é feito através da avaliação de riscos de desastres (estudo de ameaças e vulnerabilidades, hierarquização de riscos e mapeamento) e da redução de riscos (medidas preventivas não-estruturais, como planejamento do uso do solo e legislação, e estruturais, como obras de engenharia).

Os planos para a Preparação tem como meta otimizar as ações de prevenção, resposta e reconstrução. Inclui o desenvolvimento institucional e de recursos humanos, o avanço científico e tecnológico, a mudança cultural, a motivação empresarial, a coleta de informações e estudos epidemiológicos, a monitorização, alerta e alarme, e o planejamento operacional e de contingência.

Planos de Resposta consistem em ações imediatas após a ocorrência de um desastre. Abrange as fases de socorro (pré-impacto, impacto e limitação de danos), assistência às populações vitimadas (logística, assistencial e promoção da saúde) e reabilitação do cenário do desastre (avaliação de danos, remoção de escombros, limpeza e reabilitação de serviços essenciais).

Por fim, planos de Reconstrução visam restabelecer plenamente os serviços públicos, a economia da área, o moral social e o bem-estar da população. Frequentemente, essa fase se confunde com a prevenção, buscando recuperar ecossistemas, reduzir vulnerabilidades, racionalizar o uso do solo, realocar populações para áreas de menor risco, e modernizar/reforçar infraestruturas.

2.2.3. Desastres Naturais

O Manual de Desastres (SEDEC, 2003) é um material que concentra conhecimentos científicos acerca de diferentes eventos de desastres naturais para o auxílio nos esforços da SEDEC. As definições e as causas definidas de cada um dos desastres naturais estudados por este trabalho estão no Quadro 1 abaixo.

Quadro 1 – Definições e causas dos eventos de desastre

(continua)

Grupo	Evento	Definição	Causas
Hidrológico	Alagamentos	Água acumulada no leito das ruas e no perímetro urbano por fortes precipitações pluviométricas, em cidades com sistemas de drenagem deficientes.	Fortes precipitações pluviométricas em áreas urbanas com sistemas de drenagem deficientes.
Hidrológico	Enxurradas	Volume de água que escoa na superfície do terreno com grande velocidade, resultante de fortes chuvas. Corresponde a uma cheia de pequena duração com uma descarga de ponta relativamente alta.	O escoamento superficial rápido da chuva em áreas com declives e pouca infiltração de água no solo.
Hidrológico	Inundações	Transbordamento de água da calha normal de rios, mares, lagos e açudes, ou o acúmulo de água por drenagem deficiente em áreas normalmente não submersas.	Precipitações intensas e concentradas, degelo ou saturação do lençol freático. Pode ser causada também por o assoreamento de rios, impermeabilização do solo, rompimento de barragens e marés elevadas com chuvas intensas.

Quadro 1 – Definições e causas dos eventos de desastre

(conclusão)

Grupo	Evento	Definição	Causas
Meteorológico	Chuvas Intensas	Precipitação intensa de chuva, geralmente durante um período curto. Caracteriza-se pelo início e fim inesperados e por grandes e rápidas variações de intensidade.	O movimento ascendente de massas de ar, como o que ocorre em montanhas, pode originar chuvas orográficas.
Meteorológico	Ciclones	Área de concentração de energia cinética na atmosfera, resultando em ventos fortes.	O aquecimento do oceano e a instabilidade atmosférica favorecem a rotação e intensificação do vento.
Meteorológico	Granizo	Precipitação de grânulos ou pedras de gelo, com diâmetro igual ou superior a 5 mm.	Ocorre como uma forma de precipitação sólida.
Meteorológico	Vendavais	Deslocamento violento de uma massa de ar. Corresponde ao nível 10 da Escala de Beaufort, com ventos entre 88 e 102 km/h.	Formam-se pelo deslocamento de ar de uma área de alta pressão para uma de baixa pressão. Frequentemente ocorrem com a passagem de frentes frias.
Climatológico	Estiagem	Período prolongado de baixa ou ausência de pluviosidade, onde a perda de umidade do solo é maior que sua reposição.	Longos períodos sem chuva, agravados por fatores climáticos e perda da vegetação que retém umidade

Fonte: SEDEC (2003)

2.3. ALGORITMOS DE MACHINE LEARNING (ML)

Para a busca de padrões entre um conjunto de dados, bem como as relações entre diferentes colunas de dados são aplicados algoritmos de ML. Esses métodos atuam a partir da criação de funções $, y = f(x)$ onde x é um *input* (entrada) de dados, $f()$ é o método, e y é o *output* (saída) esperado. O que $f()$ faz é, com base no treinamento, identificar as correlações entre os dados de x possuem com a variável alvo e estimar qual o valor ou a classe para cada entrada de x (Bishop, 2006).

A forma como $f()$ faz as correlações e a busca por padrões pode ser dividida entre os chamados aprendizados supervisionados e não-supervisionados. O primeiro, de acordo com Cunningham, Cord e Delany (2008), trata do mapeamento do relacionamento entre as *features* do conjunto X com o conjunto da variável alvo y, e a aplicação desse mapeamento para prever os *outputs* para qualquer conjunto similar a X aplicado em $f()$.

Os problemas de ML podem ser de dois tipos: classificação e regressão. Problemas de classificação são aqueles que queremos uma variável y categórica, ou seja, quando y é uma variável com alguma representação qualitativa. Já problemas de regressão são problemas onde a variável prevista y é uma variável contínua, ou seja, um número real. (Hastie; Tibshirani; Friedman, 2009)

Matematicamente, todo problema de aprendizado de máquina pode ser descrito como um problema de otimização estatística. Os modelos de ML recebem um conjunto de pares ordenados $D = \{(x_i, y_i) \mid i \in I\}$ de treino, onde x_i pertence ao conjunto de variáveis de entrada X, e y_i pertence ao conjunto de saída Y. Como Bishop (2006) aponta, o objetivo é encontrar uma função $f : X \to Y$, de forma que minimize o risco esperado $R(f)$, dada pela Equação 1 abaixo.

$$R(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\, L(y, f(x))\,\big] \tag{1}$$

Em que:

a) $\mathbb{E}_{(x,y)\sim\mathcal{D}}$ é o valor esperado com relação à distribuição real dos dados $P(X, Y)$;

b) $L(y, f(x))$ é a função de perda que mede o “custo” do erro da previsão $f(x)$, sendo ela uma função variável de acordo com o tipo de problema de ML a ser resolvido.

Ao treinar um modelo, a distribuição real $P(X, Y)$ não é conhecida, e isso impossibilita o cálculo do risco esperado. A solução para isto é calcular o risco empírico do modelo, dado pela Equação 2 abaixo. Partindo dele, é possível estimar o erro médio daquela população partindo de amostras dos dados do conjunto D.

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} L\big(y_i, f(x_i)\big) \tag{2}$$

Em que:

a) n é o tamanho da amostra dos dados usados no treinamento do modelo.

Em suma, pode-se definir que a função matemática final dos modelos de aprendizado de máquina seguem a Equação 3. Os diferentes algoritmos ditam (ou

são configurados) para definir uma função de perda L, a estrutura paramétrica de f e a estratégia de otimização usada para minimizar $f*$.

$$f^* = min(R_n(f)) \tag{3}$$

2.3.1. Aplicação computacional

Os algoritmos de ML aplicados neste trabalho são em grande maioria parte da biblioteca *scikit-learn* baseada em *Python*, que contém diversos algoritmos (também chamados pela biblioteca de estimadores) de ML prontos para serem modelados de acordo com o problema em questão. Os algoritmos dessa biblioteca recebem hiperparâmetros, que são parâmetros definidos pelo usuário para a aplicação do algoritmo (Scikit-Learn, 2025).

Para a aplicação do modelo *XGBoost* é necessário o uso de sua biblioteca própria, de mesmo nome. Esse módulo é uma implementação otimizada e distribuída do algoritmo de *Gradient Boosting* da *scikit-learn*. Embora siga os mesmos princípios fundamentais, ela introduz várias melhorias que a tornam notavelmente mais rápida e, em muitos casos, mais precisa (XGBoost, 2024).

2.3.2. Regressão Logística

A regressão logística é um modelo estatístico amplamente utilizado para problemas de classificação binária. Diferentemente da regressão linear, a regressão logística estima a probabilidade de uma observação pertencer a uma classe, utilizando a função logística (sigmoide) para mapear qualquer valor real para um intervalo entre 0 e 1.

A estrutura de f para a regressão logística é a de uma combinação linear z_i, apresentada na Equação 4, transformada em uma probabilidade p_i de acordo com a função sigmoide $\sigma(z_i)$, mostrada na Equação 5.

$$z_i = x_i^T \beta \tag{4}$$

$$p_i = \sigma(z_i) = \frac{1}{1 + \exp(-z_i)} \tag{5}$$

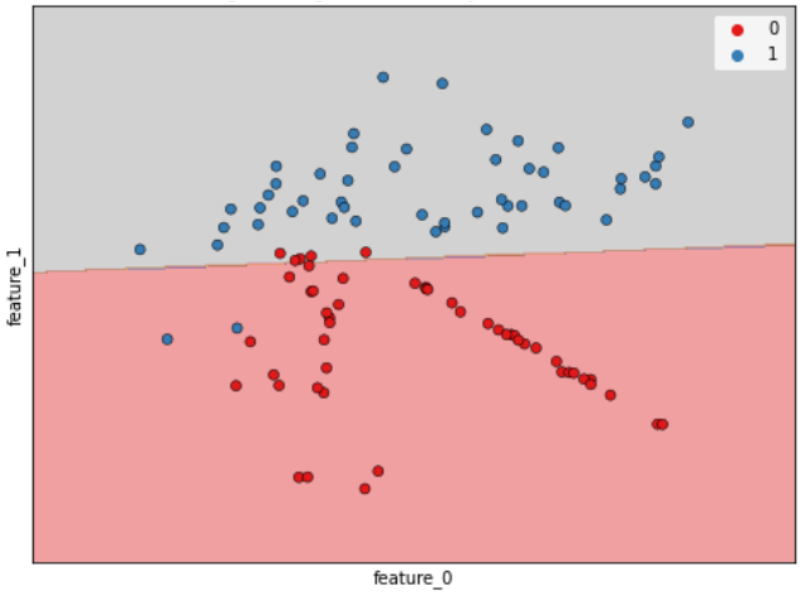
Com isso, temos que a função de perda deste algoritmo é dada pela Equação 6:

$$\hat{R}_n(\beta) = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log p_i + (1-y_i)\log(1-p_i)\right] \quad (6)$$

Segundo James et al. (2013), o modelo é especialmente útil quando o objetivo é estimar a chance de ocorrência de um evento, como a inadimplência de um cliente ou a presença de uma doença. Isso porque o algoritmo cria uma barreira – ou limite – de decisão para os valores da função sigmoide, sendo esse limite em geral 0,5.

A Figura 1 mostra uma representação gráfica do limite de decisão criado por um modelo de Regressão Logística.

Figura 1 - Regiões de Decisão de um modelo de regressão logística



Fonte: DataCamp (2023)

O algoritmo de Regressão Logística usado neste trabalho vem da biblioteca *scikit-learn.* Para sua regulagem, ele contém os hiperparâmetros explicados no Quadro 2 abaixo.

Quadro 2 - Hiperparâmetros do algoritmo de Regressão Logística da *scikit-learn*

Hiperparâmetro	Descrição
penalty	Especifica a norma a ser usada na penalização (regularização) para prevenir o sobreajuste (overfitting). As opções comuns são 'l1' (que pode zerar alguns coeficientes, útil para seleção de features), 'l2' (que diminui o valor de todos os coeficientes) e 'elasticnet' (uma combinação de L1 e L2).
C	É o inverso da força da regularização; um valor positivo e pequeno indica uma regularização mais forte, enquanto um valor grande indica uma regularização mais fraca. Controla o trade-off entre a complexidade do modelo e o ajuste aos dados.
solver	Define o algoritmo a ser usado no problema de otimização. A escolha depende do tipo de penalidade e do volume de dados. Exemplos incluem 'liblinear' (bom para pequenos datasets), 'lbfgs', 'sag' e 'saga' (melhores para grandes datasets).
max_iter	O número máximo de iterações para os solvers convergirem. Aumentar este valor pode ser necessário se o modelo não conseguir convergir com o padrão.

Fonte: Scikit-Learn (2025)

2.3.3. Random Forest

O Random Forest é um método de *ensemble learning* que utiliza múltiplas árvores de decisão para melhorar a capacidade preditiva e reduzir o risco de *overfitting*. Conforme Breiman (2001), cada árvore é construída a partir de uma amostra aleatória dos dados (*bootstrap sample)* e, para cada divisão, apenas um subconjunto aleatório de *features* é considerado.

A estrutura de f nesse algoritmo é dada pela Equação 7, onde cada T_m é uma árvore de decisão, treinada em uma amostra de *bootstrap,* pertencente ao conjunto total de árvores M.

$$f(x) = \frac{1}{M} \sum_{m=1}^{M} T_m(x) \tag{7}$$

Como função de perda (L) o algoritmo tem três diferentes possibilidades: Gini ou entropia para problemas de classificação, e MSE para problemas de regressão. As Equações 8 e 9 apresentam as funções possíveis para classificação, onde p_k é a proporção de exemplos da classe k no nó. Já a Equação 10 apresenta a fórmula do MSE utilizado em cada nó das árvores de regressão.

$$G = 1 - \sum_{k=1}^{K} p_k^2 \tag{8}$$

$$H = -\sum_{k=1}^{K} p_k \log p_k \tag{9}$$

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{10}$$

O algoritmo de *Random Forest* usado neste trabalho vem da biblioteca *scikit-learn.* Para sua regulagem, ele contém os hiperparâmetros explicados no Quadro 3 abaixo.

Quadro 3 - Hiperparâmetros do algoritmo de *Random Forest* da *scikit-learn*

Hiperparâmetro	Descrição
n_estimators	O número de árvores de decisão na floresta. Geralmente, um número maior de árvores aumenta o desempenho e a estabilidade das previsões, mas também aumenta o custo computacional.
criterion	A função para medir a qualidade de uma divisão. Para classificação, as opções são 'gini' (Impureza de Gini) e 'entropy' (Ganho de Informação). Para regressão, as opções são 'squared_error' (erro quadrático médio) e 'absolute_error' (erro absoluto médio).
max_depth	A profundidade máxima de cada árvore de decisão. Se não especificado, as árvores crescem até que as folhas sejam puras ou contenham poucas amostras. Limitar a profundidade é uma forma eficaz de combater o sobreajuste.
min_samples_split	O número mínimo de amostras necessárias para dividir um nó interno de uma árvore. Aumentar este valor pode tornar o modelo mais conservador e prevenir o sobreajuste.
min_samples_leaf	O número mínimo de amostras que devem estar em um nó folha (um nó terminal). Garante que cada "voto" final na floresta seja baseado em um número mínimo de exemplos.
max_features	O número de características a serem consideradas ao procurar a melhor divisão. Introduz aleatoriedade e diversidade entre as árvores, o que geralmente melhora o desempenho do modelo final.

Fonte: Scikit-Learn (2025)

2.3.4. Gradient Boosting

O *Gradient Boosting* é uma técnica que combina modelos fracos — geralmente árvores de decisão rasas — de forma sequencial, onde cada novo

modelo tenta corrigir os erros do anterior (Bishop, 2006). A Equação 11 apresenta a função F do algoritmo, onde $h_m(x)$ são as árvores rasas, γ_m são pesos das árvores, e ν é o hiperparâmetro *learning rate* (ver Quadro 5).

$$F_M(x) = \sum_{m=1}^{M} \nu \, \gamma_m \, h_m(x) \tag{11}$$

A função de perda segue as métricas usadas em Random Forest, dado o problema analisado. Friedman (2001) formalizou esse algoritmo como uma generalização do boosting, utilizando o gradiente negativo da função de perda como pseudo-resíduo ($r_i m$), ajustando cada função de perda L para se aproximar de $r_i m$, dado pela Equação 12.

$$r_{im} = - \left. \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} \right|_{F_{m-1}} \tag{12}$$

Com isso, o algoritmo segue a estratégia de otimização de ajuste de h_m aos resíduos $r_i m$, terminando a iteração pela atualização de F_m, dada pela Equação 13 abaixo.

$$F_m(x) = F_{m-1}(x) + \nu \, \gamma_m \, h_m(x) \tag{13}$$

O algoritmo de Gradient Boosting usado neste trabalho vem da biblioteca scikit-learn. Para sua regulagem, ele contém os hiperparâmetros explicados no Quadro 5 abaixo.

Quadro 4 - Hiperparâmetros do algoritmo de *Gradient Boosting* da *scikit-learn*

Hiperparâmetro	Descrição
n_estimators	O número de estágios de boosting (árvores) a serem executados. Diferente do Random Forest, um n_estimators muito alto pode levar ao sobreajuste, sendo crucial um ajuste fino em conjunto com a learning_rate.
learning_rate	A taxa de aprendizado que pondera a contribuição de cada árvore. Valores menores exigem um número maior de n_estimators para um bom desempenho, mas resultam em modelos mais robustos e com melhor capacidade de generalização.
max_depth	A profundidade máxima das árvores de decisão individuais. Geralmente, para Gradient Boosting, são usadas árvores "fracas" com baixa profundidade (ex: 3 a 8) para evitar que cada árvore se especialize demais.
subsample	A fração de amostras a ser usada para ajustar as árvores individuais. Se for menor que 1.0, o método é chamado de Stochastic Gradient Boosting, o que ajuda a reduzir a variância e prevenir o sobreajuste.

Fonte: Scikit-Learn (2025)

2.3.5. XGBoost

O *XGBoost* (*Extreme Gradient Boosting*) é uma versão otimizada do *Gradient Boosting* que oferece maior desempenho por meio de paralelização, regularização e outras melhorias computacionais. Sua estrutura paramétrica $f(x)$ segue a mesma do *Gradient Boosting*.

Chen e Guestrin (2016) desenvolveram essa biblioteca com foco em velocidade e performance, resultado de uma combinação de regularização L1/L2 e uso de histogramas para aceleração de cálculos. Essa regularização é dada pelo componente Ω da Equação 12, que apresenta a função de perda do algoritmo. L neste caso segue as funções de perda do *Random Forest* e do *Gradient Boosting*, que dependem do tipo de problema de ML enfrentado.

$$\mathrm{Obj} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t) \tag{12}$$

A função de penalização das árvores é dada pela Equação 13.

$$\Omega(f_t) = \gamma T_t + \frac{1}{2}\lambda \sum_{j=1}^{T_t} w_j^2 \tag{13}$$

Em que:

a) T_t é o número de folhas na árvore f_t;
b) w_j é o peso da folha j;
c) γ é a penalidade pelo número de folhas;
d) λ é a regularização L2 nos pesos das folhas.

A estratégia de otimização neste modelo, segue uma aproximação de segunda ordem da função de perda em relação às predições realizadas (Chen; Guestrin, 2016). Em cada nó é calculado o gradiente (G) e o hessiano (H) do conjunto de observação I_j, em seguida o peso ótimo (w_j) é calculado para a filha, usando a regularização L2 mostrada anteriormente. As equações a seguir apresentam esses parâmetros de otimização.

$$g_i = \frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{y}_i},\ h_i = \frac{\partial^2 L(y_i, \hat{y}_i)}{\partial \hat{y}_i^2} \tag{14), (15}$$

$$G_j = \sum_{i \in I_j} g_i, \quad H_j = \sum_{i \in I_j} h_i \tag{16), (17}$$

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{18}$$

Por fim, o ganho do nó para a previsão é dado pela Equação 19, e o algoritmo escolhe o caminho onde esse ganho é maximizado.

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{19}$$

O algoritmo de *XGBoost* usado neste trabalho vem de uma biblioteca específica para o algoritmo, chamada também de *xgboost.* Os hiperparâmetros que podem ser definidos para esse algoritmo estão no Quadro 6 abaixo.

Quadro 5 - Hiperparâmetros do algoritmo de *XGBoost* da *xgboost*

Hiperparâmetro	Descrição
n_estimators	O número de estágios de boosting (árvores) a serem executados. Diferente do Random Forest, um n_estimators muito alto pode levar ao sobreajuste, sendo crucial um ajuste fino em conjunto com a learning_rate.
learning_rate	A taxa de aprendizado que pondera a contribuição de cada árvore. Valores menores exigem um número maior de n_estimators para um bom desempenho, mas resultam em modelos mais robustos e com melhor capacidade de generalização.
max_depth	A profundidade máxima das árvores de decisão individuais. Geralmente, para Gradient Boosting, são usadas árvores "fracas" com baixa profundidade (ex: 3 a 8) para evitar que cada árvore se especialize demais.
subsample	A fração de amostras a ser usada para ajustar as árvores individuais. Se for menor que 1.0, o método é chamado de Stochastic Gradient Boosting, o que ajuda a reduzir a variância e prevenir o sobreajuste.

Fonte: *XGBoost Developers* (2025)

2.3.6. Classificação Multissaída *(Multi Output)*

A classificação multiclasse-multissaída (multi-output) caracteriza-se como uma abordagem de aprendizado supervisionado em que um mesmo conjunto de dados de entrada é utilizado para gerar múltiplas saídas de classificação. Na prática, essa técnica possibilita a definição de um único estimador capaz de atuar de forma simultânea em diversas tarefas de classificação correlacionadas, promovendo maior eficiência e coerência entre os resultados obtidos (Scikit-Learn, 2025).

2.3.7. Avaliação

As formas de avaliação para cada tipo de problema de previsão variam. Nessa seção serão discutidas as métricas, ferramentas e métodos possíveis para a avaliação dos diferentes algoritmos apresentados acima.

2.3.7.1. Em Modelos de Regressão

As principais métricas de avaliação usadas em um modelo de regressão são: Erro Médio Absoluto (*Mean Absolute Error* (MAE)) e Erro Médio Quadrático (*Mean Squared Error* (MSE)). O MAE é a média dos valores absolutos dos erros, dado pela Equação 20, sendo uma métrica robusta para outliers dado seu caráter médio (Bishop, 2006).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - f(x_i)| \tag{20}$$

O MSE, é o erro médio quadrático dentre os valores previstos e os valores esperados. Hastie, Tibshirani e Friedman (2009) descrevem essa métrica de acordo com a Equação 21:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \tag{21}$$

Por fim, para medir a efetividade de algoritmos em prever a distribuição dos valores originais, a métrica de Distância de Wasserstein (Wass) será usada. Ela é uma métrica que quantifica o "custo mínimo de transporte" necessário para transformar uma distribuição de probabilidade μ em outra ν. Em linhas gerais, mede o esforço necessário para redistribuir a massa de uma densidade até coincidir com a outra (Vallender, 1974). A Equação 22 apresenta a fórmula simplificada da métrica, considerando o custo médio de transporte:

$$W_1(\mu,\nu) = \int_{-\infty}^{+\infty}|F(x) - G(x)|, dx \tag{22}$$

Em que $F(x)$ e $G(x)$ são as funções de distribuição acumulada de μ e ν, respectivamente.

2.3.7.2. Em Modelos de Classificação

As principais métricas de avaliação usadas em um modelo de classificação são: Acurácia e *F1 score*. A acurácia mede a proporção de previsões corretas em relação ao total de casos, variando de 0 a 1 (DataCamp, 2023).

O *F1 score* é uma métrica que avalia o equilíbrio entre a precisão e o *recall* do modelo, variando de 0 a 1, sendo 1 um modelo com desempenho perfeito. Essa métrica é calculada pelas equações:

$$Precisao = \frac{VP}{VP + FP} \tag{23}$$

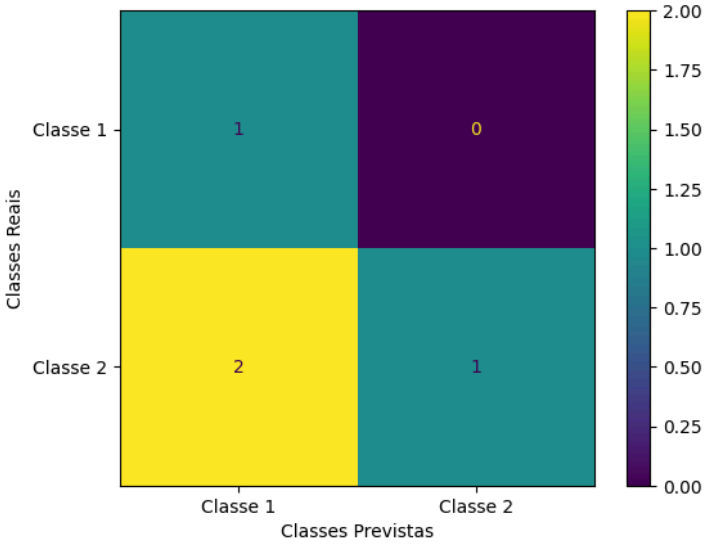
$$Recall = \frac{VP}{VP + FN} \tag{24}$$

$$F1 = 2\frac{Precisao * Recall}{Precisao + Recall} \tag{25}$$

Em que VP são Verdadeiros positivos, FN são Falsos negativos e FP são Falsos positivos. VPs correspondem a quando a classe foi corretamente prevista como positiva. Já os FPs, quando o modelo previu positivo, mas a classe real era negativa, e FNs, quando o modelo previu negativo, mas a classe real era positiva. Adicionalmente podem ser considerados também os Verdadeiros Negativos (VN), quando o modelo corretamente identificou a classe como negativa, porém não contabilizada na métrica.

Modelos de classificação podem ser avaliados também através de uma ferramenta chamada matriz de confusão. Essa ferramenta apresenta graficamente uma matriz quadrática em que o eixo ordenado aponta as classes verdadeiras e o eixo das abscissas aponta as classes previstas. Ela deve ser aplicada em um conjunto de teste da variável alvo, para que seja possível esse mapeamento. Quando construída, cada valor dentro dessa matriz representa o número de previsões feitas para a classe x_i mas que na verdade pertencem à classe y_j (Scikit-Learn, 2025). A Figura 2 abaixo representa uma dessas matrizes aplicadas no trabalho.

Figura 2 - Exemplo de uma matriz de confusão



A última métrica utilizada em algoritmos de classificação é a *Hammimg Loss*. Ela mede a fração de rótulos que foram incorretamente previstos em relação ao total de rótulos possíveis, ou seja, pode ser interpretada como uma taxa de erros. O valor da Hamming Loss varia entre 0 e 1. Quanto mais próximo de 0, melhor é o desempenho do classificador, indicando uma perda menor. (Tsoumakas; Katakis, 2007) A métrica é calculada pela equação 26 abaixo:

$$HammingLoss(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \Delta Z_i|}{|L|} \quad (26)$$

Onde $|D|$ é o número total de exemplos no conjunto de dados.$|L|$ é o número total de rótulos (classes) no problema de classificação. Y_i é o conjunto de rótulos verdadeiros para o exemplo i. Z_i é o conjunto de rótulos preditos pelo classificador H para o exemplo i. Δ representa a diferença simétrica entre os dois conjuntos (Y_i e Z_i). O valor $|Y_i \Delta Z_i|$ representa o número de rótulos incorretamente preditos para o exemplo i (os rótulos que estão em Y_i mas não em Z_i, ou que estão em Z_i mas não em Y_i).

2.3.7.3. Avaliações de diferentes modelos

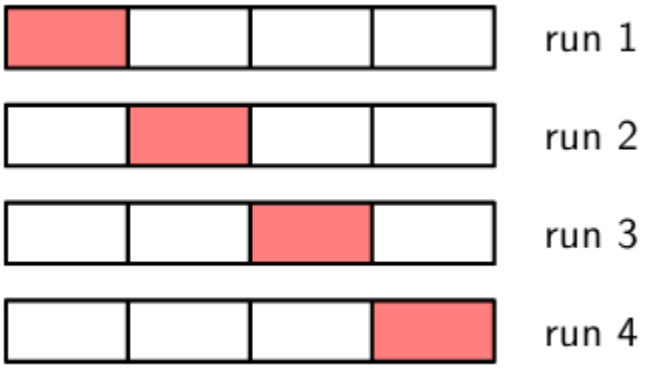
Com as métricas descritas acima é possível avaliar diferentes modelos de ML a partir de um método chamado de validação cruzada (*Cross Validation* (CV)). Essa técnica consiste em separar o conjunto de dados em múltiplas partições, ou *folds*, e executa várias rodadas de treinamento e teste. Em cada rodada, o modelo é

treinado em uma combinação de partições e testado na partição restante através das métricas apresentadas acima. O desempenho final é obtido por meio da média dos resultados de cada rodada, proporcionando uma estimativa mais robusta do desempenho do modelo e reduzindo a variabilidade causada por divisões específicas dos dados (Bishop, 2006).

Neste trabalho, a técnica de CV utilizada será o *KFold* em que o conjunto de dados é dividido em *k* subconjuntos de tamanho aproximadamente igual. O modelo é então treinado *k* vezes, cada vez utilizando *k – 1* subconjuntos como dados de treino e o subconjunto restante como dados de teste.

A Figura 3 apresenta uma forma visual da aplicação do *KFold*. No exemplo, o conjunto foi dividido em 4 subconjuntos, e em cada iteração (*run)* o conjunto de teste (em vermelho) é um subconjunto diferente, garantindo que todos os 4 conjuntos participem do treinamento do modelo.

Figura 3 - Ilustração de uma validação cruzada



Fonte: Bishop, 2006

2.3.7.4. Importância de Features

Os modelos baseados em árvores de decisão utilizam métricas de impureza para avaliar a qualidade das divisões durante o processo de construção da árvore. O objetivo é maximizar a homogeneidade dos nós resultantes, ou seja, minimizar a impureza. Entre as métricas mais empregadas, destaca-se o índice de Gini, utilizado por padrão nos classificadores do scikit-learn (Pedregosa et al., 2011). O cálculo deste índice foi dado pela Equação 8 quando tratado do seu uso no algoritmo de *Random Forest.*

A impureza do nó mede o grau de mistura das classes nele. Quanto mais homogêneo for o nó, menor será sua impureza, e sua homogeneidade é definida quando há a concentração de uma única classe nele. Assim, quanto mais equilibradas as proporções das classes, maior será o valor de Gini. Um nó totalmente puro apresenta $Gini = 0$, enquanto um nó com duas classes igualmente representadas apresenta $Gini = 0.5$.

Para a construção da árvore visando o objetivo de redução das impurezas, os algoritmos se baseiam na variação das impurezas (Δi). Hastie, Tibshirani e Friedman (2009) descrevem essa variação de acordo com a Equação 27 abaixo:

$$\Delta i = i_{\text{pai}} - (w_{\text{esq}} \times i_{\text{esq}} + w_{\text{dir}} \times i_{\text{dir}}) \tag{27}$$

Em que i_{pai} é a impureza do nó antes da divisão, i_{esq} e i_{dir} são as impurezas dos nós filhos, e w_{esq} e w_{dir} representam as proporções de amostras em cada filho.

A importância das variáveis em modelos de árvores é derivada da contribuição de cada atributo para a redução total da impureza. Essa métrica, conhecida como Redução Média da Impureza (*Mean Decrease in Impurity* (MDI)), é expressa pela soma ponderada das reduções de impureza atribuídas a cada variável, normalizada sobre todas as árvores do modelo:

$$FI_j = \frac{1}{T} \sum_{t=1}^{T} \sum_{n \in \text{nos}_j} \frac{N_n}{N} \Delta i_n \tag{28}$$

Em que FI_j representa a importância da variável j, T é o número de árvores, N_n é o número de amostras no nó n, N é o número total de amostras, e Δi_n é a redução de impureza gerada pela divisão no nó n. Esse método é eficiente, mas tende a superestimar a importância de variáveis contínuas ou com muitos valores distintos (Pedrogosa et al., 2011).

Além do MDI, existem outras formas de avaliar a importância das *features,* aplicados principalmente à algoritmos não baseados em árvores. Uma delas é a importância por permutação, que avalia o impacto da aleatorização de uma variável sobre o desempenho preditivo do modelo. O procedimento consiste em permutar aleatoriamente os valores de uma variável em um conjunto de validação e medir a variação em uma métrica de desempenho. A diferença entre o erro antes e depois

da permutação é interpretada como a importância da variável, como aponta a Equação 29:

$$FI_j = \text{Erro} * \text{perm}(j) - \text{Erro} * \text{original} \tag{29}$$

Se a permutação de uma variável causa grande degradação no desempenho, ela é considerada importante, caso contrário, tem pouca influência sobre o modelo (Breiman, 2001). Esse método é mais robusto, pois não depende da estrutura da árvore, mas sim do impacto real da variável sobre o resultado, porém tem alto custo computacional para sua aplicação.

2.4. MODELAGEM DIMENSIONAL DE DADOS

A modelagem dimensional é descrita por Kimball e Ross (2013) como uma abordagem para a organização de dados em sistemas de apoio à decisão, especialmente em *data warehouses* (DW). Seu objetivo principal é facilitar a análise e a interpretação dos dados por usuários de negócio, por meio de uma estrutura intuitiva e de alto desempenho para consultas analíticas.

A estrutura central da modelagem dimensional de Kimball e Ross (2013) é composta por tabelas fato e tabelas dimensão, geralmente organizadas no formato de estrela (star schema) ou floco de neve (*snowflake schema*). As tabelas fato armazenam os eventos quantitativos do negócio (como vendas, transações ou cliques), contendo medidas numéricas e chaves estrangeiras que se conectam às dimensões. Já as tabelas dimensão representam os atributos descritivos relacionados aos fatos (como produto, cliente, tempo ou localização), possibilitando a segmentação e agregação das métricas conforme diferentes perspectivas analíticas (Kimball; Ross, 2013).

Antes de os dados serem carregados nas tabelas fato e dimensão, eles passam por uma camada intermediária conhecida como *staging area*. A *staging area* é a área de preparação dos dados, geralmente temporária, onde ocorre a extração, limpeza, padronização e integração dos dados oriundos de múltiplas fontes. Isso garante a consistência e qualidade dos dados antes da carga final no modelo dimensional (Kimball; Ross, 2013).

2.5. CONSIDERAÇÕES FINAIS DO CAPÍTULO

A revisão teórica realizada acima tem como principal finalidade fundamentar e direcionar os passos para o objetivo geral do trabalho. Dessa forma, a seguir serão apresentadas as principais informações trazidas da pesquisa bibliográfica e sua relação com o objetivo geral da pesquisa.

A partir das definições de desastres naturais do Manual de Desastres (SEDEC, 2003) é possível entender a fundo o que caracterizam os desastres naturais que assolam o estado de Santa Catarina. Em cima disso, suas causas indicam os indicadores e pontos de atenção meteorológicos e hidrológicos que devem ser observados para tentar prever e controlar tais desastres.

Fenômenos do grupo hidrológico se relacionam principalmente com questões da capacidade do solo em absorver a água das chuvas do local, levando em conta outras características específicas a cada um dos eventos, como: a impermeabilização e deficiência de drenagem em áreas urbanas para os alagamentos; o relevo declinado do local para as enxurradas; e os súbitos despejos de água em um rio para as Inundações (SEDEC, 2003).

Já fenômenos do grupo meteorológico são afetados principalmente pelo movimento das massas de ar na atmosfera, sendo impactados por movimentos que ocorrem mesmo em alto mar. Dentre as especificidades das causas de cada evento pode-se destacar: o rápido movimento de massas de ar e consequente variação de pressão e temperatura para as chuvas intensas; as altíssimas velocidades causadas por aquecimentos em alto mar dos ciclones; as baixas temperaturas e precipitação para o granizo; e as variações de pressão para os vendavais. Próximo a este grupo temos a estiagem, fenômeno climatológico, causado principalmente por longos períodos sem chuva e de altas temperaturas (SEDEC, 2003).

O entendimento dessas causas é o fator direcionador para a modelagem dos dados visando sua aplicação científica. Kimball e Ross (2013) destacam que as tabelas fato e toda a estrutura dimensional que as apoiam começam com a pergunta de “o que queremos medir?”, e contexto deste trabalho, os registros numéricos das causas dos eventos de desastres são a resposta para essa pergunta.

Por fim, os entendimentos das funcionalidades dos modelos de *machine learning* apresentados principalmente por Bishop (2006) indicam como é possível usar dessas tabelas para a previsão dos itens. O autor também apresenta os

métodos de validação de seus resultados e melhoria do desempenho pelos métodos de *cross validation* e matrizes de confusão*,* e métricas robustas como o *f1 score, MAE, RMSE,* entre outras.

3. MÉTODO

3.1. TIPO DE PESQUISA

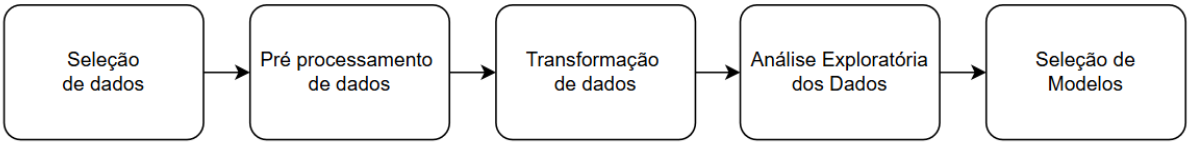
O presente estudo tem como objetivo propor um modelo preditivo baseado em algoritmos de ML para prever a demanda de itens de assistência humanitária. Dessa forma pode-se dizer que a natureza da pesquisa é aplicada e quantitativa, já que busca gerar conhecimento para a solução de problemas e dificuldades no planejamento da SEDEC/SC frente à solicitação de itens humanitários (Silva; Menezes, 2005), se embasando em dados estruturados e estatísticos para essa aplicação (Mattar, 2001).

Em relação à sua caracterização pelo objetivo, Gil (2002) aponta que pesquisas que visam identificar novas percepções sobre um problema podem ser classificadas como pesquisas exploratórias. Já quanto aos procedimentos empenhados neste trabalho – estudo histórico, investigando fenômenos naturais – a pesquisa pode ser classificada como um estudo de caso (Miguel, 2006).

3.2. PROCEDIMENTOS METODOLÓGICOS

A sequência de procedimentos realizados neste estudo segue uma adaptação da metodologia de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* (KDD)) proposto por Fayyad, Piatetsky-Shapiro e Smyth (1996). A grande adaptação dessa metodologia para o foco deste trabalho está na separação da 4° etapa do método (Mineração de Dados) em duas: Análise Exploratória dos Dados e Seleção de Modelos. Com isso, temos que o fluxo de procedimentos segue a Figura 4:

Figura 4 - Fluxograma do trabalho



a) Seleção de dados: a identificação e extração dos dados relevantes de diversas fontes para a realização do estudo.
b) Pré-processamento de dados: etapa de limpeza e preparação dos dados, ajustando seus tipos, inconsistências, valores ausentes e outros

empecilhos nos conjuntos selecionados, de acordo com a modelagem de Kimball (2013).

c) Transformação de dados: etapa transformação dos conjuntos selecionados em formas que agreguem as variáveis desejadas em formas otimizadas e aceitas pelos modelos a serem aplicados.
d) Análise Exploratória dos Dados: etapa de análises gerais dos dados transformados, visando identificar *outliers*, comportamentos não naturais, e outros possíveis pontos de atenção nos dados.
e) Seleção de Modelos: a aplicação de algoritmos de ML nos dados transformados para cada etapa, avaliando suo desempenho e selecionando o melhor modelo para o respectivo problema.

3.3. DELIMITAÇÕES

A primeira grande limitação deste trabalho refere-se à disponibilidade e qualidade dos dados. Para a construção do *pipeline* preditivo voltado à previsão de demanda de itens, foram utilizados dados públicos disponíveis na internet e dados fornecidos pela SEDEC/SC. No entanto, em alguns casos, como nas informações de urbanização do IBGE, não foi possível localizar dados específicos das áreas urbanizadas dos municípios de Santa Catarina, o que dificultou a avaliação da impermeabilização do solo. Também não foram encontrados dados sobre os tipos de solos específicos para cada cidade.

Em razão dessas limitações, as previsões de eventos hidrológicos foram consideravelmente prejudicadas, já que seus fatores causais estão ligados à esses dados, dificultando sua previsão, levando estes eventos a serem desconsiderados no treinamento dos modelos preditivos.

A segunda limitação relevante diz respeito à capacidade de processamento computacional disponível para a execução do pipeline. Os métodos mencionados na Seção 4 requerem alto poder de processamento, o qual se torna ainda mais demandado conforme aumenta o volume dos dados utilizados. Com isso, optou-se por utilizar os algoritmos computacionais sem ajustes profundos em seus hiperparâmetros.

Por fim, o presente trabalho teve de ser limitado à proposta de um modelo preditivo para a solicitação de itens de assistência humanitária sem uma devida aplicação prática em dados do clima projetados pelo INPE. Isso se deu pela

impossibilidade de conexão com os dados de projeções climáticas por meio da API do Portal de Mudanças Climáticas no Brasil do INPE/MCTI.

4. ESTUDO DE CASO

Esta seção apresenta os desenvolvimentos da pesquisa, seguindo o fluxograma da Figura 4 e as atividades adaptadas da metodologia KDD. Cada tópico abaixo descreve uma das partes descritas no método, destrinchando as atividades executadas em sua completude, iniciando-se pela apresentação do modelo proposto.

4.1. MODELO PROPOSTO

A previsão da demanda por itens humanitários exige vários passos preliminares complexos. As quantidades e tipos de itens solicitados variam significativamente de acordo com o tipo de evento de desastre. Cada tipo de evento têm causas distintas (detalhadas na Seção 2.2.3), que podem ser identificadas por meio de medições meteorológicas e sensores climáticos. Além dessa complexidade, há a incerteza inerente à ocorrência do próprio evento.

Neste contexto, modelos computacionais buscam atenuar essas incertezas através de previsões realizadas pelos algoritmos de ML apresentados na seção 2.3, sendo eles indispensáveis durante os períodos de preparação para um desastre. A aplicação desses modelos neste período concentram-se principalmente nas decisões de pré-posicionamento de insumos, que envolvem armazenar suprimentos de emergência em locais estratégicos antes da ocorrência de um desastre para melhorar a resposta (Balcik; Bozkir; Kundakcioglu, 2016).

Porém, como apontado anteriormente, cada variável de decisão importante para a preparação de resposta a desastres está sujeita a diferentes causas, e como aponta Bishop (2006), os modelos matemáticos dos algoritmos de ML recebem um conjunto Y único e o usam como base para criar suas funções de perda e otimização.

Logo, prever os itens demandados partindo apenas de dados meteorológicos, por exemplo, não se mostra eficaz. Dessa forma o modelo proposto se trata de um *pipeline* preditivo, onde os *outputs* do modelo anterior são *inputs* para os modelos subsequentes. Com isso, é possível utilizar diferentes bases de dados para o treinamento de cada modelo e garantir que o resultado de f^* seja o melhor para aquela etapa do *pipeline*.

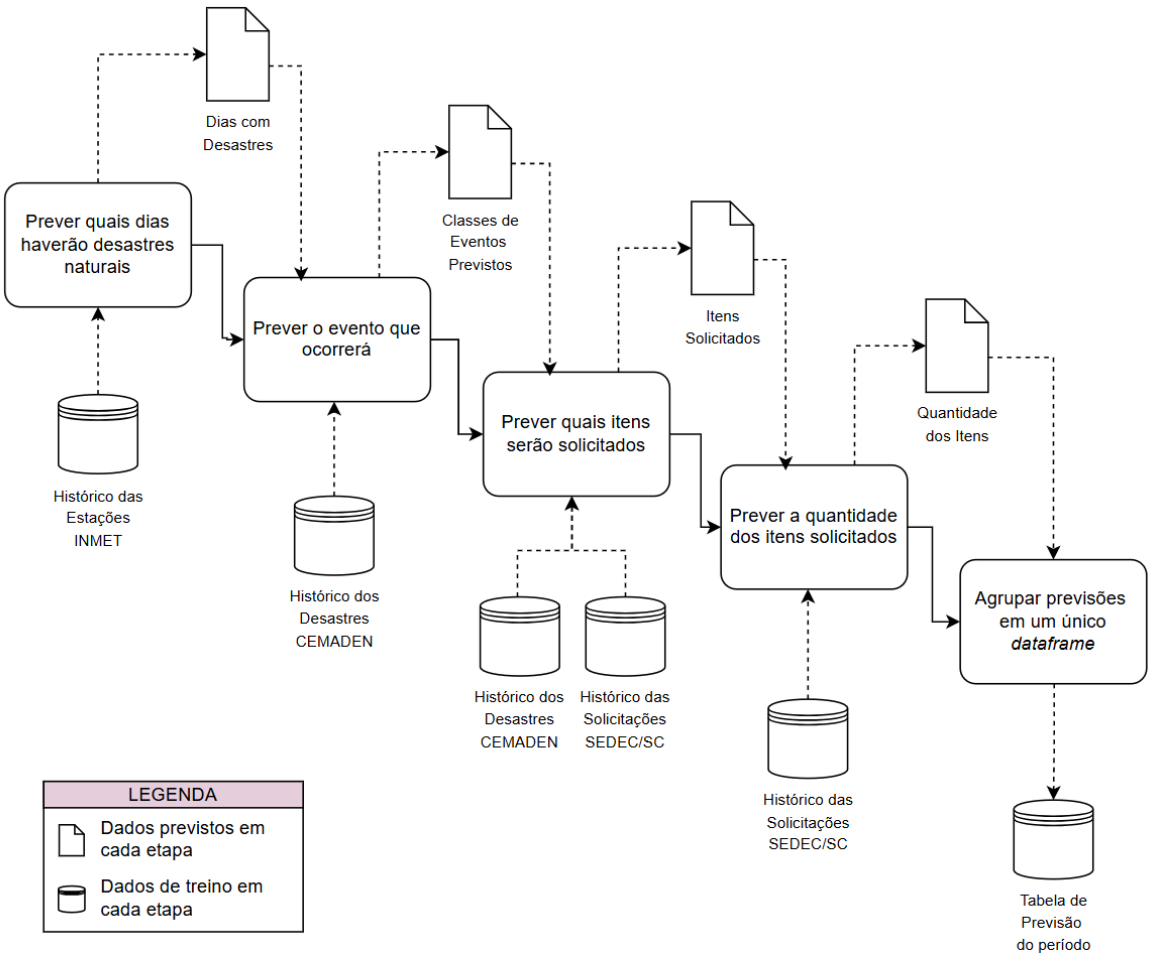
Tais etapas foram definidas com base nas variáveis que impactam sobre a solicitação de itens. As quantidades dos itens solicitados dependem do tipo do item, as quais dependem da classe do desastre previsto que podem ou não ocorrer dentro de um período. Dessa forma foram definidas as 2 macroetapas do *pipeline,* cada uma contendo 2 etapas:

I. Previsão de eventos de desastre: composta pela previsão de se haverá um desastre de acordo com o período futuro analisado; e o tipo do desastre ocorrido.

II. Previsão de itens de assistência: composta pela previsão de quais itens serão solicitados para a classe do desastre; e a quantidade deste item.

Dessa forma foi possível esquematizar as atividades do *pipeline,* disposto na Figura 5. Na figura são representados também os *outputs* de cada etapa, sendo utilizadas como *inputs* na etapa seguinte, e também os dados de treino usados em cada modelo, detalhados nas seções a seguir.

Figura 5 - *Pipeline* de Previsão



Para cada etapa, foram testados diferentes modelos e sua seleção depende da avaliação das métricas específicas para o respectivo tipo de problema de ML tratado. Cada etapa mostrada no Quadro 6 terá seu processo decisório sobre a seleção descrito na seção 4.6.

Quadro 6 - Etapas do *pipeline* de previsão

Macro Etapa	Etapa	Objetivo	Algoritmos avaliados	Métricas
Previsão de Eventos de Desastres	Haverá desastre	Prever se haverá desastres no dia simulado	*Random Forest, XGBoost, Regressão Logística*	Acurácia, F1 Score
	Classe do desastre	Prever qual tipo de evento poderá ocorrer.	*Random Forest, XGBoost, Gradient Boosting*	Acurácia, F1 Score
Previsão de Itens de Assistência	Quais itens	Prever quais itens serão solicitados	*Random Forest, XGBoost, Regressão Logística*	Acurácia, F1 Score, *Hamming Loss*
	Quantidade de cada item	Prever a quantidade solicitada de cada item	*Random Forest, XGBoost, Gradient Boosting*	MAE, MSE, Wass

4.2. SELEÇÃO DE DADOS

Para a realização deste estudo, foram selecionadas três bases de dados principais, e outras bases secundárias que servem para enriquecer as tabelas fato usadas para os treinamentos dos modelos preditivos.

A primeira base de dados coletada foi a base de dados de solicitações de itens de assistência humanitária, coletada diretamente com a SEDEC/SC. Nesse conjunto de dados, encontra-se uma listagem das solicitações de itens de 2013 a 2025, para diversos desastres naturais. Para este estudo foram disponibilizadas três planilhas eletrônicas contendo esses dados, uma contendo os dados de 2013 a 2019, outra contendo dados de 2020 a 2024, e a última com dados de 2024 à julho de 2025. Esse conjunto mapeia as relações dos eventos de desastre com as solicitações dos itens de assistência.

A segunda base coletada foi um conjunto de dados do Atlas Digital de Desastres no Brasil. Essa base trata-se de uma centralização de dados de desastres naturais por todo o país, agrupando diversas fontes nesta base seguindo uma metodologia própria (MIDR, 2023). Idealizada pelo Banco Mundial e UFSC, por meio do Centro de Estudos e Pesquisas em Engenharia e Defesa Civil – Ceped/UFSC,

atualmente é mantida pelo Ministério da Integração e Desenvolvimento Regional (MIDR). A base consiste em registros de eventos de desastres e números totais de danos humanos, materiais e ambientais causados.

Para os dados meteorológicos, foi coletado o Banco de Dados Meteorológicos (BDMET) do Instituto Nacional de Meteorologia (INMET). O BDMEP disponibiliza dados meteorológicos diários em formato digital, provenientes das séries históricas de 24 estações meteorológicas convencionais da rede do INMET em Santa Catarina. Essas informações seguem os padrões técnicos internacionais estabelecidos pela Organização Meteorológica Mundial. Elas oferecem dados horários da precipitação total, pressão atmosférica, temperatura, temperatura de orvalho, direção e velocidade do vento, umidade e radiação solar, variáveis necessárias para a previsão climática (INMET, 2023).

Além dessas bases, também foram consultados outros conjuntos de dados provenientes de órgãos públicos brasileiros. Destacam-se as bases de organização territorial do IBGE, utilizadas para mapear os dados mencionados anteriormente no nível de municípios e microrregiões definidos pelo instituto. Esses dados foram obtidos já pré-processados por meio da plataforma Base dos Dados, uma organização sem fins lucrativos que centraliza e disponibiliza dados tratados de diversas fontes em ambientes de nuvem (Base dos Dados, 2024).

4.3. PRÉ-PROCESSAMENTO DE DADOS

O pré-processamento de dados consiste no tratamento inicial de cada uma das bases selecionadas, padronizando nomes, excluindo valores inexistentes (*nan*) e corrigindo impurezas que podem afetar o desempenho dos algoritmos. As tabelas tratadas aqui são nomeadas como tabelas *stagins* como define a metodologia de Kimball (ver seção 2.4).

O Quadro 7 abaixo apresenta uma síntese dos tratamentos aplicados no pré-processamento dos dados da pesquisa, junto da estrutura final de cada uma das tabelas de dados coletadas. Em seguida, serão detalhados tais atividades para cada conjunto de dados listados no quadro.

Quadro 7 - Visão geral dos dados selecionados e tratamentos aplicados

Conjunto de Dados	Descrição do tratamento	N° de Linhas	N° de Colunas
Solicitações da SEDEC/SC	Dados manuais e inconsistentes exigiram: exclusão de linhas incompletas, padronização de nomes (eventos e itens), exclusão de colunas monetárias e mapeamento de IDs de localização.	3.475	9
Eventos de Desastre do CEMADEN	O alto volume de dados faltantes foi gerenciado por meio de: filtragem para Santa Catarina, exclusão de colunas monetárias e eventos alheios, mapeamento de IDs de localização e criação de ID de evento.	7.976	34
Meteorológicos do INMET (BDMEP)	Dados já pré-processados exigiram apenas uma reestruturação mínima para as necessidades específicas da pesquisa.	3.229.512	24

4.3.1. Dados de Solicitações da SEDEC/SC

Os dados enviados pela SDEC/SC foram todos inseridos manualmente pelos profissionais da entidade. Logo a base é suscetível a erros sendo necessário uma análise aprofundada para definição de padronizações nos dados. A Tabela 1 apresenta os detalhes das estruturas das 3 tabelas coletadas sem tratamento prévio, apresentando o número de linhas, colunas, a quantidade de entradas nela, a quantidade de entradas *nan*, ou seja, que são nulas no sentido computacional, e as quantidades de eventos, cidades e itens diferentes.

Tabela 1 - Descritivo das tabelas cruas de solicitações

Tabela	N° linhas	N° colunas	Qnt. Entradas	Qnt. Entradas *nan*	Qnt. eventos diferentes	Qnt. cidades diferentes	Qnt. itens diferentes
2013-2019	885	13	11505	2657	23	123	226
2020-2024	2549	15	38235	2341	40	205	21
2024-2025	1632	47	76704	49820	14	165	19

Pelos valores apresentados é possível identificar diversos problemas de dados faltantes em todas as tabelas, bem como uma falta de padronização perante os eventos e itens. Pelos nomes das cidades, também foi possível identificar a despadronização no nome de cada cidade do estado. Com isso, foram aplicadas os seguintes tratamentos:

I. Exclusão de todas as linhas que não continham o item solicitado;

II. Exclusão de tas as linhas que não continham a quantidade do item solicitado;
III. Padronização dos nomes de eventos de desastres de acordo com a base de dados do INMET;
IV. Padronização dos nomes dos itens solicitados para evitar divergências;
V. Exclusão das colunas de valor monetário dos itens;
VI. Renomeação das colunas;
VII. Mapeamento de IDs das colunas de município e microrregião, de acordo com a dimensão de municípios;
VIII. Criação de um identificador único para cada evento registrado;
IX. Aplicação dos tipos de dados corretos às colunas.

Com isso foi gerada a tabela *stagin* de solicitações, com a estrutura apresentada na Tabela 2.

Tabela 2 - Descritivo da tabela *stagin* solicitações

N° linhas	N° colunas	Qnt. Entradas	Qnt. Entradas nan	Qnt. eventos diferentes	Qnt. cidades diferentes	Qnt. itens diferentes
3475	9	31275	0	8	228	19

4.3.2. Dados de Eventos de Desastre do CEMADEN

Os dados coletados do CEMADEN são dados previamente tratados pelo MIDR. Logo, problemas como despadronização e erros de escrita não são vistos na base crua. Dados faltantes tem bastante ocorrência na base, dado que muitos eventos tem informações (como por exemplo, percentual de água de saneamento contaminada, ou quantidade de hospitais danificados) que alguns registros não possuem, o que gera um alto número de entradas faltantes. A estrutura da base extraída diretamente está na Tabela 3 abaixo.

Tabela 3 - Descritivo das tabelas cruas de eventos

N° linhas	N° colunas	Qnt. Entradas	Qnt. Entradas nan	Qnt. eventos diferentes	Qnt. cidades diferentes	Qnt. UFs diferentes
71929	70	5035030	630486	16	5026	28

Por conta do tratamento prévio, o pré-processamento necessário foi apenas reestruturar a base para se encaixar nas necessidades da pesquisa. Os tratamentos aplicados foram:

I. Filtragem das linhas referentes ao estado de Santa Catarina;
II. Exclusão das colunas de valores monetários;
III. Exclusão das linhas de eventos alheios à pesquisa;
IV. Mapeamento de IDs das colunas de município e microrregião, de acordo com a dimensão de municípios;
V. Criação de um identificador único para cada evento registrado;
VI. Aplicação dos tipos de dados corretos às colunas.

Com isso, temos a estrutura final da tabela *stagin* eventos, disposta na Tabela 4. Os dados faltantes ainda representam um alto valor, porém eles provém de colunas que não serão utilizadas dentro do estudo deste trabalho, porém podem servir para análises qualitativas futuras.

Tabela 4 - Descritivo da tabela *stagin* eventos

N° linhas	N° colunas	Qnt. Entradas	Qnt. Entradas nan	Qnt. eventos diferentes	Qnt. cidades diferentes
7976	34	271184	55640	8	295

4.3.3. Dados Meteorológicos do INMET

Os dados das estações meteorológicas do INMET foram coletados previamente tratados via repositório em nuvem da Base de Dados. Logo não foram executados grandes tratamentos na base. Sua estrutura está disposta na Tabela 5. O alto número de linhas se dá tanto pelo grande período de dados da tabela (2003 à 2024) e também pelo caráter de medição horária das estações. Os valores computacionalmente nulos representam próximo a 10% da quantidade total de dados na tabela, apontando baixas falhas de leituras das estações, os quais serão descartados em sua versão de tabela fato.

Tabela 5 - Descritivo da tabela *stagin* metereologia

N° linhas	N° colunas	Qnt. Entradas	Qnt. Entradas nan	Qnt. estações
3.229.512	24	77.508.288	7.030.601	24

4.4. TRANSFORMAÇÃO DE DADOS

As transformações aplicadas às tabelas *staging* visam preparar os dados para seu uso nos modelos preditivos, com o cálculo de indicadores e normais climatológicas para dados meteorológicos. Com esta etapa, são geradas as tabelas fato, de acordo com a Modelagem de Kimball (2013), as quais apresentam o que “de fato se quer medir”. Nessa seção são descritos os passos tomados para a criação das tabelas fato usadas como treinamento para os estimadores. As documentações das tabelas modeladas estão disponíveis nos primeiros Apêndices deste documento.

Para um entendimento inicial quanto às modelagens aplicadas e facilitar a leitura do trabalho, o Quadro 8 apresenta uma visão geral das tabelas modeladas para serem aplicadas no modelo proposto.

Quadro 8 - Visão geral dos dados utilizados no modelo proposto

Nome da Tabela	N° de Linhas	N° de Colunas	Descrição
Fato Solicitações	800	8	Tabela com os registros de eventos que geraram solicitações de itens de assistência à SEDEC/SC e a quantidade de itens diferentes solicitados por cada um desses eventos.
Fato Item Solicitados	2.810	5	Tabela com os itens demandados e suas quantidades por cada ocorrência registrada na base histórica da SEDEC.
Fato Eventos	7.976	11	Tabela com os registros dos eventos de desastres naturais ocorridos no estado de Santa Catarina de 1991 à 2024.
Fato Meteorológico Diário	124.402	18	Tabela com os registros das 24 estações meteorológicas em Santa Catarina diariamente de 2003 à 2024.
Fato Desastres Naturais	5.267	21	Tabela com os registros dos indicadores meteorológicos associados a cada evento.

4.4.1. Tabela Fato Solicitações

A tabela fato solicitações foi construída a partir da sua *staging* correspondente. A estrutura dos dados, que antes contemplava múltiplas linhas tratando os diferentes itens solicitados por conta de um mesmo evento foi

completamente alterada para contemplar apenas os eventos registrados e quantos itens foram solicitados em cada um.

A tabela resultante teve um total de 800 linhas e 8 colunas, logo vemos que no período da fonte coletada foram registrados 800 desastres pela SEDEC/SC com a solicitação de itens.

4.4.2. Tabela Fato Item Solicitados

A tabela fato itens solicitados vem também da *staging* solicitações mas com uma estrutura diferente: representando em cada linha um item solicitado e sua quantidade relativa por cada evento registrado. Dessa forma a tabela constitui 2.810 linhas e 5 colunas, sendo duas das colunas de valores monetários empenhados para o atendimento de cada item.

4.4.3. Tabela Fato Eventos

A tabela fato eventos corresponde ao mapeamento dos registros de eventos do CEMADEN com os dados das cidades coletados do IBGE e os dados de localização das estações meteorológicas do INMET. O objetivo deste mapeamento é rastrear os eventos registrados aos dados meteorológicos de cada estação.

Esse rastreio foi feito a partir da criação de um relacionamento entre a localização das estações meteorológicas com as cidades de cada evento registrado. Essa relação foi criada pela proximidade entre a estação e o ponto central do município, ou seja, foi calculado a distância de cada estação a cada uma das 295 cidades contidas no repositório do IBGE. A estação correspondente àquela cidade foi dada como a estação mais próxima ao centróide da cidade.

Por fim, foram agrupados os indicadores de dano humano e dano material, detalhados a nível granular na base não tratada, em somatórios de danos nas colunas: dano humano total, que corresponde ao total de pessoas afetadas pelo evento, incluindo pessoas mortas, acamadas e desalojadas; e dano material total, que corresponde ao total de prédios públicos, hospitais, escolas e residências afetadas pelo evento. Isso gerou a base final, com 7.976 linhas e 11 colunas.

4.4.4. Tabela Fato Meteorológico Diário

A tabela Fato Meteorológico Diário foi construída para agrupar as observações horárias das estações meteorológicas em um único dia. Esses agrupamentos foram feitos seguindo as diretrizes de Normais Climatológicas do Brasil, definidas pelo INMET (2022).

As diretrizes classificam variáveis em 3 diferentes grupos, com diferentes formas e fórmulas de agregação, havendo fórmulas particulares para variáveis de um mesmo grupo. Os grupos principais, suas definições e cálculos estão descritos no Quadro 9.

Quadro 9 - Grupos de Variáveis das Normais Climatológicas

Grupo	Variáveis	Cálculo Geral
I	Temperatura, pressão atmosférica a nível da estação, a nível médio do mar e de vapor, umidade relativa do ar, nebulosidade e vento	Média aritmética simples das observações no período.
II	Precipitação, evaporação e insolação.	Soma das observações no período.
III	Dias com chuva, dias com temperatura acima de limiar, etc.	Soma das observações condizentes nos períodos anteriores.

Fonte: INMET, 2022

No Quadro 10, são descritas as variáveis dos grupos com cálculos específicos para seu valor diário.

Quadro 10 - Variáveis com cálculos diários específicos

Variável	Cálculo Diário
Temperatura Média Compensada	$T_{MC,ijk} = \frac{T_{\max,ijk} + T_{\min,ijk} + T_{12,ijk} + 2 \times T_{24,ijk}}{5}$ (31)
Umidade Média Compensada	$UR_{C,ijk} = \frac{UR_{12,ijk} + UR_{18,ijk} + 2 \times UR_{24,ijk}}{4}$ (32)
Direção resultante do vento	$\begin{cases} \left\lvert \tan^{-1}\left(\frac{n(v)}{n(u)}\right) - 270° \right\rvert, & \text{se } n(u) > 0 \\ \left\lvert \tan^{-1}\left(\frac{n(v)}{n(u)}\right) - 90° \right\rvert, & \text{se } n(u) < 0 \end{cases}$ (33)
Dias com chuva	Soma dos *n* períodos anteriores com mais de 1 mm de precipitação.

Fonte: INMET, 2022

Dessa forma, as linhas da *staging* meteorológica foram agrupadas por dia e por estação, seguindo as equações descritas. Os valores de máximas e mínimas foram calculados como a máxima ou o mínimo daquele dia naquela estação. E por fim, a tabela foi acrescida de uma coluna binária que indica a ocorrência ou não de um evento de desastre natural, mapeada de acordo com os registros contidos nas tabelas fato eventos e fato solicitações.

A tabela resultante teve um total de 124.402 linhas e 18 colunas, um alto volume, já esperado dado o alto número de estações meteorológicas e de dias observados.

4.4.5. Tabela Fato Desastres Naturais

A tabela fato desastres naturais representa a união das tabelas fato meteorológico diário, fato eventos e fato solicitações, a fim de atribuir aos desastres registrados os indicadores meteorológicos que o geraram. Essa união foi feita com base no rastreamento das datas e cidades da tabela fato meteorológico diário nas datas e cidades dos eventos nas outras duas tabelas.

A tabela resultante teve um total de 5.267 linhas e 21 colunas, apontando que grande parte dos eventos registrados nas tabelas fato eventos e solicitações foram contemplados.

4.5. ANÁLISE EXPLORATÓRIA DOS DADOS (AED)

Para cada uma das bases construídas, foram aplicadas técnicas de AED diferentes para investigar e entender os comportamentos de cada base, o que fornecerá embasamento para decisões na etapa de Seleção de Modelos. Cada seção a seguir representa a AED realizada em cada uma das bases, seguidas de uma análise dos dados de projeções climáticas coletadas do INPE.

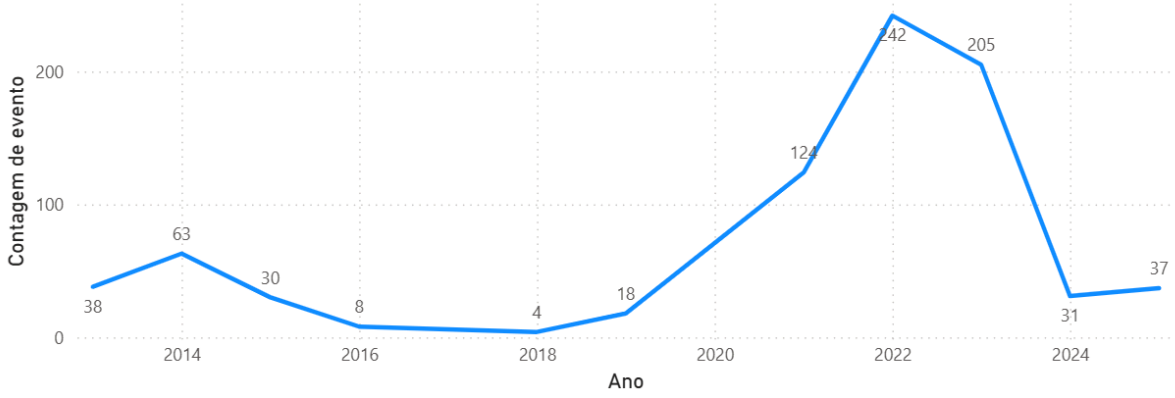
4.5.1. Tabela Fato Solicitações

A Fato Solicitações reúne um total de 800 registros ao longo do período analisado. As solicitações possuem datas variando entre 26 de junho de 2013 e 2 de julho de 2025, o que representa mais de uma década de informações coletadas. Em média, cada ocorrência está associada a 4,34 itens diferentes solicitados, evidenciando a diversidade de necessidades apresentadas em cada evento. Além

disso, as solicitações contemplam 228 municípios distintos do estado, mostrando boa representação geográfica.

Para apoiar a análise temporal, foi elaborado um gráfico de linhas que apresenta a evolução do número de solicitações ao longo dos anos, disposto na Figura 6. A análise temporal das solicitações revela que, entre 2013 e 2018, o número de registros anuais manteve-se relativamente baixo, oscilando entre poucos eventos e um máximo de 63 solicitações em 2014. A partir de 2019, observa-se uma tendência de crescimento acentuado, culminando em um pico expressivo no ano de 2022, com 242 solicitações. Após esse ápice, em 2023, o volume ainda se manteve elevado, com 205 solicitações, mas voltou a cair drasticamente em 2024, registrando apenas 31 eventos. O comportamento em 2025 sugere uma retomada, embora em menor escala até a data considerada.

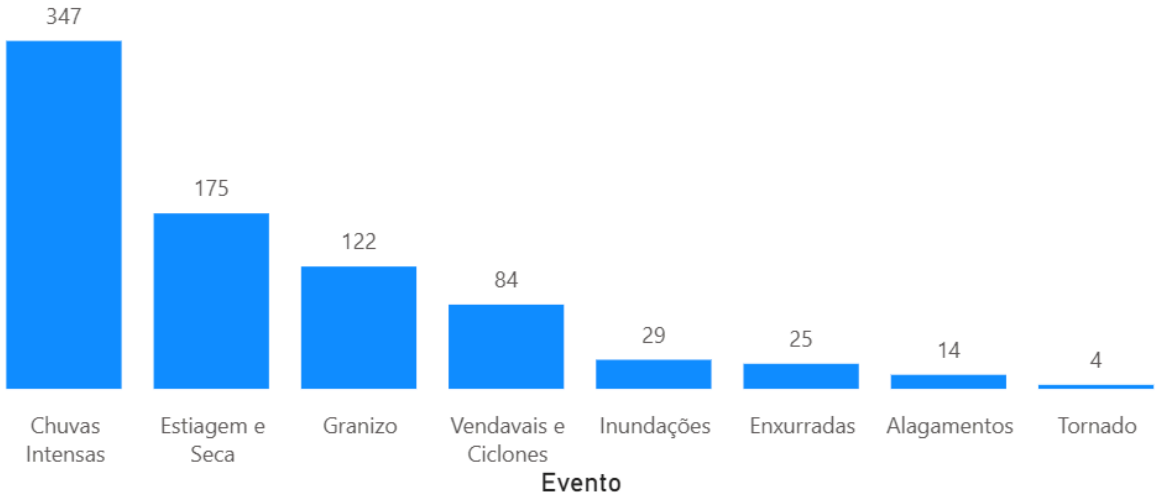
Figura 6 - Quantidade de eventos registrados por ano na Fato Solicitações



O comportamento oscilante dos registros de eventos podem sugerir baixas ocorrências de desastres na década de 2010, ou o baixo número de eventos que geram a necessidade de solicitações de itens, entretanto, ainda pode apontar a falta de registros oficiais dessas solicitações.

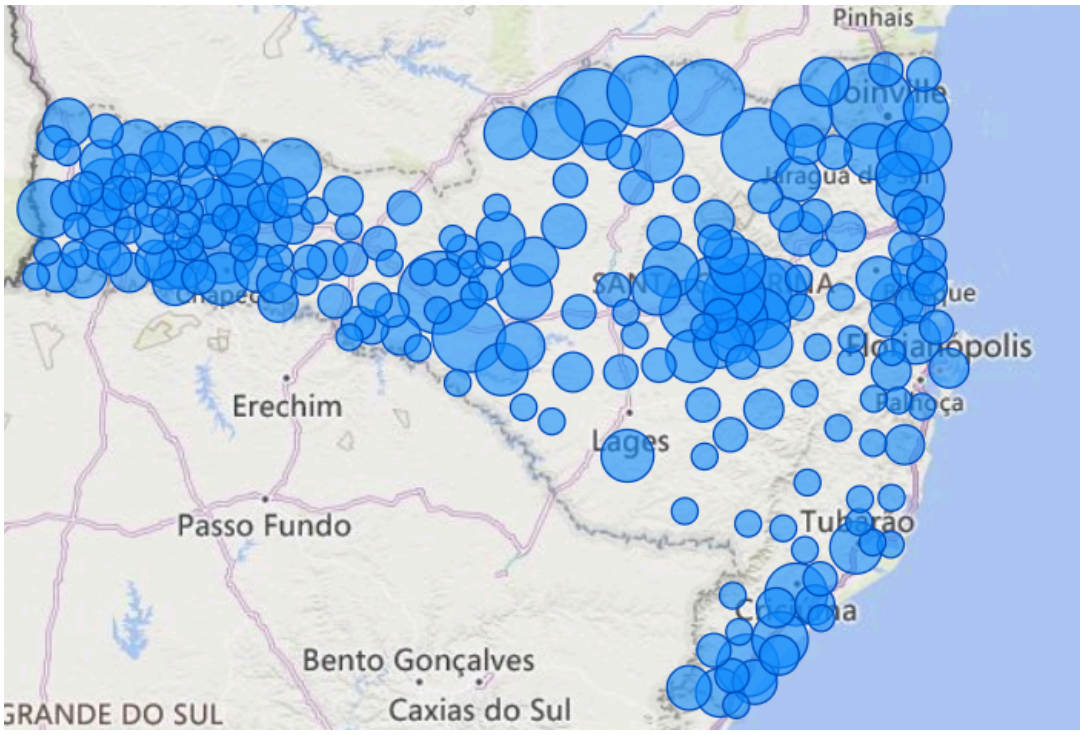
Complementarmente, o gráfico de barras na Figura 7 ilustra a distribuição das solicitações segundo o tipo de evento associado. Por meio dela podemos ver que os eventos hidrológicos (Inundações, Enxurradas e Alagamentos) representam apenas 17% das solicitações de todo o período, sendo o período marcado muito mais por fenômenos climatológicos. E entre eles, destaca-se as Chuvas Intensas que correspondem a quase metade de todas as solicitações, seguidas por fenômenos de Estiagem e Seca.

Figura 7 - Quantidade de solicitações por evento



Quanto à distribuição geográfica dessas solicitações, temos na Figura 8 um mapa de bolhas das ocorrências de solicitações por município, onde o tamanho de cada bolha representa o número de solicitações naquela localidade. Nele podemos ver a representatividade do Oeste catarinense nas solicitações, seguido pelo litoral, Norte, Centro-Oeste e Sul do estado.

Figura 8 - Mapa de bolhas das solicitações de itens



4.5.2. Tabela Fato Itens Solicitados

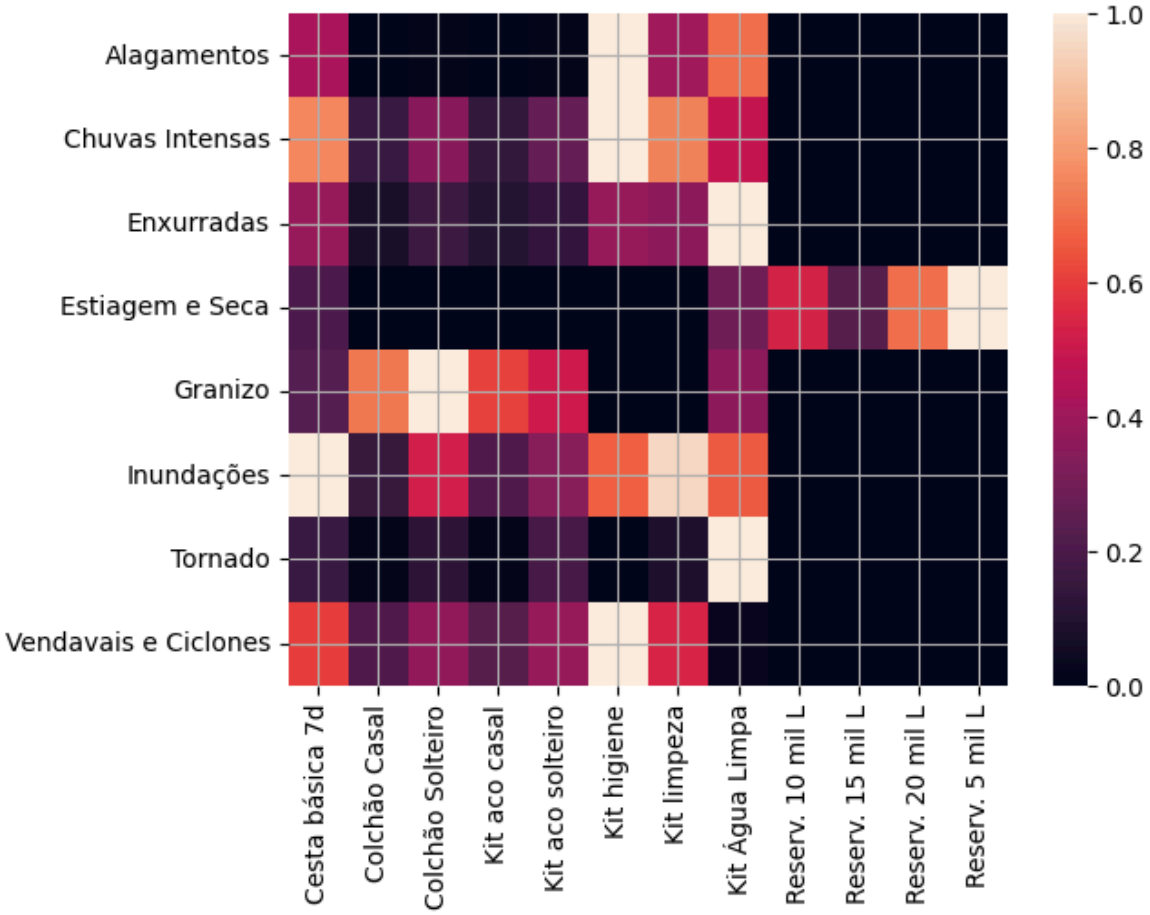
A Fato Itens Solicitados totaliza 2.810 registros, representando os múltiplos itens associados às 800 solicitações previamente analisadas. Cada ocorrência pode conter mais de um item demandado, e a consolidação desses dados permitiu a construção da Tabela 6, que apresenta para cada item, a quantidade de solicitações em que aparece, o total solicitado (considerando as quantidades pedidas em cada registro) e a média de unidades por solicitação.

Tabela 6 - Indicadores das solicitações por cada item

Item	Pedidos	Quant. Pedida	Média por Pedido
Cesta básica 7d	404	70.020	173
Colchão Casal	280	15.243	54
Colchão Solteiro	338	34.136	100
Kit aco casal	255	14.269	56
Kit aco solteiro	288	26.383	92
Kit higiene	266	91.199	343
Kit limpeza	377	67.912	180
Kit Água Limpa	224	56.944	254
Reserv. 10 mil L	93	397	4
Reserv. 15 mil L	62	167	3
Reserv. 20 mil L	86	523	6
Reserv. 5 mil L	137	742	5

Para apoiar a análise, foi elaborado um gráfico de calor que relaciona os itens solicitados a cada tipo de evento, disposto na Figura 9. Nessa visualização, os valores foram normalizados por evento, assim, é possível destacar itens que, embora não sejam os mais frequentes no geral, ganham importância em eventos específicos. Dessa forma, pode-se ver a concentração da solicitação de reservatórios apenas em fenômenos de seca e estiagem, e o foco de solicitações de itens característicos de desalojamento, como colchões e kits de acomodação, para fenômenos de granizo e vendavais.

Figura 9 - Mapa de calor de itens solicitados por evento



A Tabela 7 compõe os números não normalizados que construiram o gráfico da Figura 11. Nela, podemos ver a discrepância entre os valores de solicitação, em que vemos o grande volume de solicitações decorrentes de chuvas intensas no estado.

Tabela 7 - Quantidade total solicitado de cada item por evento

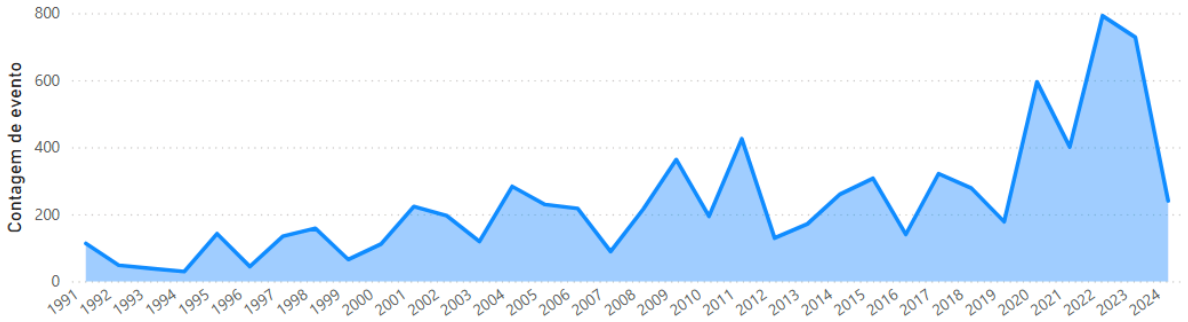
	Alagamentos	Chuvas Intensas	Enxurradas	Estiagem e Seca	Granizo	Inundações	Tornado	Vendavais e Ciclones
Cesta básica 7d	2.083	56.627	3.628	152	435	3.399	630	3.066
Colchão Casal	0	11.588	704	0	1.343	508	35	1.065
Colchão Solteiro	60	26.432	1.614	0	1.859	1.781	512	1.878
Kit aco casal	0	10.310	917	0	1.135	704	35	1.168
Kit aco solteiro	60	20.218	1.271	0	943	1.167	762	1.962
Kit higiene	4.976	75.324	3.551	0	0	2.258	0	5.090
Kit limpeza	2.004	56.262	3.332	0	0	3.228	330	2.756
Kit Água Limpa	3.460	36.738	9.400	212	669	2.220	4.100	145
Reserv. 10 mil L	0	0	0	397	0	0	0	0
Reserv. 15 mil L	0	0	0	167	0	0	0	0
Reserv. 20 mil L	0	4	0	519	0	0	0	0
Reserv. 5 mil L	0	0	0	742	0	0	0	0

4.5.3. Tabela Fato Eventos

O conjunto de dados analisado compreende 7.976 registros de eventos, abrangendo o período de 7 de janeiro de 1991 a 31 de dezembro de 2024.

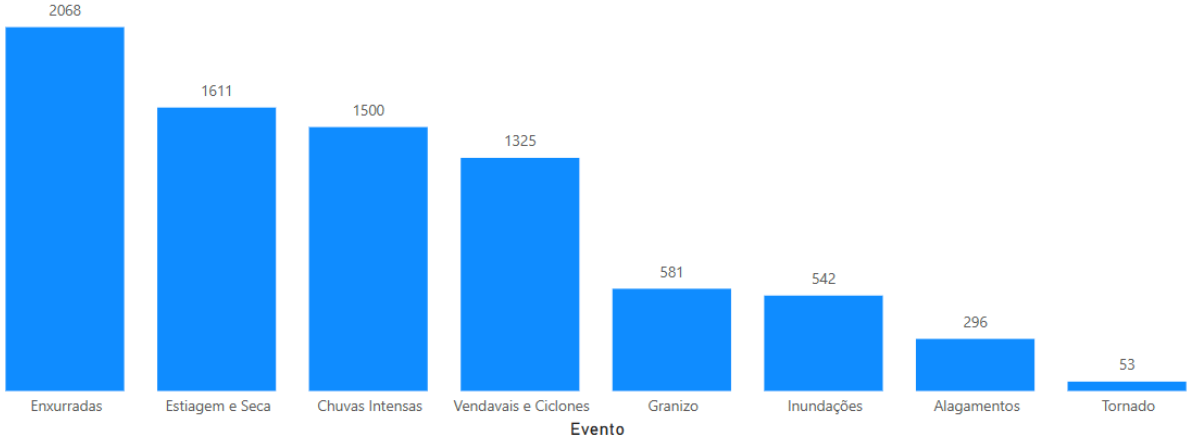
A evolução anual da contagem de eventos é apresentada em um gráfico de linha na Figura 10. Observa-se que a quantidade de eventos varia de um ano para outro, porém apresenta uma tendência crescente ao longo do tempo. Destaca-se o aumento mais expressivo a partir de 2020, atingindo seu ápice em 2022, com 792 eventos registrados.

Figura 10 - Eventos por ano (1991-2024)



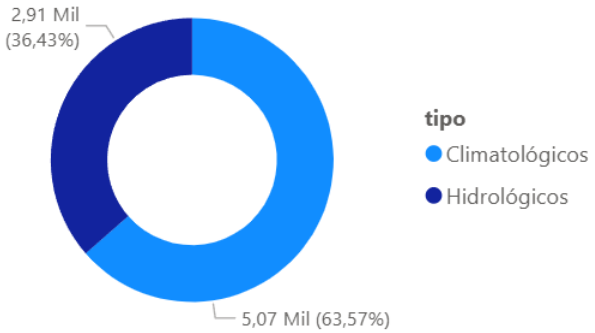
Um gráfico de barras na Figura 11 ilustra a distribuição dos registros por tipo de evento. Notavelmente, os eventos de enxurradas são os mais frequentes, seguidos por eventos de seca. A quantidade de registros varia conforme a localização dos eventos, e observa-se também a baixa ocorrência de outros eventos hidrológicos, como inundações e alagamentos. De forma geral, os eventos climatológicos representam a maior parte dos registros analisados.

Figura 11 - Distribuição de registros por tipo de evento



Olhando para os totais dos tipos de eventos, podemos ver a predominância dos eventos climatológicos nessa base, apresentados na Figura 12.

Figura 12 - Proporção de tipo de eventos registrados na série histórica

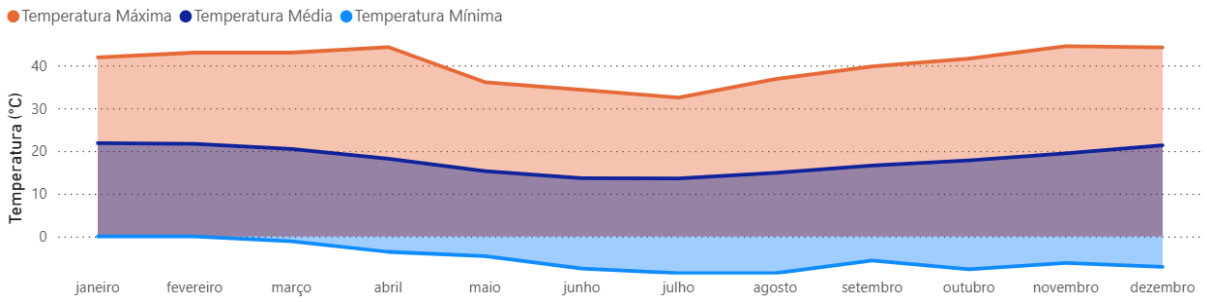


4.5.4. Tabela Fato Meteorológico Dia

O conjunto de dados analisado compreende 126.402 registros, abrangendo o período de 17 de março de 2003 a 30 de novembro de 2024.

A análise da sazonalidade das temperaturas foi realizada por meio de um gráfico de área mensal, representado na Figura 13, considerando a temperatura máxima, média e mínima. Observa-se que a temperatura média se mantém próxima de 20°C ao longo do ano. Destacam-se, no período analisado, temperaturas extremas, como a máxima de 44,3°C em abril e as mínimas registradas em julho e agosto, chegando a -8°C. De forma geral, o comportamento das temperaturas acompanha a sazonalidade típica das estações, com os valores mais baixos concentrados no meio do ano.

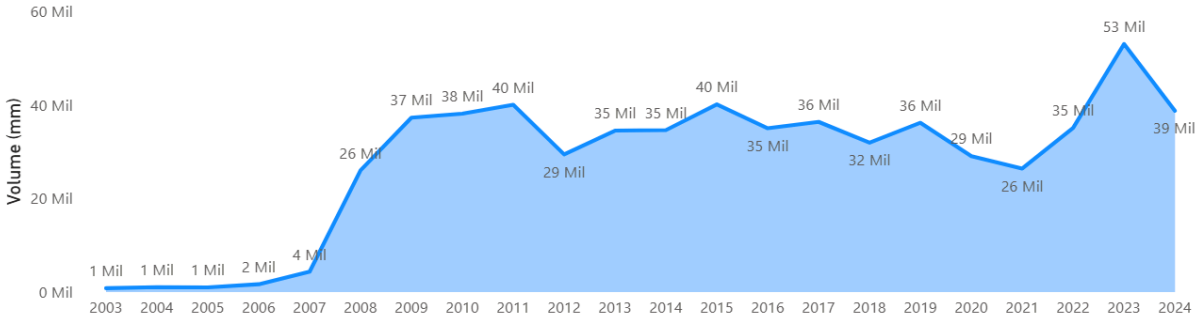
Figura 13 - Temperatura máxima, média e mínima por mês na série histórica



Em relação à precipitação total, foi construído um gráfico de linhas mostrando a soma anual por ano, disposto na Figura 14. Nota-se que, até o início da década de 2010, os registros apresentam baixa visibilidade devido a dados faltantes em algumas estações. Apesar disso, é possível identificar uma sazonalidade nos

volumes de precipitação, com variações de um ano para outro. O ano de 2023 se destaca por apresentar o maior volume de precipitação registrado, representando um pico em relação aos demais anos analisados.

Figura 14 - Precipitação total por ano (2003-2024)



4.5.5. Tabela Fato Desastres Naturais

A Fato Desastres Naturais conta com a junção dos desastres registrados na fato eventos com seus indicadores climáticos na fato meteorológicos dia, contemplando também desastres observados apenas nas bases de dados da SEDEC/SC. Ao todo, a base reúne 5.267 registros, cobrindo o período compreendido entre 20 de agosto de 2003 e 27 de novembro de 2024.

A distribuição dos eventos pode ser observada no gráfico da Figura 15, no qual se destaca a predominância de chuvas intensas como principal tipo de desastre registrado. Ainda assim, há uma quantidade expressiva de ocorrências de enxurradas, indicando que ambos os fenômenos possuem relevância significativa na base. Já o gráfico de rosca da Figura 16 evidencia a proporção dos tipos de desastres observada na Fato Eventos, porém com uma leve alteração, sendo 70% classificados como climatológicos e 30% como hidrológicos.

Figura 15 - Distribuição da quantidade de eventos na série histórica

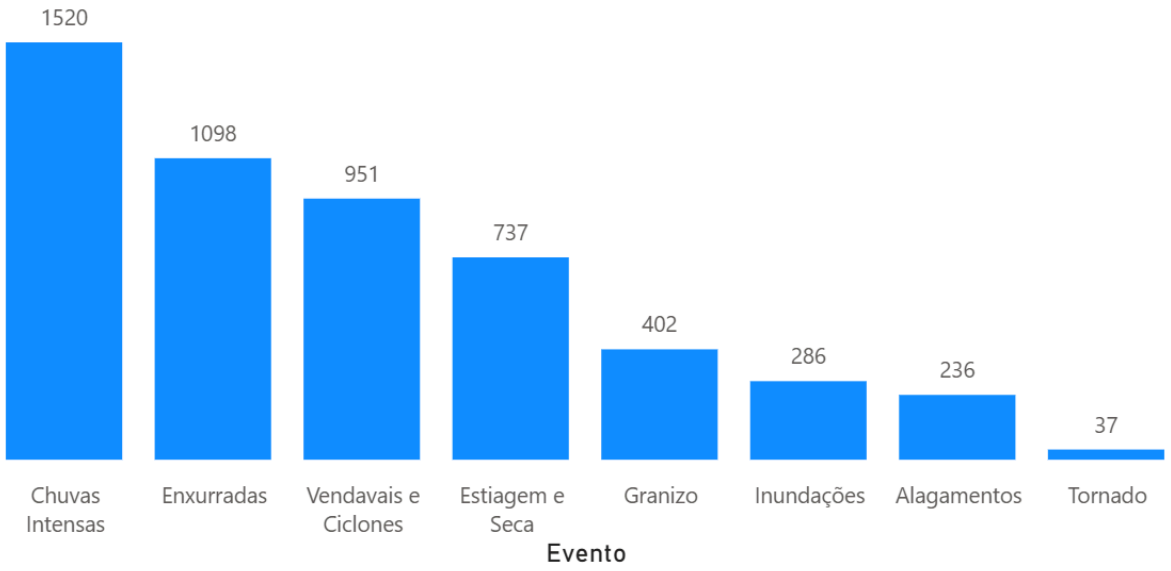
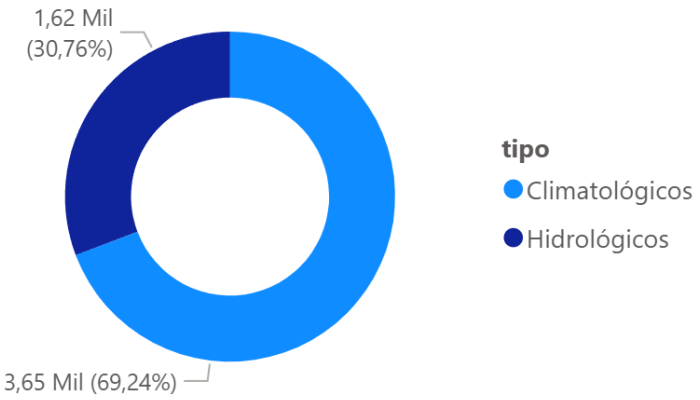


Figura 16 - Proporção de eventos por tipo



A fim de avaliar o comportamento das variáveis em cada evento, a Tabela 8 foi criada a partir das médias dos indicadores meteorológicos por evento. Os resultados mostram o que se era esperado: uma média de precipitação mais elevada em eventos de chuva, enquanto em casos de estiagem esse valor se torna praticamente nulo. Por outro lado, a estiagem se destaca nos valores médios de radiação global e temperatura média, refletindo condições típicas de períodos prolongados de seca.

Tabela 8 - Média de cada indicador meteorológico por evento

Evento	Precipitação (mm)	Pressão Atm. (hPa)	Radiação Glob. (kJ/m²)	Temperatura (°C)	Umidade Relat. (%)	Vel. Vento (m/s)
Alagamentos	29,87	912,73	120,50	18,00	76,51	0,56
Chuvas Intensas	32,88	888,93	128,35	17,92	76,39	0,61
Enxurradas	33,08	928,86	129,89	18,41	80,66	0,62
Estiagem e Seca	2,69	894,46	208,79	19,52	65,52	0,67
Granizo	9,71	920,11	208,17	18,42	74,75	0,68
Inundações	20,73	927,21	132,46	15,68	78,12	0,44
Tornado	17,65	927,33	123,45	17,67	84,77	0,64
Vendavais e Ciclones	13,54	917,10	147,69	17,14	74,53	0,71

4.5.6. Conclusões da AED

A AED das bases da modeladas trouxe informações importantes para a construção do modelo proposto. O primeiro ponto importante que a AED mostrou é o grande desequilíbrio entre os tipos de desastres. Os eventos de origem climatológica e meteorológica, principalmente Chuvas Intensas e Estiagem/Seca, ocorrem muito mais nas informações históricas, especialmente nas solicitações de itens. Esse desequilíbrio exige que sejam usadas formas de avaliação mais rigorosas, como o F1-Score, nas fases de classificação, pois a simples acurácia poderia dar resultados falsos por causa da classe mais frequente.

As análises quanto aos item pedidos mostram grande diferença nas quantidade solicitadas entre cada item. A variação dos valores é muito grande, indo de grandes volumes em itens básicos, como cestas básicas e kits de higiene, até volumes pequenos de itens mais específicos e caros (como reservatórios de água). Essa diferença na quantidade de pedidos cria um desafio sério na fase de regressão, onde se espera que os erros de quantidade (MAE e RMSE) sejam altos. Por isso, é fundamental avaliar a forma como o modelo capta a distribuição dos valores, e não apenas a precisão dos valores pontuais. Dessa forma, a métrica de Distância de Wasserstein é uma boa ferramenta para captar isso.

A análise das médias dos indicadores meteorológicos confirmou que as informações de entrada são úteis para diferenciar os eventos. Por exemplo, grandes volumes de precipitação se ligam a Chuvas Intensas, e a baixa precipitação junto

com a alta radiação global estão ligadas à Estiagem. Isso mostra que as variáveis que o modelo usa têm a capacidade necessária para distinguir os diferentes tipos de desastres na fase de classificação.

Por fim, a AED também ajudou a decidir o que o trabalho incluiria. Foi confirmada a pouca importância dos fenômenos hidrológicos (Inundações, Enxurradas e Alagamentos) para o volume de pedidos de itens, pois eles representam uma parte pequena dos registros. Essa baixa quantidade de dados, somada à falta de dados sobre as causas desses eventos (como informações detalhadas sobre solo), justifica a decisão de não incluir essas classes no treinamento do modelo. Isso faz com que o modelo se concentre nos fenômenos mais comuns e com informações de entrada mais seguras.

4.6. SELEÇÃO DE MODELOS

Nesta seção serão apresentados os processos de criação e seleção dos modelos para cada uma das etapas definidas para o *pipeline*.

Para este desenvolvimento, o tamanho do conjunto de testes foi de 20% da amostra original, em todos os casos, e a *seed* de aleatoriedade passada para todos os componentes foi 42. Para a execução de cada teste foi realizada uma validação cruzada (CV) para determinar as amplitudes das métricas principais de análise, e em seguida uma das iterações da CV foi utilizada para gerar a matriz de confusão e os cálculos das importâncias de cada *feature* de X.

4.6.1. Dias com Desastre

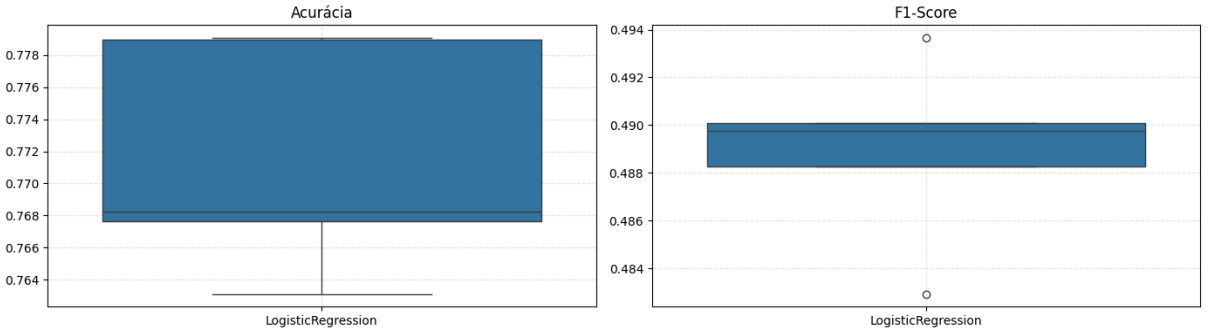
A primeira etapa do *pipeline* consiste na previsão de se haverá um evento no local, sendo este um problema de classificação binária. Para isso, a tabela utilizada foi a Fato Meteorológico Diário, com o mapeamento junto da tabela Fato Eventos para identificar em quais dos registros houveram evento. Essa identificação foi feita por meio de uma coluna binária, chamada “houve_evento”, a qual é a variável alvo desta etapa.

Foram testados os algoritmos *LogisticRegression, RandomForestClassifier e XGBoostClassifier.* Por se tratar de um problema de classificação binária, o modelo de regressão logística teve de ser testado, visando avaliar a possibilidade de seu uso, já que se trata de um modelo com gasto computacional menor. Os resultados da aplicação de cada algoritmo estão abaixo.

4.6.1.1. Regressão Logística

O primeiro modelo testado foi a Regressão Logística. O teste de validação gerou os resultados das métricas acurácia e *F1-Score* apresentados nos *boxplots* da Figura 17. O gráfico da acurácia apresenta uma amplitude razoável, não tendo nenhum *outlier* nos testes. Já o teste de *F1* mostra a existência de dois *outliers*, que fogem dos quartis da figura.

Figura 17 - Resultados da validação cruzada da etapa 1 para regressão logística



A média da acurácia foi de 77,14%, com um baixo desvio padrão, de 0,72%, sua mediana representada no gráfico, entretanto foi menor que a média, com um valor de 76,82%. Já o *F1* teve uma média próxima da mediana: 0,4889 de média e 0,4898 de mediana.

Analisando agora a matriz de confusão do experimento, apresentada na Figura 18, vemos uma grande concentração de verdadeiros negativos, acompanhados dos outros valores tímidos. Ao olhar para a Tabela 9, que contém os números da matriz de confusão, pode-se ver a discrepante existência de verdadeiros negativos em comparação ao resto, com um destaque ainda para a quantidade de falsos negativos, que chegou a mais de 2% dos resultados.

Figura 18 - Matriz de confusão da regressão logística na etapa 1

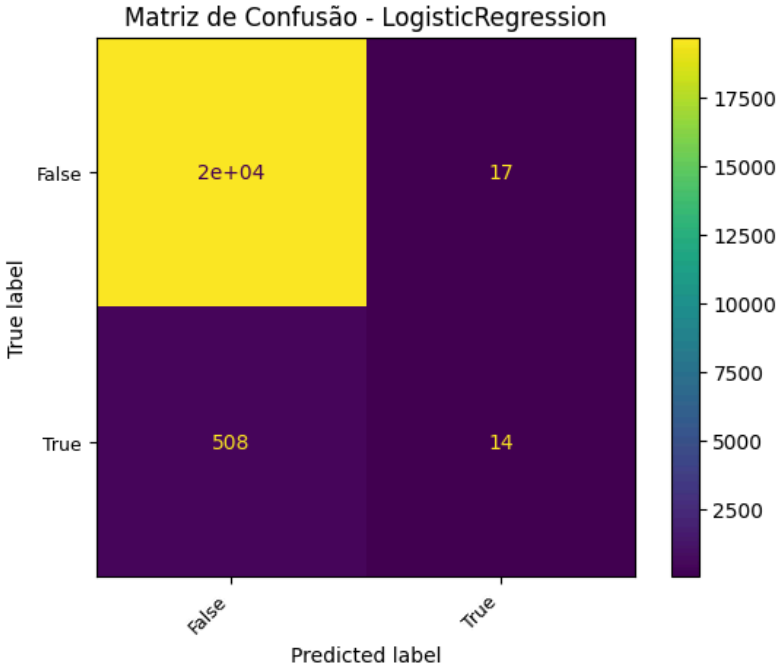
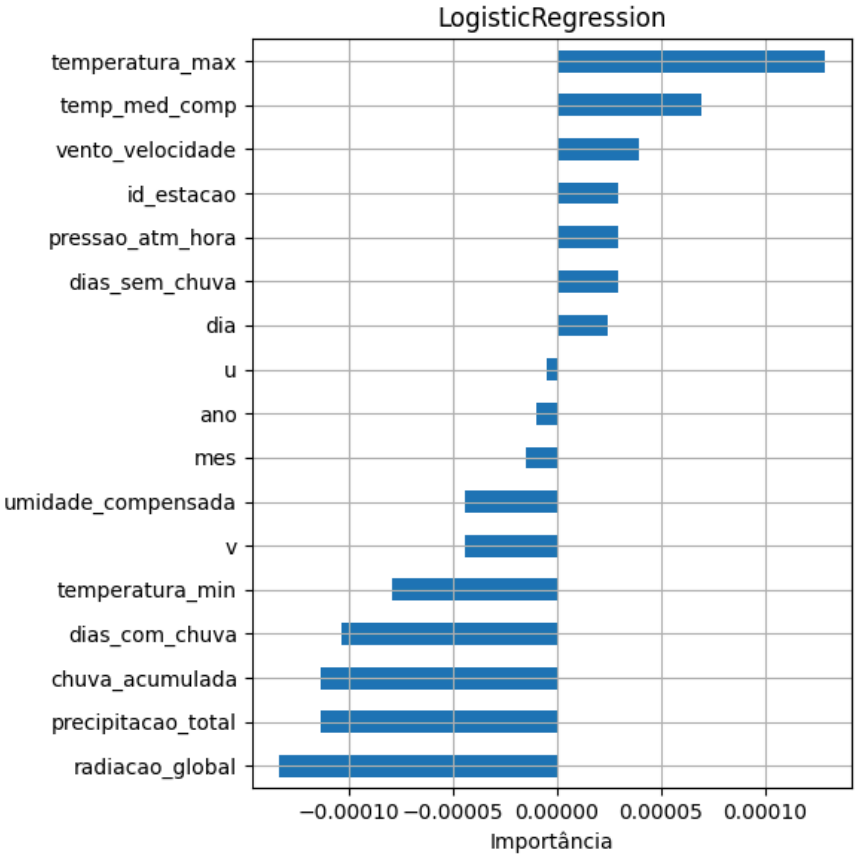


Tabela 9 - VPs, VNs, FPs e FNs da regressão logística na etapa 1

Verdadeiros Positivos		**Verdadeiros Negativos**		**Falsos Positivos**		**Falsos Negativos**	
Num.	**%**	**Num.**	**%**	**Num.**	**%**	**Num.**	**%**
14	0,07%	19.686	97,33%	17	0,08%	508	2,51%

Para análise de como o algoritmo se adaptou a X, foi feito um gráfico das importâncias de cada *feature* de X nos resultados do modelo, disposto na Figura 19. Para tal foi usado o método de permutação, já que o algoritmo de Regressão Logística se trata de um estimador linear e não contabiliza a importância de cada *feature* para sua função de perda. Com isso, podemos ver que as *features* de temperatura, vento e localidade (dada pela “id_estacao”) são as mais influentes para o modelo, enquanto as *features* de chuva e de radiação global tem menor importância para a previsão correta do algoritmo.

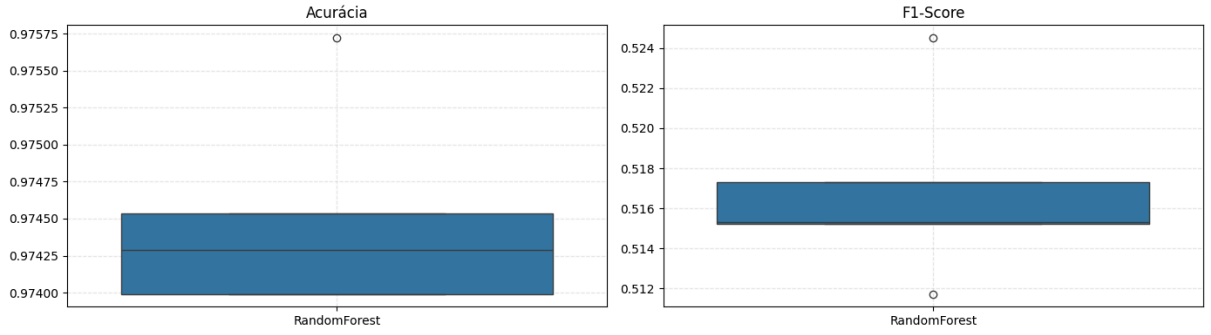
Figura 19 - Importância das *features* para a regressão logística na etapa 1



4.6.1.2. Random Forest

O segundo algoritmo testado foi o *Random Forest*. Os resultados das métricas do experimento estão dispostos na Figura 20. Nele, podemos ver valores quase perfeitos de acurácia, com um valor mínimo de 97,4%. Olhando, agora, para o *F1-Score*, seu valor ainda indica umo desempenho mediana do modelo, com valores se mantendo abaixo dos 0,60.

Figura 20 - Resultados da validação cruzada da etapa 1 para *random forest*



A matriz de confusão para o algoritmo está disposta na Figura 212 abaixo. Nela podemos ver a predominância de verdadeiros negativos perante o resto,

seguida pelo grande número de falsos negativos. As porcentagens de cada classe estão na Tabela 10, onde pode-se ver uma baixa expressividade de verdadeiros positivos e falsos positivos perante as outras classes.

Figura 21 - Matriz de confusão do *Random Forest* na etapa 1

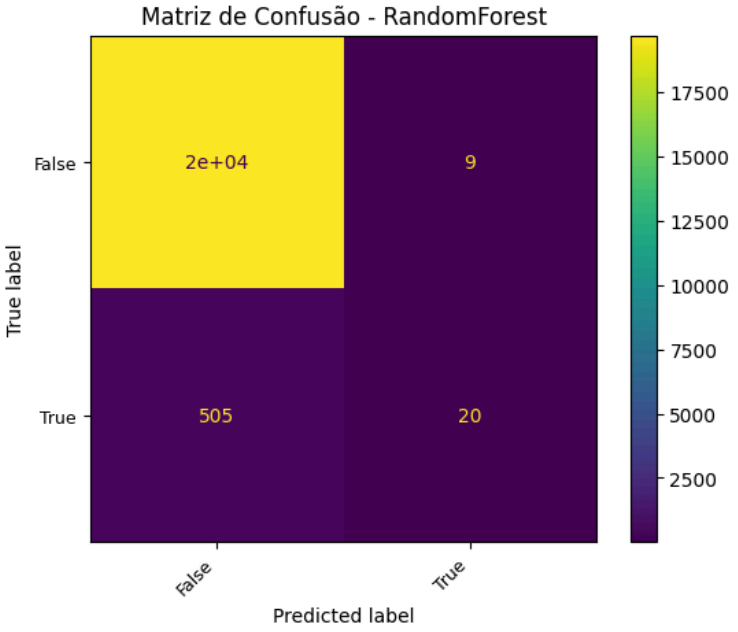
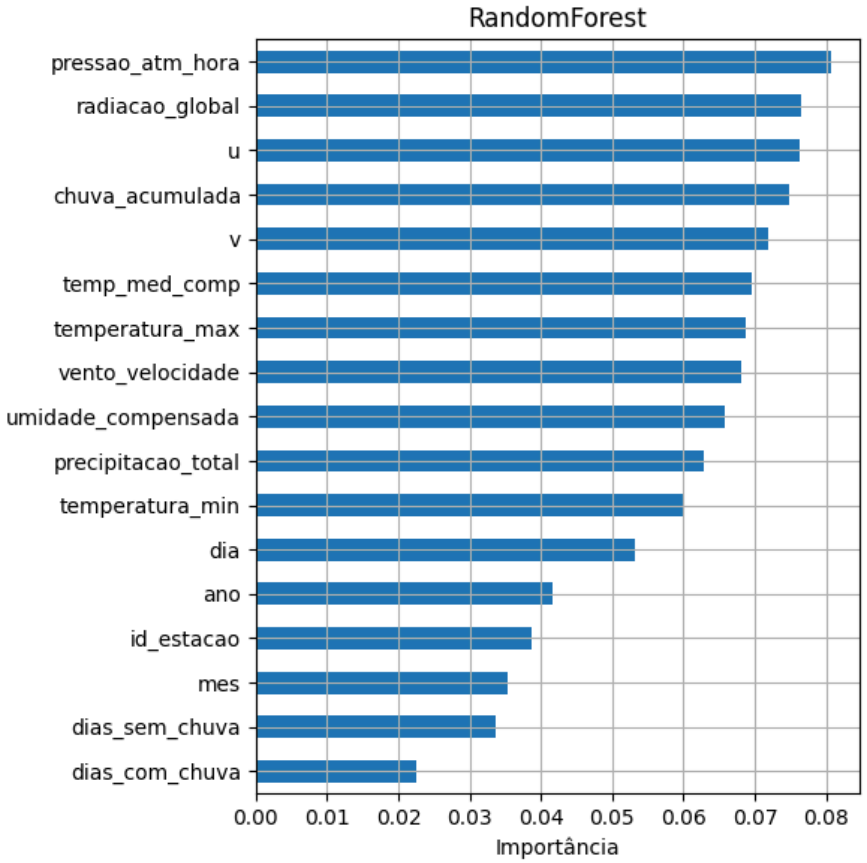


Tabela 10 - VPs, VNs, FPs e FNs do *Random Forest* na etapa 1

Verdadeiros Positivos		**Verdadeiros Negativos**		**Falsos Positivos**		**Falsos Negativos**	
Num.	**%**	**Num.**	**%**	**Num.**	**%**	**Num.**	**%**
20	0,10%	19.691	97,36%	9	0,04%	505	2,50%

Por fim, as importâncias de cada *feature* estão apresentadas na Figura 22. Neste algoritmo é possível ver a importância que cada *feature* trás para os resultados, e mostra a grande importância das variáveis de pressão, radiação e componentes do vento, enquanto as *features* de data e de dias com e sem chuva tem uma força consideravelmente menor na minimização da função de perda.

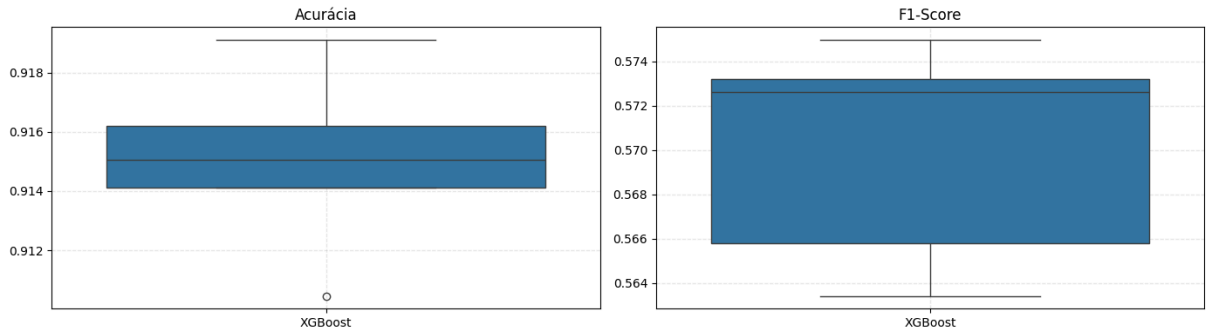
Figura 22 - Importância das *features* para o *random forest* na etapa 1



4.6.1.3. XGBoost

O último algoritmo testado nesta etapa foi o *XGBoost.* Os resultados de suas métricas para as 5 validações cruzada estão na Figura 23 abaixo. Nele pode-se ver a boa acurácia do experimento, passando dos 90% em todos os testes. Já o *F1-Score* tem uma amplitude maior, de 0,10, se mantendo em valores médios na escala da métrica.

Figura 23 - Resultados da validação cruzada da etapa 1 para *XGBoost*



A matriz de confusão do experimento está apresentada na Figura 24. Ela segue o padrão apresentado nos últimos experimentos, com uma grande presença

de verdadeiros negativos e falsos negativos. Os números dessa matriz estão dispostos na Tabela 11, e detalham as baixas porcentagens de positivos que o algoritmo prevê, efeito possivelmente causado pelo desbalanceamento de classes para a previsão binária.

Figura 24 - Matriz de confusão do *XGBoost* na etapa 1

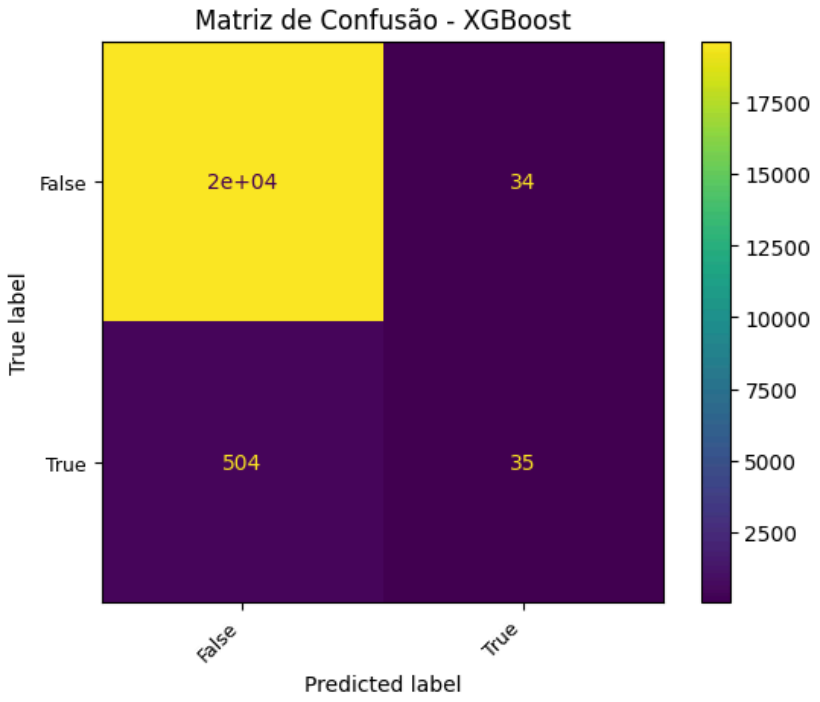
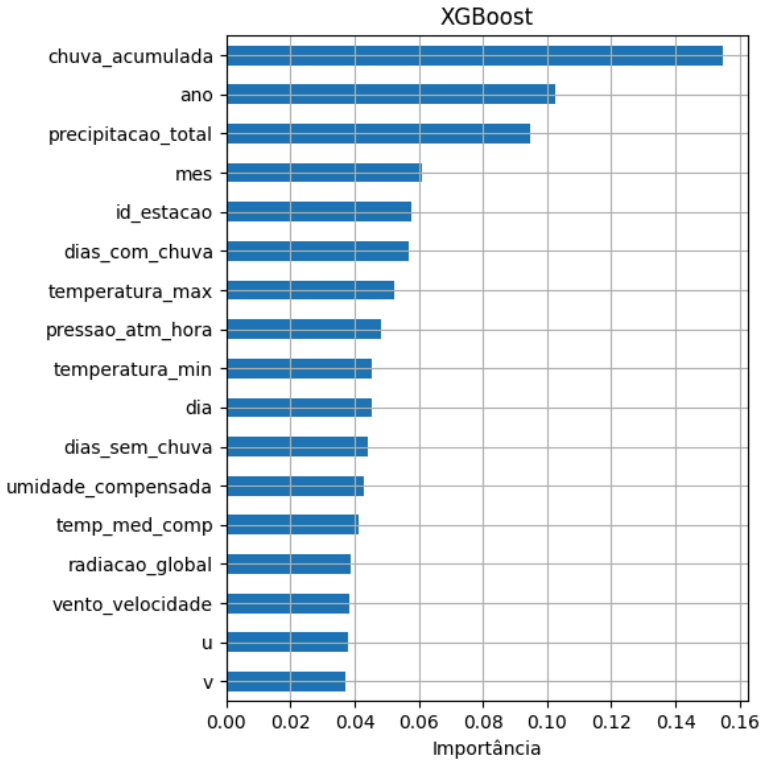


Tabela 11 - VPs, VNs, FPs e FNs do *XGBoost* na etapa 1

Verdadeiros Positivos		Verdadeiros Negativos		Falsos Positivos		Falsos Negativos	
Num.	**%**	**Num.**	**%**	**Num.**	**%**	**Num.**	**%**
35	0,17%	19.652	97,17%	34	0,17%	504	2,49%

Por fim, a análise das importâncias de cada *feature* está disposta na Figura 25. Nela podemos ver a esmagadora importância da variável de chuva acumulada, mostrando a dependência do modelo nessa variável, seguida pelo ano e pela precipitação total. As outras variáveis climáticas têm importâncias médias para as previsões, com destaque para as variáveis de vento, que têm as menor importância para as decisões do algoritmo.

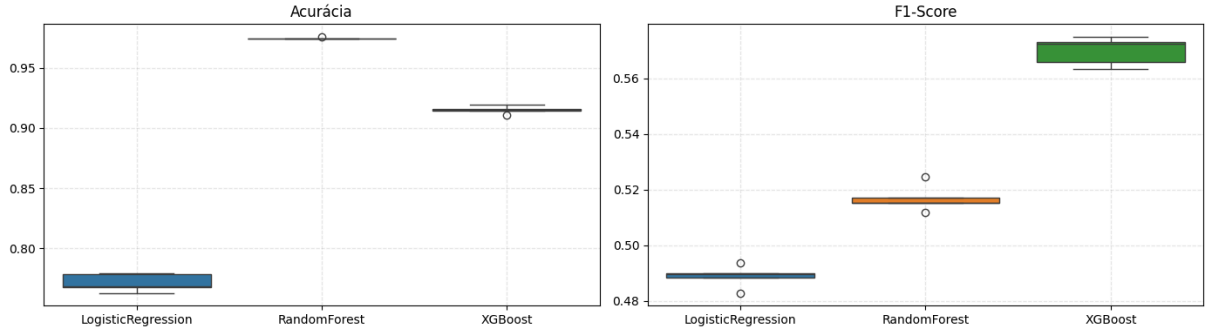
Figura 25 - Importância das *features* para o *XGBoost* na etapa 1



4.6.1.4. Algoritmo Selecionado

A Figura 26r esume os testes de acurácia e *F1-score* comparativo entre cada um dos algoritmos usados na etapa. Com ela é possível ver a qualidade superior dos modelos baseados em árvore quando comparado ao modelo linear, o qual não atingiu 80% de acurácia. Dessa forma a regressão logística foi descartada da seleção.

Figura 26 - Resultados da validação cruzada da etapa 1 para todos os algoritmos



O *Random Forest* foi o algoritmo superior para a métrica de acurácia, passando em 5% o segundo colocado, entretanto, seu *F1-Score* ficou mais próximo do algoritmo de regressão logística, que não atingiu 0,50.

Como ambos os modelos baseados em árvores tiveram uma acurácia considerada ótima, acima de 90%, o *F1-Score* e o grande risco de *overfitting* por parte do *Random Forest* são os pontos que fazem este modelo ser descartado. Logo, o modelo selecionado para realizar a previsão nesta etapa foi o *XGBoost.*

4.6.2. Tipo do Evento de Desastre

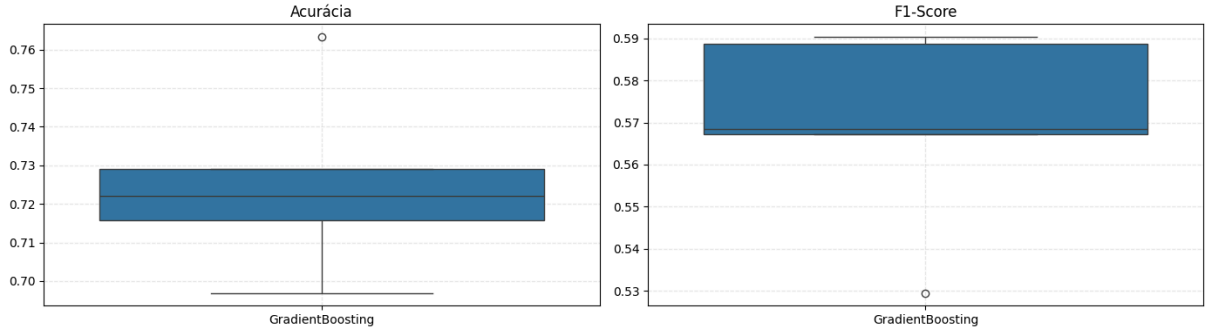
A segunda etapa do *pipeline* consiste na previsão do tipo de evento de desastre, sendo esse um problema de classificação com mais de uma classe. Para isso, a tabela utilizada foi a Fato Desastres Naturais, contemplando apenas os fenômenos climatológicos. Essa segregação se dá pela baixa proporção de eventos hidrológicos tanto na Fato Desastres Naturais, quanto na Fato Solicitações, como apontado na AED realizada na Seção 4.5.

Foram aplicados os algoritmos *GradientBoostingClassifier, RandomForestClassifier e XGBoostClassifier* diretamente aos dados, avaliando seu desempenho com as métricas acurácia e *F1-Score.*

4.6.2.1. Gradient Boosting

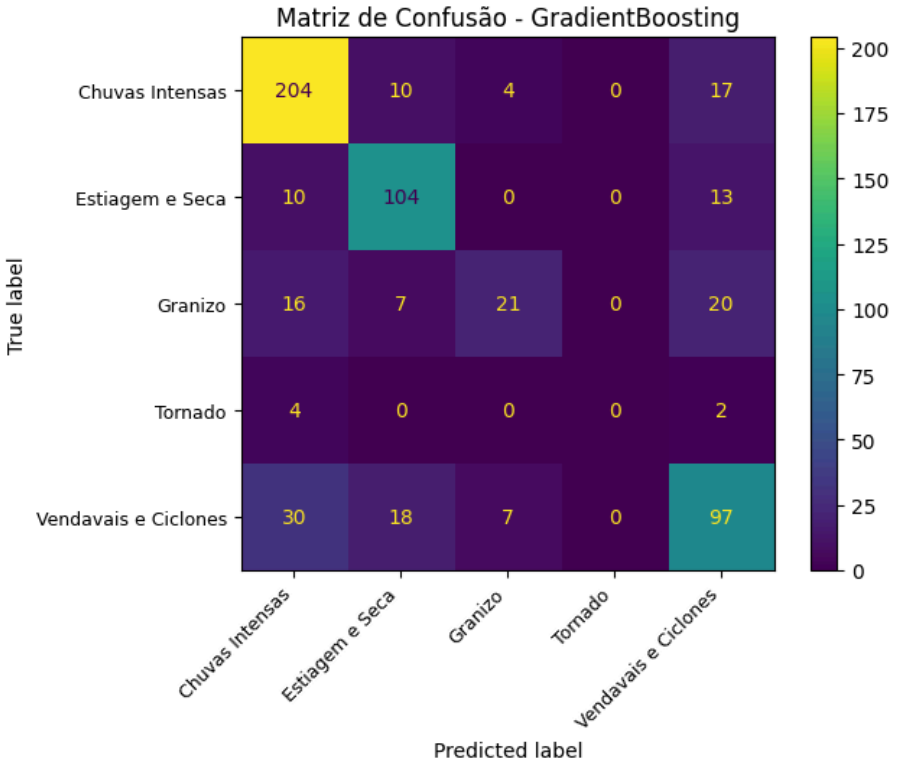
O primeiro algoritmo testado foi o *Gradient Boosting.* A Figura 27 mostra os *boxplots* de cada uma das validações cruzadas realizadas no algoritmo. A acurácia teve grande amplitude no experimento, variando desde abaixo de 70% até acima 76%, e algo similar ocorreu com o *F1-Score,* com uma amplitude de 0,06. Ainda sim, os valores apresentados são satisfatórios, passando de 70% de acurácia e resultados de precisão dessas previsões acima dos 0,5 no *F1.*

Figura 27 - Resultados da validação cruzada da etapa 2 para *gradient boosting*



A matriz de confusão de uma das iterações da validação cruzada está disposta na Figura 28 abaixo, e os resultados detalhados para cada tipo de evento estão na Tabela 12 em seguida. Primeiramente, pela imagem é possível ver uma tendência de verdadeiros positivos nos eventos de chuva, estiagem e vendavais, porém valores baixos para granizo e Tornado. Destaque também para o alto número de vezes que o algoritmo prevê chuvas quando são outros tipos de eventos meteorológicos, e até mesmo a prevendo para casos de estiagem.

Figura 28 - Matriz de confusão do *Gradient Boosting* na etapa 2



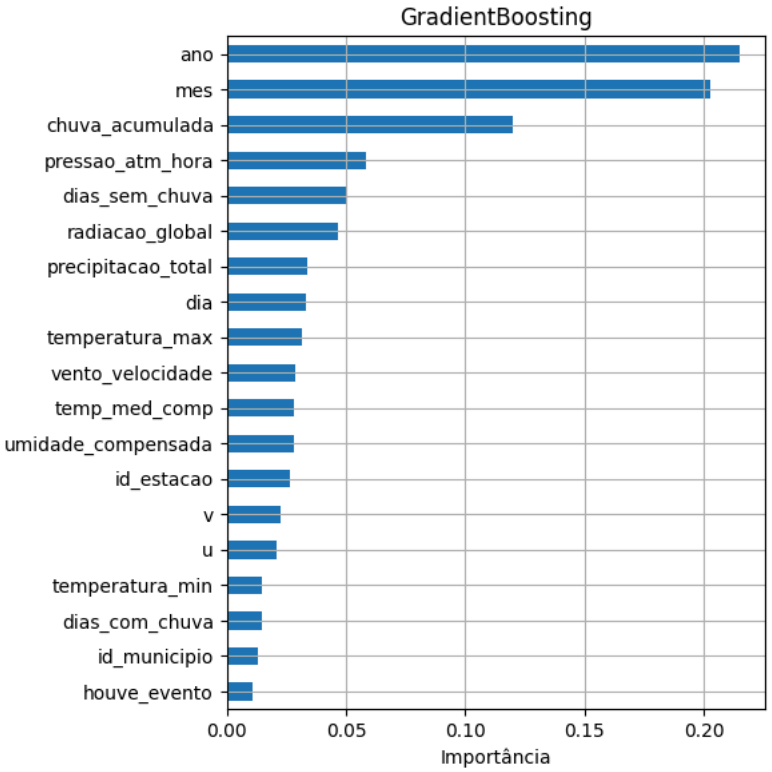
Olhando os números do gráfico acima de forma tabular, vemos o baixo desempenho do algoritmo no evento de Tornado, com 0 previsões verdadeiramente positivas. Entretanto, com os outros eventos o desempenho foi satisfatório, com valores de quantidades previstas totais condizentes com as quantidades de eventos naquela amostra. Olhando para os índices de verdadeiros e falsos positivos, os valores de previsões corretas se mostram acima dos 70% em todos os eventos, com exceção do tornado.

Tabela 12 - VPs e FPs do *Gradient Boosting* na etapa 2

Label	Qnt. Eventos Previstos	Verdadeiros Positivos		Falsos Positivos	
		Num.	%	Num.	%
Chuvas Intensas	271	205	75,65%	66	24,35%
Estiagem e Seca	119	95	79,83%	24	20,17%
Granizo	35	26	74,29%	9	25,71%
Tornado	3	0	0,00%	3	100,00%
Vendavais e Ciclones	156	110	70,51%	46	29,49%

Por fim, analisando as *features* que mais impactam no algoritmo, é possível ver que as variáveis de tempo “ano” e “mês” são as que mais ajudam o algoritmo a reduzir sua função de perda, tendo elas uma importância quase duas vezes maior que a terceira variável de chuva acumulada. As *features* das componentes do vento e de temperatura se mostram menos impactantes, assim como as variáveis de localidade.

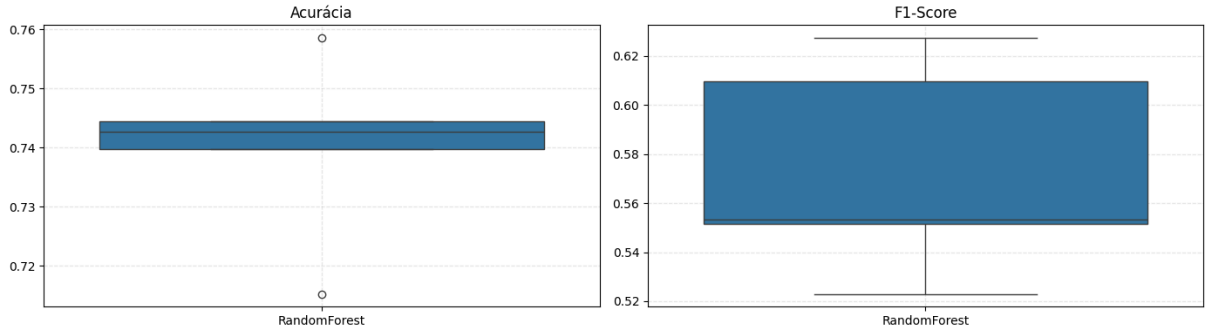
Figura 29 - Importância das *features* para o *Gradient Boosting* na etapa 2



4.6.2.2. Random Forest

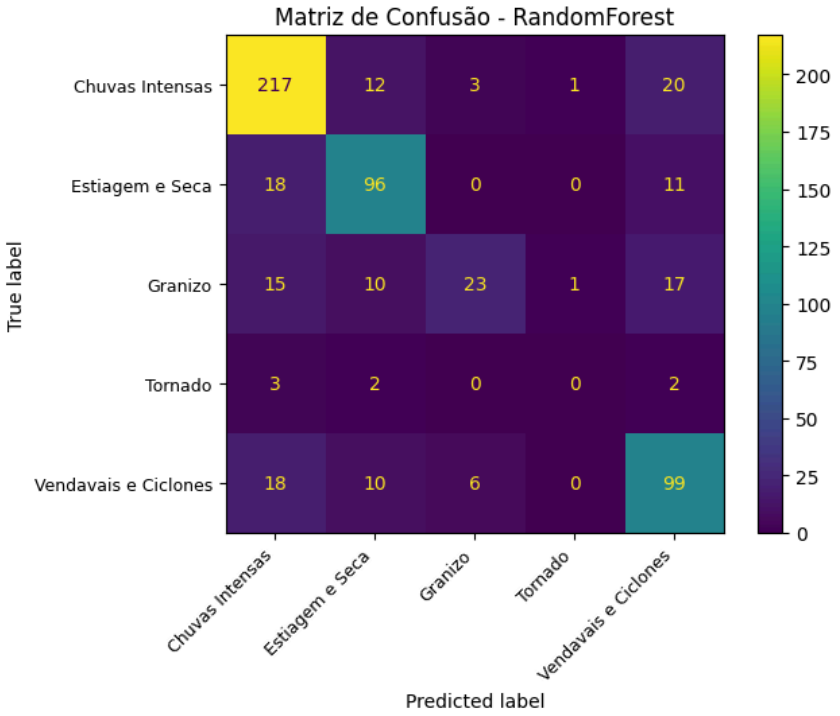
O segundo algoritmo testado nessa etapa foi o *Random Forest.* Nele é possível ver um aumento na amplitude do *F1-Score*, e uma acurácia similar ao algoritmo anterior. A acurácia se manteve acima dos 70% em todas as iterações, tendo outliers que chegam até 76%. Já o *F1-Score* conta com uma amplitude de 0,1, mas com um valor mediano abaixo de 0,56, o que indica um modelo com desempenho ainda mediano.

Figura 30 - Resultados da validação cruzada da etapa 2 para *random forest*



Olhando agora para os resultados da matriz de confusão disposta na Figura 31, é possível notar a similaridade deste algoritmo com o anterior, com a ótima capacidade em previsão de chuvas, estiagem e vendavais, porém a dificuldade para os outros tipos de evento. Destaque novamente para a quantidade de falsos positivos para chuvas, e desta vez para vendavais, sendo esses dois eventos pontos de confusão para o algoritmo.

Figura 31 - Matriz de confusão do *Random Forest* na etapa 2



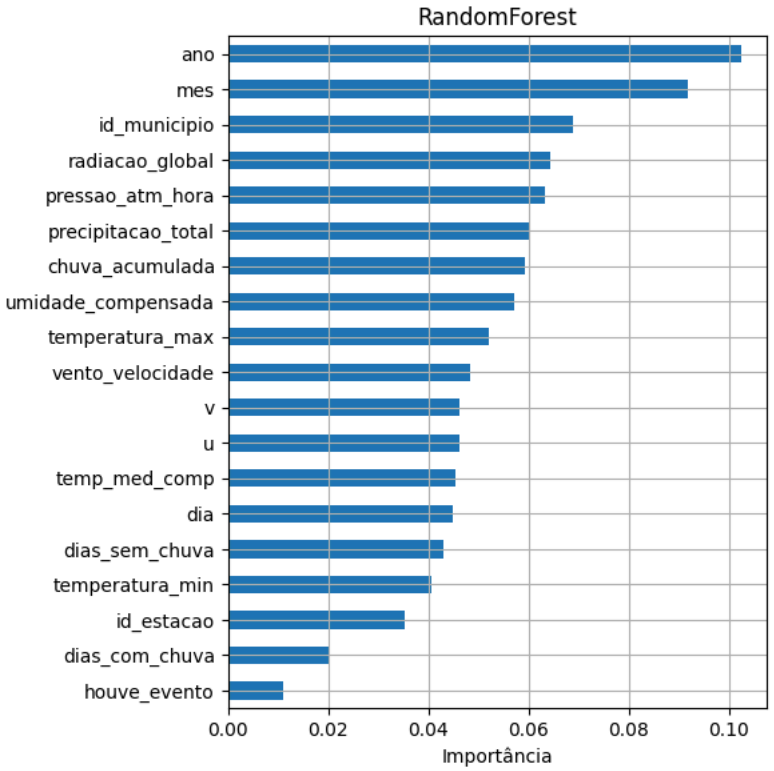
Olhando em números as informações da matriz acima, pode-se ver índices de verdadeiros positivos que ultrapassam os 80%, indicando um bom desempenho do algoritmo para o evento de chuva. Porém, ao olharmos para para o evento de vendaval, seu desempenho é baixo comparada aos outros, sendo 40% de seus falsos positivos, previsões de chuvas.

Tabela 13 - VPs e FPs do *Random Forest* na etapa 2

Label	**Qnt. Eventos Previstos**	**Verdadeiros Positivos**		**Falsos Positivos**	
		Num.	**%**	**Num.**	**%**
Chuvas Intensas	271	217	80,07%	54	19,93%
Estiagem e Seca	130	96	73,85%	34	26,15%
Granizo	32	23	71,88%	9	28,13%
Tornado	2	0	0,00%	1	50,00%
Vendavais e Ciclones	149	99	66,44%	50	33,56%

Por fim, a análise de importâncias para o algoritmo apresentada na Figura 32 mostra, novamente, a relevância das variáveis de tempo dentro do algoritmo de previsão de eventos. Seguida por essas variáveis temos a localidade do município sendo um fator de relevância para o acerto da previsão.

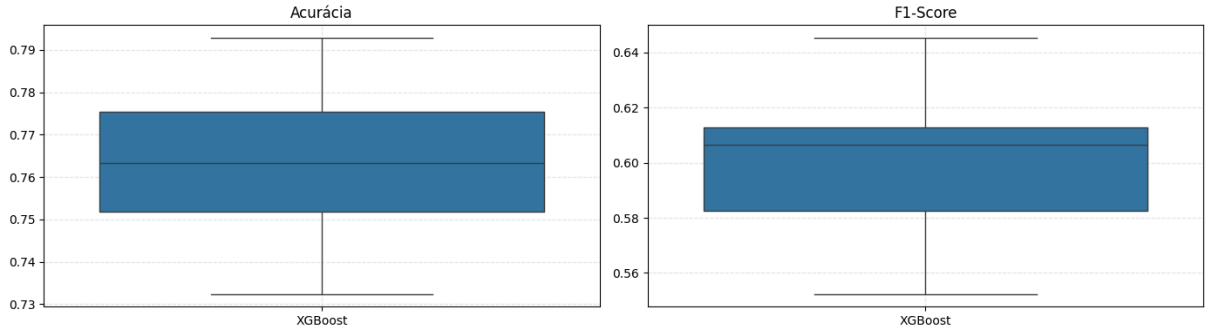
Figura 32 - Importância das *features* para o *Random Forest* na etapa 2



4.6.2.3. XGBoost

O último algoritmo testado nessa etapa foi o *XGBoost.* Os resultados das métricas de avaliação deste algoritmo estão na Figura 33 abaixo. Com ele vemos uma grande amplitude para ambas as métricas, com destaque positivo para os máximos de cada série, chegando a mais de 79% de acurácia e mais de 0,64 de f1 score.

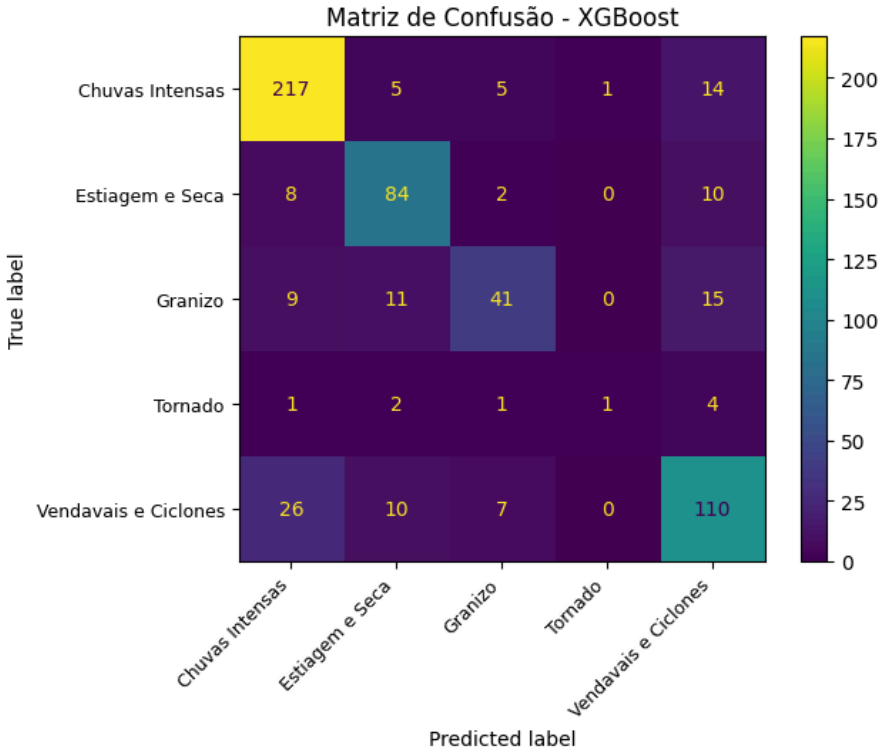
Figura 33 - Resultados da validação cruzada da etapa 2 para *XGBoost*



A matriz de confusão para este algoritmo está na Figura 34. Ela segue o padrão das anteriores, com destaque para um maior número de previsões de chuva para eventos reais de vendavais, e uma redução do caminho contrário: o algoritmo

prevê menos vendavais para eventos reais de chuvas. Um destaque negativo é a grande presença de falsos negativos para o evento de tornado, um ponto alarmante para a previsão de itens, já que tornados são eventos de alta calamidade.

Figura 34 - Matriz de confusão do *XGBoost* na etapa 2



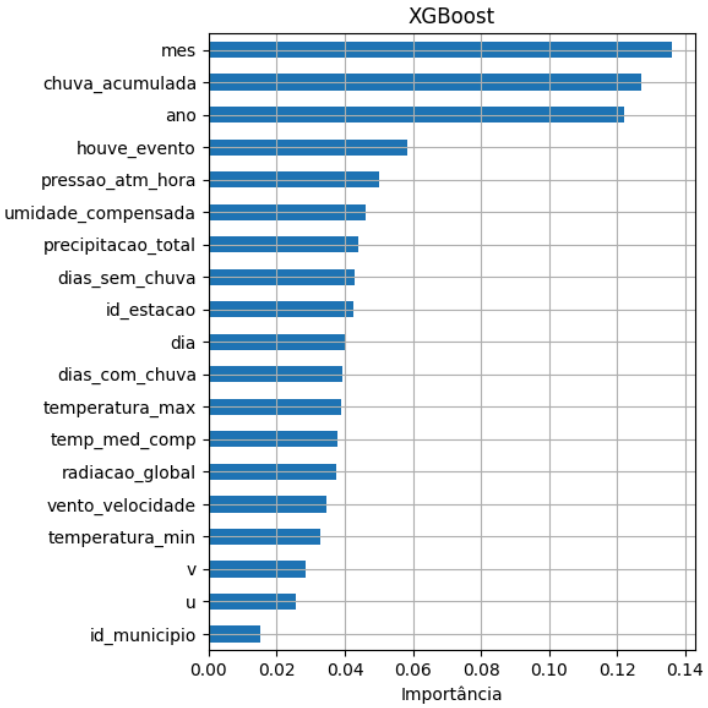
Olhando para os números do experimento, na Tabela 14, o algoritmo se mostra com um bom desempenho dado seus valores altos da proporção de verdadeiros positivos para todos os eventos. Destaque para o evento de tornado, o qual o *XGBoost* foi o único capaz de prevê-lo com assertividade, ao menos uma vez.

Tabela 14 - VPs e FPs do *XGBoost* na etapa 2

Label	**Qnt. Eventos Previstos**	**Verdadeiros Positivos**		**Falsos Positivos**	
		Num.	**%**	**Num.**	**%**
Chuvas Intensas	261	217	83,14%	44	16,86%
Estiagem e Seca	112	84	75,00%	28	25,00%
Granizo	56	41	73,21%	15	26,79%
Tornado	2	1	50,00%	1	50,00%
Vendavais e Ciclones	153	110	71,90%	43	28,10%

Enfim, as importâncias deste algoritmo tem diferenças quando vistas em comparação com os anteriores. As variáveis de tempo ainda estão entre os 3 mais importantes, porém a variável de chuva acumulada se mostra muito decisiva para ajudar a guiar as boas classificações. Novamente as variáveis de vento se mostram pouco expressivas para a classificação dos eventos.

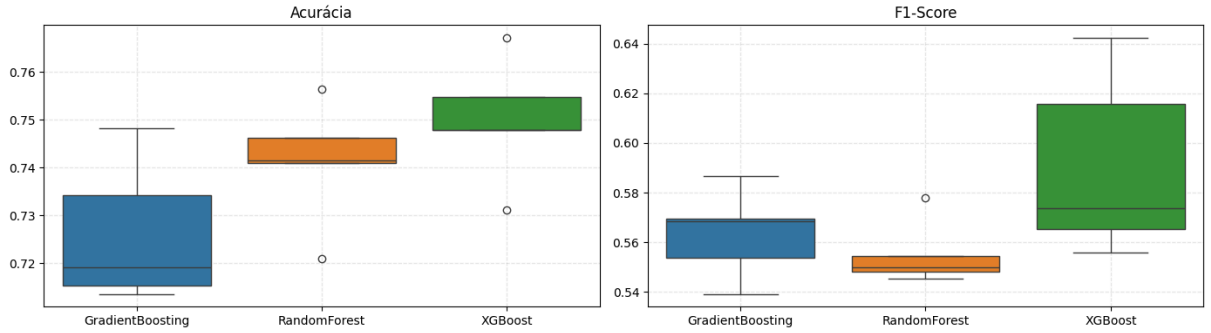
Figura 35 - Importância das *features* para o *XGBoost* na etapa 2



4.6.2.4. Algoritmo Selecionado

A Figura 36 resume os testes de acurácia e *F1-Score* entre cada um dos algoritmos usados na etapa. A partir dela vemos a grande amplitude na métrica de acurácia em todos os algoritmos, mas com valores superiores no *XGBoost,* tendo seu menor ponto sendo maior que a mediana do *Gradient Boosting* e próximo da mediana do *Random Forest*. A análise do F1 demonstra uma alta variabilidade no desempenho do XGBoost, contrastando com seus concorrentes. Estes se mantiveram com uma baixa amplitude de resultados e medianas próximas a 0,55. Por outro lado, o algoritmo otimizado de boosting alcançou valores de F1 de até 0,64, com sua mediana próxima aos valores de pico atingidos pelos demais modelos.

Figura 36 - Resultados da validação cruzada da etapa 2 para todos os algoritmos



Dessa forma, pelo melhor desempenho em acurácia e *F1-Score*, bem como um melhor desempenho unitário para os diferentes tipos de eventos, o modelo selecionado para a etapa foi o *XGBoost*.

4.6.3. Itens Solicitados

A primeira etapa da previsão de itens consiste na previsão se haverá ou não a solicitação de cada um dos 12 itens de assistência humanitária tratados neste trabalho. O problema tratado aqui se trata de uma classificação binária para cada um dos 12 itens, dado os dados meteorológicos e o tipo de evento previsto nas etapas anteriores.

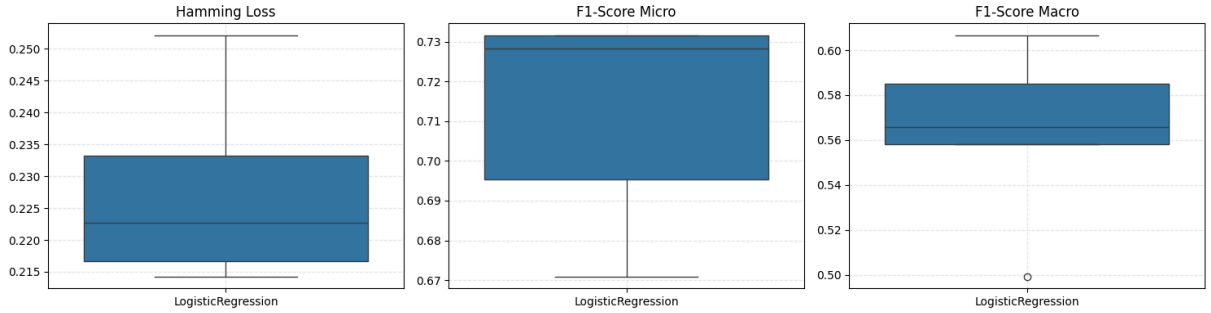
Para essa etapa, são usados os dados de desastres naturais, junto com a tabela de itens solicitados. Para o encaixe dos dados ao problema desenhado, a tabela de itens foi transformada em colunas binárias na tabela de desastres naturais, tendo assim 12 novas colunas na tabela de desastres, uma com o nome de cada item, com valores 0 e 1: 0 para quando o item da coluna não foi solicitado naquele evento; 1 para quando o item foi solicitado na ocorrência.

Aqui será empregada a estratégia de *Multi Output*, disponível na biblioteca *sklearn* (ver Seção 2.3.6). Apesar das diferenças computacionais, essa estratégia segue o mesmo padrão dos testes de seleção apresentados anteriormente, repetindo-os para cada um dos itens na base, testando os algoritmos de *LogisticRegression, RandomForestClassifier e XGBoostClassifier*. O uso dessa estratégia acarreta em mudanças nas métricas de base usadas para cada tipo de problema, sendo aqui usadas as métricas referentes ao problema com múltiplas saídas (*multi output).*

4.6.3.1. Regressão Logística

Tratando-se de uma previsão binária, o primeiro algoritmo a ser testado é a Regressão Logística. Em relação aos valores de proporção de erros, o *Hammimg Loss* apontou um valor de 25,21% de erros na previsão, um valor consideravelmente baixo. Já nos testes de *F1-Score* há uma certa discrepância: ponderando os valores de F1 (*F1-Score Micro*) há um aumento da métrica, enquanto atribuindo pesos iguais para as métricas individuais (*F1-Score Macro*) diminui o valor de F1.

Figura 37 - Resultados da validação cruzada da etapa 3 para regressão logística



Para investigar as razões da diferenciação entre os F1s, as métricas individuais para cada item foram colocadas na Tabela 15. Nela pode-se ver o bom desempenho do algoritmo em produtos muito pedidos, como Cesta básica, kit limpeza e o colchão de solteiro. Porém, itens com menos pedidos e com menos quantidade de amostras não tiveram uma boa métrica, com destaque para o kit água limpa e o reservatório de 20 mil L que não chegaram à 0,3.

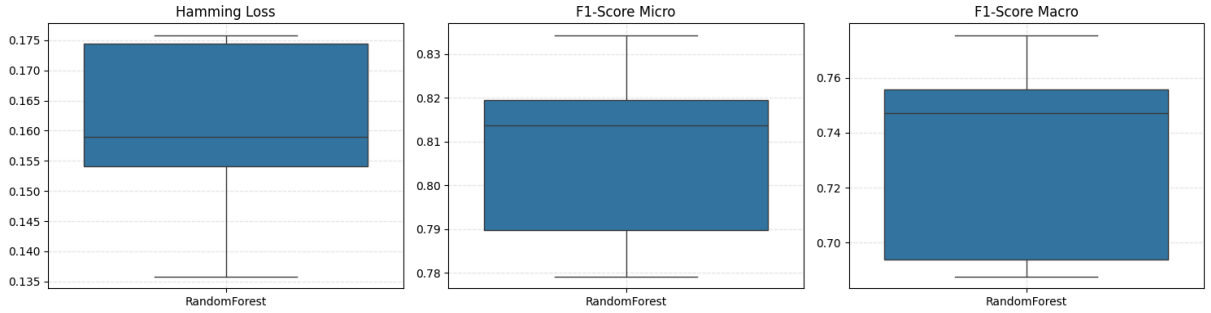
Tabela 15 - *F1-Scores* por item pela regressão logística

Item	Amostras	Precisão	Recall	F1-Score
Cesta básica 7d	84	0,9000	0,8571	0,8781
Colchão Casal	55	0,5373	0,6546	0,5902
Colchão Solteiro	66	0,6667	0,8182	0,7347
Kit aco casal	52	0,5161	0,6154	0,5614
Kit aco solteiro	57	0,5854	0,8421	0,6907
Kit higiene	60	0,7368	0,7000	0,7180
Kit limpeza	75	0,8171	0,8933	0,8535
Kit Água Limpa	49	0,5263	0,2041	0,2941
Reserv. 10 mil L	12	0,8571	0,5000	0,6316
Reserv. 15 mil L	8	0,7500	0,3750	0,5000
Reserv. 20 mil L	14	0,5000	0,1429	0,2222
Reserv. 5 mil	17	0,5600	0,8235	0,6667

4.6.3.2. Random Forest

O segundo modelo testado foi o *Random Forest*. O modelo teve um bom desempenho nas três métricas avaliadas, dispostas na Figura 38. O percentual de erros apontado pela Hamming Loss teve alta amplitude, porém ficou abaixo de 20%, com uma mediana próxima de 16%. Já nas métricas de F1 há um ótimo desempenho do algoritmo. Pela média geral dos F1 de cada item o experimento resultou em valores que ultrapassam 0,76, e sendo mais criterioso com a análise micro, a mediana foi de mais de 0,81, podendo chegar até 0,83.

Figura 38 - Resultados da validação cruzada da etapa 3 para *random forest*



Os valores de F1 para cada um dos itens foram colocados na Tabela 16 abaixo. Nela vemos o bom desempenho do algoritmo com a maioria dos itens. O Reservatório de 15mil L de água teve o pior desempenho, passando por pouco a marca de 0,3 na métrica. Os outros reservatórios com menos de 10 amostras também tiveram seu desempenho afetado, com valores que não ultrapassam 60%. Já o reservatório de 5 mil L, com 14 amostras, teve um desempenho bom, com 0,75. Destaque, também, para os itens de Cesta Básica e Kit Limpeza, os quais ultrapassaram a barreira do 0,9 na métrica.

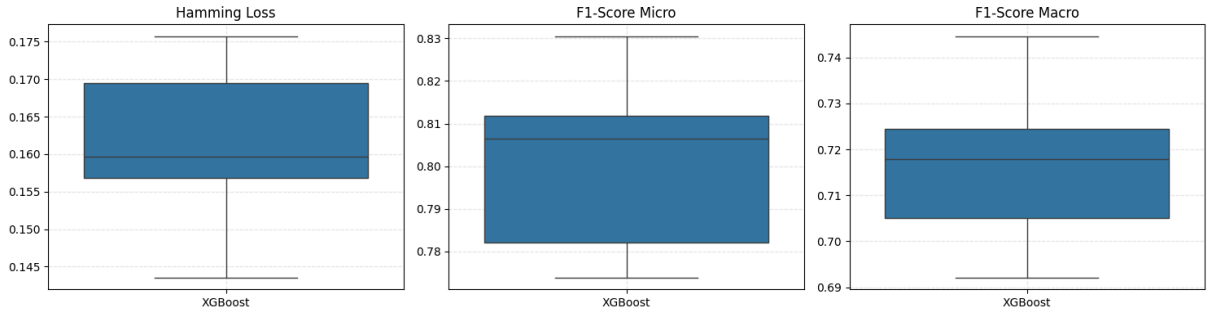
Tabela 16 - *F1-Scores* por item pelo *random forest*

Item	Amostras	Precisão	Recall	F1-Score
Cesta básica 7d	80	0,8876	0,9875	0,9349
Colchão Casal	74	0,8077	0,8514	0,8290
Colchão Solteiro	85	0,8427	0,8824	0,8621
Kit aco casal	66	0,7500	0,8182	0,7826
Kit aco solteiro	71	0,7619	0,9014	0,8258
Kit higiene	66	0,8305	0,7424	0,7840
Kit limpeza	75	0,8675	0,9600	0,9114
Kit Água Limpa	48	0,5349	0,4792	0,5055
Reserv. 10 mil L	9	0,4546	0,5556	0,5000
Reserv. 15 mil L	6	0,2857	0,3333	0,3077
Reserv. 20 mil L	8	0,5000	0,6250	0,5556
Reserv. 5 mil	14	0,7333	0,7857	0,7586

4.6.3.3. XGBoost

Por fim foi testado o *XGBoost* para prever quais itens seriam solicitados dado o evento previsto anteriormente. O algoritmo teve uma baixa taxa de erros, com um piso máximo de 17,56% e valor mediano de 15,96%, valores abaixo do que seus concorrentes. Nas métricas de F1, a análise micro e macro tiveram valores similares ao *Random Forest*, com mediana acima de 0,8 na métrica micro e próximo de 0,72 para o macro.

Figura 39 - Resultados da validação cruzada da etapa 3 para *XGBoost*



As métricas de F1 para cada item previsto estão dispostas na Tabela 17. O desempenho geral foi melhor que o de seus concorrentes, com destaque para números baixos de falsos negativos como aponta os altos valores de recall para a maioria dos itens. Entretanto a quantidade de amostras dos reservatórios de água não se mostraram muito benéficas para o desempenho do algoritmo com esses

itens, se mantendo valores razoáveis para os reservatórios de 10, 15 e 20 mil L, mas com um ótimo valor de 0,8387 para o reservatório de 5 mil L.

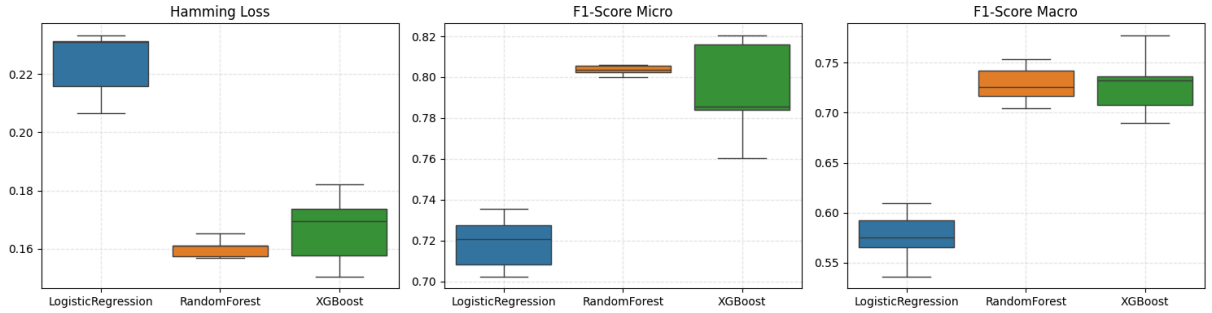
Tabela 17 - *F1-Scores* por item pelo *XGBoost*

Item	Amostras	Precisão	Recall	F1-Score
Cesta básica 7d	83	0,8977	0,9518	0,9240
Colchão Casal	67	0,8356	0,9105	0,8714
Colchão Solteiro	79	0,8471	0,9114	0,8781
Kit aco casal	60	0,7286	0,8500	0,7846
Kit aco solteiro	70	0,7895	0,8571	0,8219
Kit higiene	58	0,8305	0,8448	0,8376
Kit limpeza	80	0,9167	0,9625	0,9390
Kit Água Limpa	46	0,5000	0,4565	0,4773
Reserv. 10 mil L	13	0,6154	0,6154	0,6154
Reserv. 15 mil L	10	0,5000	0,5000	0,5000
Reserv. 20 mil L	11	0,5000	0,4546	0,4762
Reserv. 5 mil	15	0,8125	0,8667	0,8387

4.6.3.4. Algoritmo Selecionado

Com todos os algoritmos testados, suas métricas principais por cada uma das validações cruzadas foram dispostas na Figura 40. Nela, é possível ver a superioridade dos modelos baseados em árvore em relação ao modelo linear, como já havia sido apontado na etapa 1 do *pipeline*. Dessa forma, o algoritmo de regressão logística é facilmente descartado da seleção para esta etapa.

Figura 40 - Resultados da validação cruzada da etapa 3 para todos os algoritmos



Ao analisar os outros algoritmos, é possível ver que ambos têm medianas similares nas métricas de Hammimg Loss e F1 macro, porém o *XGBoost* tem muito mais amplitude do que seu concorrente. Ainda sim, ele não apresentou nenhum F1

abaixo de 0,45 no F1 por item, logo este se mostra mais apto a ser aplicado no *pipeline* do que seu concorrente.

4.6.4. Quantidade dos Itens Solicitados

O *pipeline* tem seu fim com o único problema de regressão dentre as etapas. Os valores a serem previstos são as quantidades solicitadas de cada item. Essa previsão faz uso dos mesmos dados da etapa anterior: o histórico de solicitações da SEDEC.

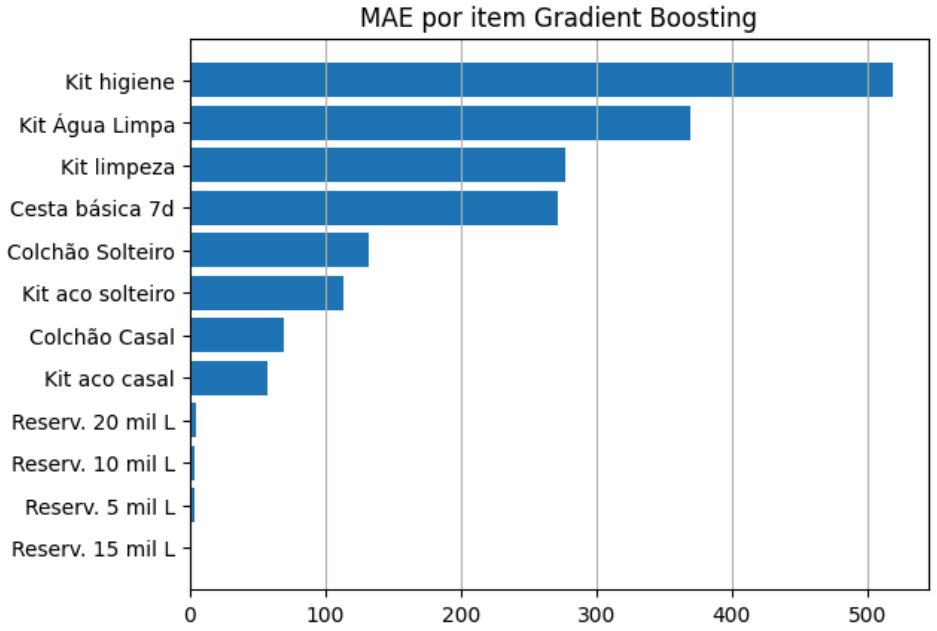
A etapa toma uma estratégia similar à usada na sua predecessora, sendo testado um algoritmo em cada item, porém a biblioteca *sklearn* não suporta o *Multi Output* para algoritmos de regressão, sendo necessário o treino individual de cada algoritmo para cada item.

Nesta etapa, os modelos testados foram o *GradientBoostingRegressor*, *RandomForestRegressor e XGBRegressor.* Os modelos foram avaliados dadas as métricas de regressão adotadas: Erro Médio Absoluto (MAE) e Raiz do Erro Médio Quadrático (RMSE) e a distância de Wasserstein.

4.6.4.1. Gradient Boosting

O primeiro algoritmo testado foi o *Gradient Boosting*. A métrica de MAE por item foi feita a partir da média dos MAEs das 5 validações cruzadas realizadas pelo modelo, e seus resultados estão dispostos na Figura 41. Através dela, podemos ver os grandes erros que o estimador cometeu em itens pedidos em grandes e pequenas quantidades ao longo da série histórica.

Figura 41 - Resultados da validação cruzada da etapa 4 para *gradient boosting*



Para uma análise mais robusta do desempenho do algoritmo, a Tabela 18 apresenta os valores do MAE apresentados na figura acima, do RMSE e da distância de Wass para a comparação da distribuição. A partir dos valores, é possível ver os altos erros no MAE e RMSE, e valores mais tímidos para as distâncias entre distribuições. Destaque para os itens de reservatórios de água, os quais tiveram altos índices nos modelos de classificação, porém tem a melhor performance dentre os itens na etapa de regressão do *pipeline*.

Tabela 18 - Métricas por item pelo *Gradient Boosting*

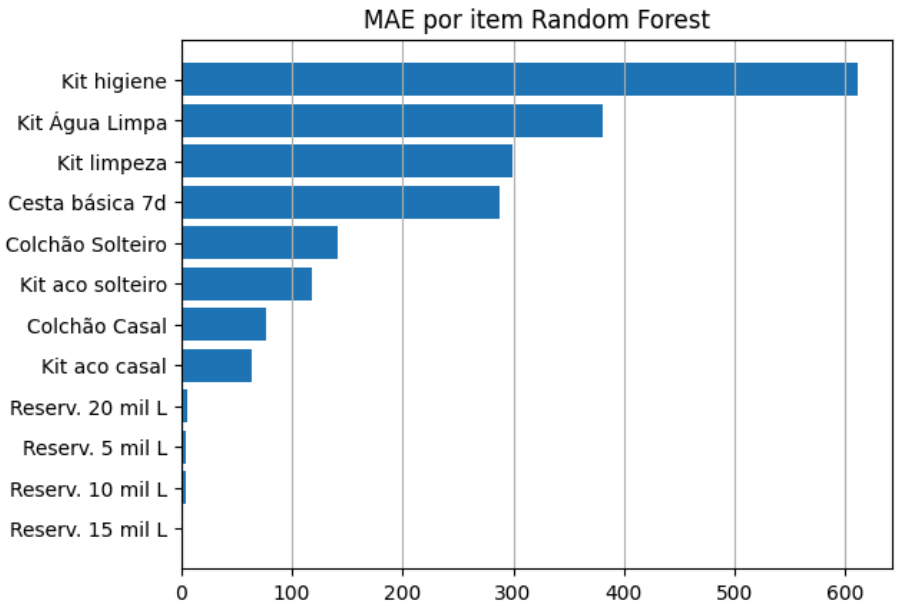
Item	MAE	RMSE	Wass
Cesta básica 7d	271,87	460,18	140,23
Colchão Casal	69,02	107,82	37,57
Colchão Solteiro	132,12	202,27	67,25
Kit aco casal	57,50	86,67	33,52
Kit aco solteiro	113,20	174,45	59,83
Kit higiene	518,82	873,48	271,79
Kit limpeza	276,65	450,18	153,00
Kit Água Limpa	368,91	678,87	205,55
Reserv. 10 mil L	3,87	5,23	2,17
Reserv. 15 mil L	1,73	2,57	0,79
Reserv. 20 mil L	5,10	7,86	3,47
Reserv. 5 mil L	3,86	5,81	2,07

Com isso é possível apontar que os erros absoluto e quadrático são ambos muito altos, e junto a eles, os altos valores de distância entre as distribuições apontam que o algoritmo não performou bem para captar os valores absolutos de solicitação, nem sua distribuição.

4.6.4.2. Random Forest

O segundo estimador testado foi o *Random Forest.* Seu desempenho era esperado que seria melhor do que o algoritmo de boosting testado anteriormente, porém, ao olharmos para o MAE por item na Figura 42, é possível ver valores ainda maiores que do predecessor. Ambos experimentos foram parecidos, com destaque para o Kit Higiene que tem um erro média quase duas vezes maior que o kit água limpa. E novamente é possível ver os itens de reservatório de água com baixos valores de erro.

Figura 42 - Resultados da validação cruzada da etapa 4 para *random forest*



Os valores detalhados das métricas de erro e de distância de distribuições estão na Tabela 19. Todas as métricas têm valores muito altos, com destaque para o RMSE do kit higiene que ultrapassa 1000 itens. Olhando para as distâncias entre as distribuições, elas se mostram proporcionais ao erro médio e acompanham seu aumento e diminuição, com valores que representam, aproximadamente, a metade do erro médio absoluto.

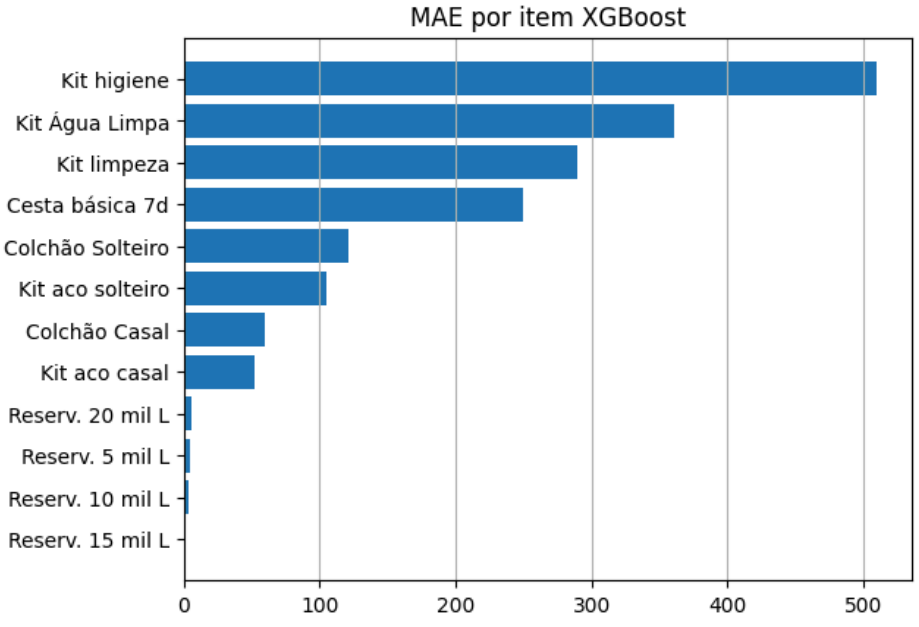
Tabela 19 - Métricas por item pelo *random forest*

Item	MAE	RMSE	Wass
Cesta básica 7d	287,80	492,72	178,42
Colchão Casal	76,96	121,75	40,67
Colchão Solteiro	141,95	220,57	78,13
Kit aco casal	63,77	93,03	33,66
Kit aco solteiro	118,03	179,20	62,03
Kit higiene	611,39	1045,15	345,02
Kit limpeza	299,82	481,14	184,22
Kit Água Limpa	381,33	710,05	238,78
Reserv. 10 mil L	3,75	5,00	2,36
Reserv. 15 mil L	1,83	2,54	1,12
Reserv. 20 mil L	5,06	7,12	2,17
Reserv. 5 mil L	3,91	5,59	2,25

4.6.4.3. XGBoost

Por fim, o algoritmo de regressão do XGBoost foi testado. Os resultados do MAE por item estão na Figura 43. Novamente, os itens de kits tiveram um desempenho ruim, com grandes erros que ficam próximos dos erros apresentados pelo algoritmo irmão, o *Gradient Boosting.* Os kits higiene, água limpa e de limpeza novamente tiveram os maiores erros, sendo o MAE do primeiro significativamente menor que no experimento com o *Random Forest*.

Figura 43 - Resultados da validação cruzada da etapa 4 para *XGBoost*



Contudo, os valores da Tabela 22 mostram o ponto de superioridade do *XGBoost* em relação a seus concorrentes. A métrica de Wasserstein em 3 dos 12 itens previstos ficou abaixo de 150, e não houveram distâncias maiores de 300, apontando que este algoritmo conseguiu captar as distribuições das quantidades de cada item melhor que seus concorrentes.

Tabela 20 - Métricas por item pelo *XGBoost*

Item	MAE	RMSE	Wass
Cesta básica 7d	250,16	451,03	101,12
Colchão Casal	60,12	101,61	24,69
Colchão Solteiro	121,51	195,79	52,00
Kit aco casal	52,50	82,05	21,69
Kit aco solteiro	104,80	177,58	41,47
Kit higiene	509,98	883,78	245,04
Kit limpeza	289,28	478,36	117,46
Kit Água Limpa	361,30	733,76	154,47
Reserv. 10 mil L	3,50	4,84	2,10
Reserv. 15 mil L	1,55	2,27	0,90
Reserv. 20 mil L	5,18	7,88	2,45
Reserv. 5 mil L	4,21	6,49	1,71

4.6.4.4. Algoritmo Selecionado

Todos os algoritmos tiveram um desempenho abaixo da média nesta etapa. O *Gradient Boosting* mostrou erros muito altos nas quantidades solicitadas, com destaque para o alto valor dos kits de limpeza e higiene. Entretanto, ele foi capaz de captar as distribuições de quantidades de forma satisfatória.

Já o *Random Forest* teve os maiores erros dos 3 estimadores. Com um RMSE do kit higiene ultrapassando as 1000 unidades e o maior MAE dos três algoritmos para o item: 611,89. Além disso, as distâncias de wasserstein tiveram valores altos, com 4 valores acima de 150. Com isso, o algoritmo foi descartado da seleção.

Por fim, o *XGBoost* performou de forma similar ao *Gradient Boosting.* Os erros foram altos, com destaque para um RMSE do kit higiene acima de 800 unidades. Entretanto, o estimador foi o que melhor captou a distribuição das

quantidades dos itens, como é possível ver pelos seus valores baixos de distância de Wasserstein.

Olhando as médias dos valores das métricas de distância entre o *Gradient* e o *XGBoost,* temos que o primeiro tem uma distância média entre distribuições de 81,44, enquanto o segundo tem uma média de 63,76. Assim é seguro selecionar o *XGBoost* como algoritmo da etapa, dado que os valores pontuais previstos dos itens são passíveis de correção via partícula atenuante, porém a forma do algoritmo perceber a distribuição não pode ser alterada externamente.

4.7. PROPOSTA DE USO

Com os diferentes estimadores testados e selecionados para cada uma das etapas, resta juntá-los em um único *pipeline* para seu uso no planejamento operacional da SEDEC. Dos algoritmos testados anteriormente, o XGBoost teve um desempenho melhor em todas as etapas do *pipeline*, sendo o escolhido para todo o trabalho. No Quadro 11 abaixo, estão dispostas as justificativas desse desempenho em cada uma das etapas.

Quadro 11 - Quadro resumo da seleção de modelos

Macro Etapa	Etapa	Algoritmo Selecionado	Justificativa
Previsão de Eventos de Desastres	Haverá desastre	*XGBoost*	Obteve a melhor métrica de F1-Score (0,573), sendo superior ao Random Forest (cerca de 0,50), e alcançando a segunda melhor acurácia (91,5%).
	Classe do desastre	*XGBoost*	Apresentou desempenho superior em acurácia (75%) e F1-Score (mediana entre 0,56 e 0,64), com melhor capacidade de prever eventos de baixa ocorrência, como o Tornado.
Previsão de Itens de Assistência	Quais itens	*XGBoost*	Teve melhor desempenho nas métricas de classificação multi-output, com F1-Score Micro e Macro superiores (mediana acima de 0,8 e próximo de 0,72, respectivamente) e não apresentou nenhum F1-Score por item abaixo de 0,45.
	Quantidade de cada item	*XGBoost*	Embora os erros (MAE e RMSE) tenham sido altos, foi o que melhor capturou a distribuição das quantidades solicitadas, apresentando a menor média de Distância de Wasserstein (63,76), mais próxima da distribuição original.

O modelo obteve altos valores em ambos os F1-Score para prever quais itens serão solicitados, e uma taxa de erro de aproximadamente 16% como apontado pelo

resultado de *Hamming Loss* do algoritmo. Em relação às quantidades solicitadas, o modelo apresentou desafios: valores de erros muito altos em relação às quantidades reais. Entretanto, ao olhar no aspecto geral, alguns valores pontuais foram previstos erroneamente, porém a distribuição das quantidades solicitadas, pela distância de Wass, se manteve numa média de 80 itens da distribuição original, apontando que apesar dos erros pontuais, as proporções das quantidades solicitadas foram mantidas próximas.

As variáveis usadas para o treinamento do modelo foram escolhidas de forma que se alinhem às variáveis de dados de projeções climáticas que, diferentemente de previsões climáticas, são simulações da resposta do sistema climático a cenários hipotéticos de emissões futuras de gases de efeito estufa e aerossóis (IPCC, 2023).

O modelo de projeção levantado para ser usado como dados de entrada para o *pipeline* é o ETA. O modelo ETA é um modelo regional de projeção aplicado nos estudos nacionais sobre projeção de clima, e com ele é levada em conta as características topográficas do ambiente analisado, favorecendo as previsões mais localizadas (Lyra et al., 2022). Os dados do ETA estavam disponíveis através do Portal de Projeções Climáticas do INPE, porém, devido a alterações na infraestrutura tecnológica do instituto, eles não puderam ser coletados para serem testados neste estudo.

Dentro do Portal, é possível selecionar as coordenadas específicas dentro do estado de SC, como por exemplo os centros de cada cidade do estado, e o período projetado para análise junto da variável desejada. Coletando os dados dessa forma, eles podem ser aplicados ao modelo e gerar previsões para o curto e longo prazo, auxiliando o planejamento operacional da SEDEC/SC.

5. CONCLUSÃO E TRABALHOS FUTUROS

5.1. CONCLUSÕES

As projeções do IPCC sobre as mudanças climáticas apontam que o número de desastres naturais aumentará vertiginosamente nas próximas décadas (IPCC, 2023). Logo, haverão mais pessoas afetadas por tais desastres, gerando uma maior necessidade de assistencialismo em situações de emergência provido por entidades como a SEDEC.

Nesse contexto, o presente trabalho se propôs a desenvolver um modelo computacional capaz de estimar qual será a demanda do assistencialismo necessário em relação à itens solicitados para a SEDEC em situações de desastres naturais. O projeto atingiu este objetivo a partir da criação de um *pipeline* preditivo que, treinado via de dados históricos de estações meteorológicas, de registros de eventos de desastres naturais, e da solicitações de itens de assistência nesses eventos, gera uma previsão de quando haverá desastres, qual desastre ocorrerá, quais itens serão solicitados e quantas unidades de cada item serão solicitados.

Para conseguir desenvolver tal *pipeline*, primeiramente o trabalho teve de organizar dados de desastres naturais e itens solicitados para a SEDEC. Os dados de desastres coletados do portal do CEMADEN, e os dados de solicitações de itens coletados da SEDEC, foram tratados e modelados seguindo a modelagem de Kimball, o que garantiu à essas bases um relacionamento facilitado com outras bases, como as do BDMET, e consequentemente uma maior facilidade para seu uso nos objetivos seguintes.

Com os dados modelados, foi possível atingir o segundo objetivo específico: analisar a distribuição de desastres naturais pelo estado de Santa Catarina. Como os eventos do estado e os eventos que geravam a solicitação de itens eram majoritariamente climatológicos e meteorológicos, os eventos hidrológicos puderam ser descartados para garantir que não houvessem grandes impurezas nos algoritmos treinados.

Pelas análises exploratória realizada nos dados e os direcionamentos tirados dela, foi possível iniciar o teste de diferentes modelos de ML frente aos dados organizados, atingindo o terceiro objetivo específico. Foram testados ao todo 7 algoritmos de ML diferentes, sendo 4 usados nos problemas de classificação (Regressão Logística, *Random Forest Classifier, Gradient Boosting Classifier* e

XGBoost Classifier), e 3 utilizados no problema de regressão da última etapa (*Random Forest Regressor, Gradient Boosting Regressor* e *XGBoost Regressor).*

A partir desses algoritmos, foi estruturado, na primeira macroetapa do *pipeline, um modelo* para a previsão da ocorrência e classificação de desastres naturais, com base nos dados tratados do CEMADEN e do INMET. Com isso o penúltimo objetivo específico da pesquisa foi atingido, gerando previsões mais de 75% de acurácia e valores acima de 0,6 de *F1-Score* com o *XGBoost.*

Este algoritmo também se mostrou superior em todas as outras etapas do *pipeline*, e foi o selecionado como melhor algoritmo em todas elas. Para a primeira etapa, o ele teve a segunda melhor acurácia (91,5%), porém se destacou como tendo os melhores resultados de *F1-Score* (0,573). Na etapa 2, o algoritmo também se sobressaiu com folga dos concorrentes, com uma acurácia de 75% e *F1-Score* variando de 0,56 a 0,64, sendo seu mínimo maior que as medianas dos outros algoritmos. Na terceira etapa, o *Random Forest* teve um desempenho extremamente similar ao *XGBoost*, porém teve valores abaixo de 0,45 para os *F1-Scores* de cada item, sendo descartado. Por fim, na etapa de regressão, todos os algoritmos tiveram erros próximos quanto às quantidades, porém o algoritmo de Chen e Guestrin (2016) foi o melhor em captar a distribuição das solicitações.

O modelo proposto, então, é um *pipeline* composto de 4 etapas de previsão executadas pelo algoritmo *XGBoost*. O *pipeline* obteve uma acurácia para apontar quais itens serão solicitados de 75%, um valor alto, que está longe de sofrer de *underfitting* e *overfitting*. Em relação às quantidades solicitadas, o modelo apresentou desafios: valores de erros muito altos em relação às quantidades reais. Entretanto, ao olhar no aspecto geral, alguns valores pontuais foram previstos erroneamente, porém a distribuição das quantidades solicitadas, pela distância de Wass, se manteve numa média de 80 itens da distribuição original, apontando que apesar dos erros pontuais, as proporções das quantidades solicitadas foram mantidas próximas.

5.2. SUGESTÕES PARA TRABALHOS FUTUROS

Como sugestões para trabalhos futuros, recomenda-se o teste do *pipeline* frente aos dados projetados pelo INPE do clima de Santa Catarina. Dessa forma será possível avaliar como o modelo trabalha frente a esses dados, e posteriormente, validar sua eficácia em prever os itens solicitados.

Posteriormente a isso, pode-se sugerir a evolução deste *pipeline*. Como ponto de partida, é recomendado a otimização dos hiperparâmetros do XGBoost, através de funcionalidades como a *Grid Search*, da biblioteca *sklearn*, sendo recomendado que isso seja feito em sistemas de computação em nuvem, a fim de garantir que essa otimização seja feita de forma mais rápida.

Outra abordagem possível para sua evolução é a aplicação de métodos de clusterização durante a etapa de transformação dos dados, criando novas *features* nas tabelas que ajudem o modelo a performar melhor. Junto a isso, também é recomendado os testes com algoritmos de modelagem não supervisionada, como redes neurais.

Ainda sobre o modelo, o tratamento do desbalanceamento de classes nos dados de treino é uma melhoria que deve ser aplicada, a qual pode ser tratada com métodos de sub ou super amostragem, por métodos mais robustos como o SMOTE (Super-amostragem sintética de minorias, adaptado do inglês), ou pela adição de critérios de pesos às classes, o que é possível para os algoritmos aplicados neste trabalho através do atributo *sample_weights,* do método *fit,* nativo da biblioteca *sklearn*.

Para a última macroetapa do *pipeline*, que prevê a solicitação de itens propriamente, é recomendada a adição de mais *features* e bases de dados para tais previsões. Critérios socioeconômicos, perfil da população e topografia da região são critérios que devem ser analisados por métodos de AED para confirmar sua relação com as solicitações de itens, para então serem adicionados como bases de treino para o modelo.

Outros dados, junto da topografia da região, como taxa de impermeabilização do solo, tipos de solo, áreas sob influência da agropecuária, presença de grandes corpos de água, entre outros, devem ser considerados em pesquisas futuras também. Isso porque, com esses dados, eventos hidrológicos podem se tornar mais fáceis de prever, elevando assim a capacidade do modelo de antecipar desastres para o planejamento da SEDEC.

Por fim, a recomendação é a criação de um novo painel de controle para a SEDEC/SC, que apresentará os dados históricos e do modelo preditivo para facilitar o controle, avaliação e tomada de decisão dos profissionais da secretaria sobre os itens de assistência solicitados.

REFERÊNCIAS

ALTAY, N. et al. Innovation in humanitarian logistics and supply chain management: a systematic review. **Annals of Operations Research**, [S. l.], v. 335, n. 3, p. 965-987, abr. 2024. DOI: 10.1007/s10479-023-05208-6. Disponível em: http://link.springer.com/10.1007/s10479-023-05208-6. Acesso em: 3 nov. 2025.

BALCIK, B.; BOZKIR, C. D. C.; KUNDAKCIOGLU, O. E. A literature review on inventory management in humanitarian supply chains. **Surveys in Operations Research and Management Science**, [S. l.], v. 21, n. 2, p. 101-116, 2016.

BASE DOS DADOS. **Sobre nós**. 2024. Disponível em: https://basedosdados.org/about-us. Acesso em: 24 maio 2025.

BISHOP, C. M. **Pattern recognition and machine learning**. New York: Springer, 2006.

BRASIL. Ministério da Integração e do Desenvolvimento Regional. Secretaria de Proteção e Defesa Civil. **Atlas Digital de Desastres no Brasil**. Brasília, DF: MIDR; Florianópolis: CEPED UFSC, 2023.

BRIAN, Magadi Wanner; SHALE, Noor. Role of humanitarian logistics on supply chain performance in non-governmental organizations in Kenya: a case of ACTED Kenya. **International Journal of Management and Commerce Innovations** , [S. l.], v. 5, n. 2, p. 277-287, 2017.

BREIMAN, L. Random Forests. **Machine Learning**, [S. l.], v. 45, n. 1, p. 5–32, 2001.

CASTRO, Antônio Luiz Coimbra de. **Glossário de defesa civil: estudos de riscos e medicina de desastres**. 5. ed. [Brasília, DF]: Ministério da Integração Nacional, Secretaria Nacional de Defesa Civil, [20-?].

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. In: **ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING**, 22., 2016, San Francisco. **Anais** [...]. New York: ACM, 2016. p. 785–794. DOI: 10.1145/2939672.2939785.

CISLAGHI, Tatiane; FERNANDES, Elieti Biques; DORISCAT, Estherlin. **Logística humanitária:** práticas e desafios em tempos de tensões geopolíticas, desigualdades sociais e crise ambiental. São Bernardo do Campo, v. 20, n. 40, p. 127-164, 2024.

CONFEDERAÇÃO NACIONAL DE MUNICÍPIOS (CNM). **Estudo técnico sobre danos e prejuízos causados por desastres no Brasil entre 2013 a 2023**. Brasília, DF: CNM, 2023.

CUNNINGHAM, P.; CORD, M.; DELANY, S. J. Supervised learning. In: CORD, M.; CUNNINGHAM, P. (ed.). **Machine learning techniques for multimedia**. Berlin; Heidelberg: Springer, 2008. p. 21-49. (Cognitive Technologies). DOI: 10.1007/978-3-540-75171-7_2.

DATACAMP. **Machine Learning with Tree-Based Models in Python**. [S. l.]: DataCamp, 2023. Curso online. Disponível em:

https://campus.datacamp.com/pt/courses/machine-learning-with-tree-based-models-in-python. Acesso em: 25 maio 2025.

DATACAMP. **Supervised Learning with scikit-learn**. [S. l.]: DataCamp, 2023. Curso online. Disponível em: https://campus.datacamp.com/pt/courses/supervised-learning-with-scikit-learn. Acesso em: 25 maio 2025.

DAVENPORT, Thomas H.; HARRIS, Jeanne G. **Competição analítica: vencendo através da nova ciência**. Rio de Janeiro: Alta Books, 2018.

FAYYAD, U. et al. From data mining to knowledge discovery: an overview. In: FAYYAD, U. et al. (ed.). **Advances in Knowledge Discovery and Data Mining**. Menlo Park: AAAI Press; MIT Press, 1996. p. 1-34.

FRIEDMAN, J. H. Greedy Function Approximation: A Gradient Boosting Machine. **Annals of Statistics**, [S. l.], v. 29, n. 5, p. 1189–1232, 2001.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

GROBMAN, Ricardo Arippol; CUGNASCA, Prof. Dr. Carlos Eduardo. Análise preditiva de demanda com machine learning, aplicado ao varejo com sazonalidade. In: **SIMPÓSIO DE ENGENHARIA DE PRODUÇÃO**, 13., 2025, João Pessoa, PB. **Anais** [...]. João Pessoa, PB: [s. n.], 2024. ISSN: 2318-9258. Disponível em: https://dspace.sti.ufcg.edu.br/handle/riufcg/43850.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. New York: Springer, 2009.

INSTITUTO NACIONAL DE METEOROLOGIA (INMET). **Normais climatológicas do Brasil: 1991–2020**. Brasília, DF: INMET, 2023. Disponível em: https://portal.inmet.gov.br/uploads/normais/NORMAISCLIMATOLOGICAS.pdf. Acesso em: 24 maio 2025.

INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE (IPCC). **AR6 Synthesis Report: Climate Change 2023**. Geneva: IPCC, 2023. Disponível em: https://www.ipcc.ch/report/ar6/syr/. Acesso em: 22 jun. 2025.

INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE (IPCC). **Climate Change 2013: The Physical Science Basis**. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press, 2013. 1535 p.

JAMES, G. et al. **An introduction to statistical learning: with applications in R**. New York: Springer, 2013.

KIMBALL, R.; ROSS, M. **The data warehouse toolkit: the definitive guide to dimensional modeling**. 3. ed. Indianapolis: Wiley, 2013.

LYRA, A. de A. et al. **Manual Modelo Eta – Versão 1.4.2**. São Paulo: CPTEC/INPE, 2022. 103 p.

MATTAR, F. N. **Pesquisa de marketing**. 3. ed. São Paulo: Atlas, 2001.

MEIRIM, H. **Logística humanitária e logística empresarial**. Administradores.com, [S. l.], 2006. Disponível em: https://www.administradores.com.br/artigos/logistica-humanitaria-logistica-empresarial. Acesso em: 19 abr. 2025.

MIGUEL, P. A. C. Estudo e caso na engenharia de produção: estruturação e recomendações para sua condução. In: **CONGRESSO BRASILEIRO DE CUSTOS**, 13., 2006, Belo Horizonte, MG. **Anais** [...]. [S. l.: s. n.], 2006.

NOGUEIRA, C.; GONÇALVES, M. B.; NOVAES, A. G. **Logística humanitária e logística empresarial: relações, conceitos e desafios**. 2012. Trabalho de Conclusão de Curso (Especialização em Engenharia de Produção) - Programa de Pós-graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2012.

NOGUEIRA, C.; GONÇALVES, M. B.; NOVAES, A. G. **A LOGÍSTICA HUMANITÁRIA E MEDIDAS DE DESEMPENHO:** A PERSPECTIVA DA CADEIA DE ASSISTÊNCIA HUMANITÁRIA. 2008. 12 f. Universidade Federal de Santa Catarina, Florianópolis, 2008.

PAINEL BRASILEIRO DE MUDANÇAS CLIMÁTICAS (PBMC). **Segundo Relatório de Avaliação Nacional sobre Mudanças Climáticas**. Rio de Janeiro: PBMC, 2014. Disponível em: https://www.pbmc.coppe.ufrj.br/. Acesso em: 22 jun. 2025.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, [S. l.], v. 12, p. 2825–2830, 2011.

REJEB, Abderahman; REJEB, Karim; SIMSKE, Steve; TREIBLMAIER, Horst. Humanitarian Drones: A Review and Research Agenda. **Internet of Things**, [S. l.], v. 16, p. 1-20, 2021. DOI: 10.1016/j.iot.2021.100434. Disponível em: https://doi.org/10.1016/j.iot.2021.100434. Acesso em: 11 nov. 2025.

RUSHTON, A.; CROUCHER, P.; BAKER, P. **The Handbook of Logistics & Distribution Management: understanding the supply chain**. 4. ed. London: Kogan Page, 2010.

SANTA CATARINA. Secretaria de Estado da Defesa Civil. **Relatório técnico estiagem no Oeste Catarinense: diagnóstico e resiliência**. Florianópolis: SEDEC, 2017.

SCIKIT-LEARN. **3.3. Model evaluation: quantifying the quality of predictions – Which scoring function should I use?** 2025. Disponível em: https://scikit-learn.org/stable/modules/model_evaluation.html#which-scoring-function-should-i-use. Acesso em: 22 jun. 2025.

SCIKIT-LEARN. **Multiclass-multioutput classification**. [2025]. Disponível em: https://scikit-learn.org/stable/modules/multiclass.html#multiclass-multioutput-classification. Acesso em: 13 out. 2025.

SCIKIT-LEARN. **sklearn.metrics.hamming_loss**. [2025]. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.hamming_loss.html. Acesso em: 13 out. 2025.

SCIKIT-LEARN. **Tuning the hyper-parameters of an estimator**. 2025. Disponível em: https://scikit-learn.org/stable/modules/grid_search.html. Acesso em: 22 jun. 2025.

SCIPY. **scipy.stats.wasserstein_distance (versão 1.11.4). Documentação oficial**. [2025]. Disponível em: https://docs.scipy.org/doc/scipy-1.11.4/reference/generated/scipy.stats.wasserstein_distance.html. Acesso em: 13 out. 2025.

SILVA, E. L.; MENEZES, E. M. **Metodologia da pesquisa e elaboração de dissertação**. 4. ed. Florianópolis: UFSC, 2005. 138 p.

SILVA, J.; OLIVEIRA, M. Estudo sobre técnicas de engenharia naval. In: **SIMPÓSIO DE ENGENHARIA MARÍTIMA**, 1., 2019, São Paulo. **Anais** [...]. São Paulo: Blucher, 2019. Disponível em: https://pdf.blucher.com.br/marineengineeringproceedings/spolm2019/104.pdf. Acesso em: 27 abr. 2025.

SILVEIRA, B. C. da. **Validação de modelos climáticos globais e análise de projeções futuras para o Rio Grande do Sul**. 2021. Trabalho de Conclusão de Curso (Graduação em Geografia) – Instituto de Geociências, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2021.

SOUZA, F. de. **Gargalos burocráticos na logística humanitária pela Defesa Civil de Santa Catarina em eventos hídricos extremos**. 2012. Monografia (Graduação em Administração) – Universidade do Estado de Santa Catarina, Florianópolis, 2012.

TSOUMAKAS, Grigorios; KATAKIS, Ioannis. Multi-Label Classification: An Overview. **International Journal of Data Warehousing & Mining**, [S. l.], v. 3, n. 3, p. 1-13, 2007. Disponível em: https://www.researchgate.net/publication/273859036. Acesso em: 13 out. 2025.

VALLENDER, S. S. Calculation of the Wasserstein Distance Between Probability Distributions on the Line. **Theory of Probability & Its Applications**, [S. l.], v. 18, n. 4, p. 784–786, 1974.

WORLD WEATHER ATTRIBUTION. **Likely influence of climate change in severe flooding in Santa Catarina, January 2025**. [S.l.]: World Weather Attribution, 2025. Disponível em: https://noticias.paginas.ufsc.br/files/2025/02/WWA_FloodsSC.pdf. Acesso em: 11 nov. 2025.

XGBOOST DEVELOPERS. **XGBoost Python Package – Setting Parameters**. 2025. Disponível em: https://xgboost.readthedocs.io/en/stable/python/python_intro.html#setting-parameters. Acesso em: 22 jun. 2025.

XGBOOST: eXtreme Gradient Boosting. [S. l.]: DMLC, 2024. Disponível em: https://xgboost.readthedocs.io/en/stable/. Acesso em: 11 set. 2025.

.

APÊNDICE A – DESCRITIVO FATO SOLICITAÇÕES

Nome da Tabela:	Fato Itens Solicitados
Descrição:	Tabela com os registros de eventos que geraram solicitações de itens de assistência à SEDEC/SC e a quantidade de itens diferentes solicitados por cada um desses eventos.
Fonte(s):	Tabelas de registros manuais dos itens solicitados da SEDEC.
Total de entradas:	800 entradas.

Descritivo das Colunas

#	Coluna	Descrição	Tipo	Unidade de Medida	Transformação Aplicada
1	id_ocorrencia	ID da ocorrência registrada, baseado na data, evento e localização	String	-	-
2	data	Data da solicitação	Date	-	-
3	id_municipio	ID Município - IBGE 7 Dígitos	Int	-	-
4	id_micro	ID Microrregiao - IBGE 5 dígitos	Int	-	-
5	name_muni	Nome do Município	String	-	-
6	name_micro	Nome da Microrregião	String	-	-
7	evento	Nome do evento de desastre	String	-	-
8	qnt_itens	Quantidade de itens diferentes solicitados no evento	Int	Itens	Contagem dos itens diferentes solicitados por cada registro de evento diferente.

APÊNDICE B – DESCRITIVO FATO ITENS SOLICITADOS

Nome da Tabela:	Fato Itens Solicitados
Descrição:	Tabela com os itens demandados e suas quantidades por cada ocorrência registrada na base histórica da SEDEC.
Fonte(s):	Tabelas de registros manuais dos itens solicitados da SEDEC.
Total de entradas:	2.810 entradas.

Descritivo das Colunas

#	Coluna	Descrição	Tipo	Unidade de Medida	Transformação Aplicada
1	id_ocorrencia	ID da ocorrência registrada, baseado na data, evento e localização	String	-	-
2	item	Nome do item solicitado	String	-	-
3	val_unit	Valor unitário do item	Float	R$	-
4	quant	Quantidade demandada do item	Int	Item	-
5	val_total	Valor total da solicitação	Float	R$	-

APÊNDICE C – DESCRITIVO FATO EVENTOS

Nome da Tabela:	Fato Eventos
Descrição:	Tabela com os registros dos eventos de desastres naturais ocorridos no estado de Santa Catarina de 1991 à 2024
Fonte(s):	Observatório de Desastres Naturais e Tabelas de registros manuais dos itens solicitados da SEDEC.
Total de entradas:	7.976 entradas.

Descritivo das Colunas

#	Coluna	Descrição	Tipo	Unidade de Medida	Transformação Aplicada
1	id_ocorrencia	ID da ocorrência registrada, baseado na data, evento e localização	String	-	-
2	id_estacao	ID da estação meteorológica que mapeia a área do evento	String	-	-
3	id_municipio	ID Município - IBGE 7 Dígitos	Int	-	-
4	id_microrregiao	ID Microrregiao - IBGE 5 dígitos	Int	-	-
5	data	Data do evento	Date	-	-
6	name_muni	Nome do Município	Int	-	-
7	name_micro	Nome da Microrregião	Float	-	-
8	evento	Nome do evento de desastre	String	-	-
9	grupo_evento	Grupo ao qual o evento pertence, de acordo com o Manual de Desastres da SEDEC (2003)	String	-	-
10	dh_total	Total de pessoas afetadas (dano humano) decorrente do desastre	Int	Pessoas	Soma dos diferentes tipos de dano humano por cada ocorrência.
11	dm_total	Total de estruturas e objetos materiais afetadas pelo desastre.	Int	Materiais	Soma dos diferentes tipos de dano material por cada ocorrência.

APÊNDICE D – DESCRITIVO FATO METEOROLÓGICO DIÁRIO

Nome da Tabela:	Fato Meteorológico Diário
Descrição:	Tabela com os registros das 24 estações meteorológicas em Santa Catarina diariamente de 2003 à 2024.
Fonte(s):	BDMEP
Total de entradas:	126.402 entradas.

Descritivo das Colunas

#	Coluna	Descrição	Tipo	Unidade de Medida	Transformação Aplicada
1	id_estacao	ID da estação meteorológica	Int	-	-
2	id_municipio	ID Município - IBGE 7 Dígitos	Int	-	-
3	id_municipio_nome	Nome do município	String	-	-
4	data	Data da leitura	Date	-	-
6	precipitacao_total	Precipitação total no dia	Float	mm	-
7	pressao_atm_hora	Pressão atmosférica média ao nível da estação	Float	hPa	-
8	radiacao_global	Radiação global média do dia	Float	J/cm³	-
9	temperatura_max	Temperatura máxima do dia	Float	Graus Celsius (°C)	-
10	temperatura_min	Temperatura mínima do dia	Float	Graus Celsius (°C)	-
11	vento_velocidade	Velocidade do vento	Float	hPa	-
12	v	Componente V do vento	Float	m/s	Equação 33
13	u	Componente U do vento	Float	m/s	Equação 33
14	temp_med_comp	Temperatura média compensada do dia	Float	m/s	Equação 31
15	umidade_compensada	Umidade média compensada do dia	Float	%	Equação 32
16	dias_com_chuva	Dias com chuva até o respectivo dia da leitura	Float	Dias	Contagem dos dias anteriores com precipitação acima de 1mm
17	dias_sem_chuva	Dias sem chuva até o respectivo dia da leitura	Float	Dias	Contagem dos dias anteriores com precipitação de até 1mm
18	chuva_acumulada	Precipitação acumulada dos ultimos dias em que houveram chuva	Float	mm	Soma da precipitação diária dos últimos n dias com chuva

19	houve_evento	Caso verdadeira, indica que houve evento de desastre no dia em questão	Boolean	-	Mapeamento com a Fato Eventos

APÊNDICE E – DESCRITIVO FATO DESASTRES NATURAIS

Nome da Tabela:	Fato Desastres Naturais
Descrição:	Tabela com os registros dos indicadores meteorlógicos associados a cada evento.
Fonte(s):	BDMEP, Fato Eventos
Total de entradas:	5.267 entradas.

Descritivo das Colunas

#	Coluna	Descrição	Tipo	Unidade de Medida	Transformação Aplicada
1	id_ocorrencia	ID da ocorrência registrada, baseado na data, evento e localização	String	-	-
2	id_municipio	ID Município - IBGE 7 Dígitos	Int	-	-
3	id_estacao	ID da estação meteorológica	Int	-	-
4	evento	Nome do evento de desastre	String	-	-
6	precipitacao_total	Precipitação total no dia	Float	mm	-
7	pressao_atm_hora	Pressão atmosférica média ao nível da estação	Float	hPa	-
8	radiacao_global	Radiação global média do dia	Float	J/cm³	-
9	temperatura_max	Temperatura máxima do dia	Float	Graus Celsius (°C)	-
10	temperatura_min	Temperatura mínima do dia	Float	Graus Celsius (°C)	-
11	vento_velocidade	Velocidade do vento	Float	hPa	-
12	v	Componente V do vento	Float	m/s	Equação 33
13	u	Componente U do vento	Float	m/s	Equação 33
14	temp_med_comp	Temperatura média compensada do dia	Float	m/s	Equação 31
15	umidade_compensada	Umidade média compensada do dia	Float	%	Equação 32
16	dias_com_chuva	Dias com chuva até o respectivo dia da leitura	Float	Dias	Contagem dos dias anteriores com precipitacao acima de 1mm
17	dias_sem_chuva	Dias sem chuva até o respectivo dia da leitura	Float	Dias	Contagem dos dias anteriores com precipitacao de até 1mm

18	chuva_acumulada	Precipitação acumulada dos ultimos dias em que houveram chuva	Float	mm	Soma da precipitação diária dos últimos n dias com chuva
19	houve_evento	Caso verdadeira, indica que houve evento de desastre no dia em questão	Boolean	-	Mapeamento com a Fato Eventos
20	dia	Dia do evento	Int	Dias	Decomposição da coluna de data
21	mes	Mês do evento	Int	Meses	Decomposição da coluna de data
22	ano	Ano do evento	Int	Anos	Decomposição da coluna de data

APÊNDICE F – LINK PARA REPOSITÓRIO DE CÓDIGOS

Todo o desenvolvimento de código foi versionado pela plataforma GitHub. Abaixo está o link do repositório com os códigos utilizados para este trabalho.

https://github.com/loobato/projeto

APÊNDICE G - APRESENTAÇÃO PARA A BANCA

Abaixo segue o link para a apresentação realizada para a banca examinadora no dia 04/12/2025. O link segue para uma apresentação do Google Slides, e em casos de problemas de acesso, entre em contato com o autor.

https://docs.google.com/presentation/d/1J-7kfMrXyIL6dWXLIZLUjqubgerWqEVxrBRejlMSfHU/edit?usp=sharing