PANDORA: Sistema Estatístico de Previsão de Eventos Climáticos Extremos

Luís Fernando Cezar dos Santos, Flavia Maristela S. Nascimento

[1] Departamento Acadêmico de Computação – Instituto Federal da Bahia (IFBA)

20212160053@ifba.edu.br, flaviamsn@ifba.edu.br

Abstract. *This paper highlights the importance of extreme weather forecasting systems and presents* `Pandora`*, a forecasting tool for extreme weather events.* `Pandora` *uses the Emergency Events Database (EM-DAT), focusing on the categories of extreme weather events relevant to the Brazilian territory. Preliminary results show that* `Pandora` *is a valid tool for forecasting and correlating data from extreme events in Brazil.*

Resumo. *Este trabalho ressalta a importância dos sistemas de previsão de eventos extremos e apresenta* `Pandora`*, uma ferramenta de previsão de eventos climátios extremos.* `Pandora` *utiliza a base de dados do Emergency Events Database(EM-DAT), com foco nas categorias de eventos extremos relevantes para o território brasileiro. Resultados preliminares mostram que* `Pandora` *é uma ferramenta válida para previsão e correlação de dados de eventos extremos no Brasil.*

1. Introdução

Os eventos climáticos, como tempestades, inundações e ondas de calor, têm sido objeto de crescente preocupação devido à relevância seu impacto socioeconômico e ambiental [Intergovernmental Panel on Climate Change (IPCC) 2021]. Esses eventos, também conhecidos como eventos extremos, são caracterizados por sua ocorrência fora das variações climáticas normais em termos de intensidade, duração, frequência ou distribuição espacial. Vários estudos científicos têm investigado a relação entre mudanças climáticas e a ocorrência de eventos extremos, amparado na observada tendência crescente de ocorrência destes eventos em muitas partes do mundo, inclusive em regiões onde tipicamente não ocorriam com frequência, como o Brasil. De fato, modelos matemáticos climáticos indicam que as mudanças climáticas antropogênicas têm contribuído significativamente para o aumento da probabilidade de eventos extremos em várias regiões do mundo [Hansen et al. 2012].

Este trabalho apresenta `Pandora`, uma aplicação inovadora, que apresenta uma visualização probabilística de eventos extremos em potencial, baseada em um processo de Extração, Transformação e Carregamento de Dados (ETL) [Almeida et al. 2024], além de proporcionar acesso a dados históricos relacionados à ocorrência de destes eventos. Os dados históricos acessados pela aplicação são armazenados em uma base de dados não relacional, com o objetivo de prover escalabilidade e prover melhor desempenho nas consultas, por se tratar de um volume de dados com alta granularidade [Soares and Matos 2017]. O objetivo principal do ferramenta é fornecer aos usuários uma maneira intuitiva para antever os eventos extremos, capacitando-os a tomar decisões proativas diante de situações críticas.

O restante deste documento está estruturado da seguinte forma. A Seção 2 apresenta os trabalhos relacionados destacando as principais características de cada projeto. As Seções 3 e

4 descrevem, respectivamente, as fontes de dados utilizadas e as características do cálculo de correlação quando comparado ao aprendizado de máquina, justificando a escolha deste projeto. Na sequência, a Seção 5 apresenta a ferramenta `Pandora`, sua visão arquitetural e características. A Seção 6 apresenta os resultados preliminares e a Seção 7 apresenta a conclusão e próximos passos.

2. Trabalhos relacionados

As plataformas Meteored [Meteored 2024] e Climatempo [Climatempo 2024] são hoje as duas principais fontes de informações meteorológicas mais gerais (previsão de chuvas, umidade relativa do ar, temperaturas e condições de vento). Apesar de tratar os mesmos tipos de dados, estas plataformas apresentam entre si diferenças significativas em termos de abordagem, recursos e foco.

A plataforma Meteored se destaca por oferecer uma abrangência global, fornecendo previsões detalhadas e análises climáticas para uma ampla gama de localidades em todo o mundo. Além disso, conta com recursos extras como mapas interativos, gráficos e aplicativos móveis, que auxiliam na visualização e compreensão das condições climáticas.

A plataforma Climatempo se concentra nas informações climáticas do território brasileiro, disponibilizando análises climáticas detalhadas e previsões específicas para as diversas regiões do país, embora também forneça informações de outras partes do mundo. A Climatempo disponibiliza ainda recursos como notícias sobre meteorologia, vídeos explicativos e conteúdo educativo sobre o clima, visando oferecer informações mais abrangentes e contextualizadas.

Essas diferenças refletem as distintas abordagens e foco das plataformas Meteored e Climatempo, permitindo que os usuários escolham aquela que melhor atenda às suas necessidades e preferências específicas. Apesar da incontestável importância destas aplicações, nenhuma delas foca na previsão e análise de dados climáticos relativos à eventos extremos. A Tabela 1, apresenta um resumo comparativo entre as principais características das plataformas apresentadas nesta seção com a proposta apresentada neste trabalho.

Tabela 1. Recursos disponíveis

***	PANDORA	METEORED	CLIMATEMPO
PREVISÃO	X	X	X
CROSS-PLATAFORM	X	X	-
EVENTOS EXTREMOS	X	-	-
SISTEMA DE ALERTA	-	X	-
RADAR	X	X	X
NOTICIAS	X	X	X

3. Fonte de Dados

A fonte de dados é a base fundamental para qualquer processo de ETL, pois é a partir dela que se originam todas as informações que serão utilizadas. De fato, a qualidade, integridade e consistência dos dados na origem da aplicação afetam diretamente a confiabilidade e o valor do resultado final [Diouf et al. 2018, Almeida et al. 2024]. `Pandora` utiliza duas fontes de

dados principais: EM-DAT e Open-Meteo, que por suas características garantem precisão, confiabilidade e qualidade para as funcionalidades da ferramenta.

A base de dados **EM-DAT (The International Disaster Database)** é utilizada para racionalizar a preparação para catástrofes e a tomada de decisões, ao mesmo tempo que fornece uma base objetiva para a avaliação de vulnerabilidade e riscos. EM-DAT contém registros de catástrofes em massa, bem como os seus impactos na saúde e na economia a nível nacional. Além disso, contém dados essenciais sobre a ocorrência e os efeitos de 26.000 desastres em todo o mundo, desde 1900 até hoje. A base de dados é compilada a partir de diversas fontes de informação, incluindo agências da ONU, organizações não governamentais, companhias de seguros, institutos de investigação e agências de imprensa [CRED 2024].

Open-Meteo é uma API *open-source* que provê informações meteorológicas precisas para qualquer localidade. Particularmente, dois serviços desta API são importantes: (a) *Historical Weather API*, que permite acessar informações meteorológicas históricas e *Weather Forecast API*, relacionada a dados de previsão do tempo.

A API *Historical Weather* se baseia em conjuntos de dados de reanálise e usa dados oriundos de diferentes fontes de captação, como estação meteorológica, aeronaves, bóias, radares e satélites, para criar um registro abrangente das condições climáticas passadas. Esses conjuntos de dados são capazes de preencher lacunas usando modelos matemáticos e estatísticos para estimar os valores de várias variáveis meteorológicas [Zippenfenig 2023]. Já a API *Weather Forecast* utiliza modelos meteorológicos de vários provedores de uma mesma localidade. A ideia é que para cada localidade, os melhores modelos sejam combinados para fornecer a melhor previsão possível[Zippenfenig 2023].

4. Cálculo de correlação e Aprendizado de Máquina

O cálculo de correlação entre variáveis é uma técnica estatística fundamental que permite avaliar a relação linear entre duas ou mais variáveis [Montgomery and Runger 2021]. Essa técnica oferece algumas vantagens quando comparada ao uso de técnicas de aprendizado de máquina para previsão de dados, dependendo do contexto e dos objetivos do projeto. Por exemplo, a **interpretabilidade** fornece uma medida direta da relação linear entre as variáveis, o que facilita a interpretação dos resultados.

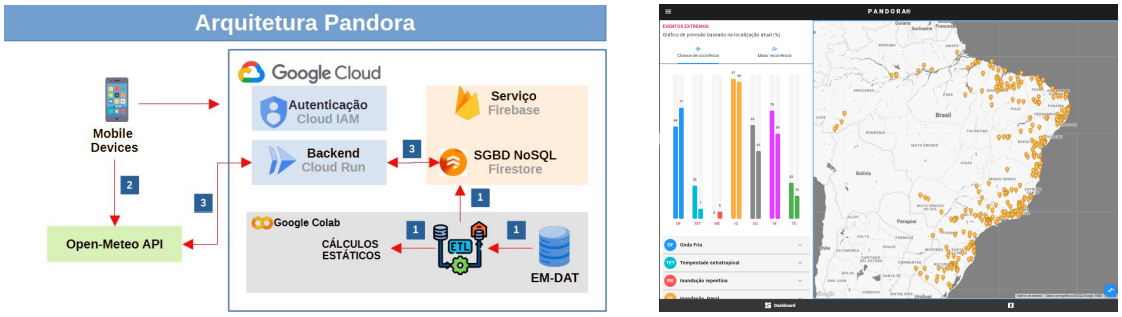
Outra característica do cálculo de correlação é a **simplicidade** que é simples e rápida de ser aplicada, especialmente quando se trata de um número pequeno a moderado de variáveis. Por fim, o cálculo de correlação também permite **identificar relações lineares**, que possibilita usar uma variável para prever outra de forma mais direta, , sem a necessidade de métodos mais complexos, caso haja uma forte correlação entre elas.

É importante ressaltar que o cálculo de correlação tem suas limitações e pode não capturar relações não lineares entre as variáveis. Além disso, em cenários onde existem muitas variáveis ou onde as relações são complexas e não lineares, técnicas de aprendizado de máquina podem ser mais adequadas para construir modelos de previsão precisos. Considerando a quantidade de variáveis relacionadas à previsão de eventos extremos, a opção neste trabalho foi pelo cálculo de correlação.

5. Arquitetura e Funcionamento do Sistema PANDORA

O sistema PANDORA será uma solução integrada para previsão e análise de eventos climáticos extremos, disponibilizada por meio de aplicações web e mobile. A experiência do usuário final será simples e intuitiva, oferecendo visualizações claras do risco probabilístico de desastres naturais para uma localidade específica.

A arquitetura do PANDORA está fundamentada no ecossistema da Google Cloud, garantindo escalabilidade, segurança e alta disponibilidade. Os cálculos complexos de correlação estatística entre dados históricos serão pré-processados e armazenados no Firebase Firestore, um banco NoSQL que permite consultas rápidas e estruturadas por documento. O backend, implementado em Python e hospedado no Cloud Run, será responsável por realizar o cálculo percentual de similaridade entre as correlações pré-armazenadas e os dados climáticos coletados em tempo real.



(a) Pandora: Arquitetura (b) Pandora: Painel principal

Figura 1. Arquitetura e interface principal do sistema PANDORA

A autenticação dos usuários será gerenciada pelo Firebase Authentication, garantindo controle seguro de acesso aos recursos da aplicação e protegendo as informações sensíveis.

Assim, o usuário poderá, a qualquer momento, consultar via web ou aplicativo móvel a probabilidade atualizada de ocorrência de eventos extremos em sua região, baseada em análises estatísticas robustas e dados históricos consolidados. A infraestrutura baseada em Cloud Run e Firebase assegura que o sistema seja escalável e responsivo, mesmo com aumento no volume de usuários e dados, proporcionando uma ferramenta confiável para suporte à tomada de decisão em situações críticas relacionadas ao clima [Google 2024a, Google 2024b, Intergovernmental Panel on Climate Change (IPCC) 2021].

5.1. Fluxo de execução detalhado

O projeto PANDORA emprega técnicas de ETL essenciais para o processamento eficiente de dados climáticos e de desastres naturais. Inicialmente, realiza-se a extração (Extract) dos dados históricos da base EM-DAT, armazenados em arquivos Excel, que contêm informações heterogêneas e frequentemente incompletas. Em seguida, ocorre a transformação (Transform), que envolve o tratamento dos dados — como preenchimento de valores ausentes, padronização de tipos, geocodificação para obtenção de coordenadas geográficas via API OpenCage, e o enriquecimento com dados meteorológicos históricos provenientes da API Open-Meteo. Esta etapa também contempla o cálculo de correlações estatísticas entre variáveis climáticas segmentadas

por tipo de desastre. Finalmente, o carregamento (Load) consiste na persistência dos dados tratados e das correlações pré-calculadas em um banco de dados NoSQL, especificamente o Firebase Firestore, organizado em coleções e documentos que garantem escalabilidade e flexibilidade, como exemplicado na Figura 2.

Dentre os principais desafios enfrentados destacam-se o tratamento de dados heterogêneos e incompletos, a obtenção precisa das coordenadas geográficas para múltiplas localizações, a gestão do grande volume de dados meteorológicos diários e a necessidade de realizar cálculos estatísticos complexos em grandes conjuntos de dados. Tais desafios foram superados por meio de técnicas como preenchimento e padronização dos dados, filtragem rigorosa das informações geográficas, implementação de cache e estratégias de retry para otimizar as chamadas às APIs externas, além do uso eficiente das bibliotecas Pandas e Numpy para o processamento estatístico.

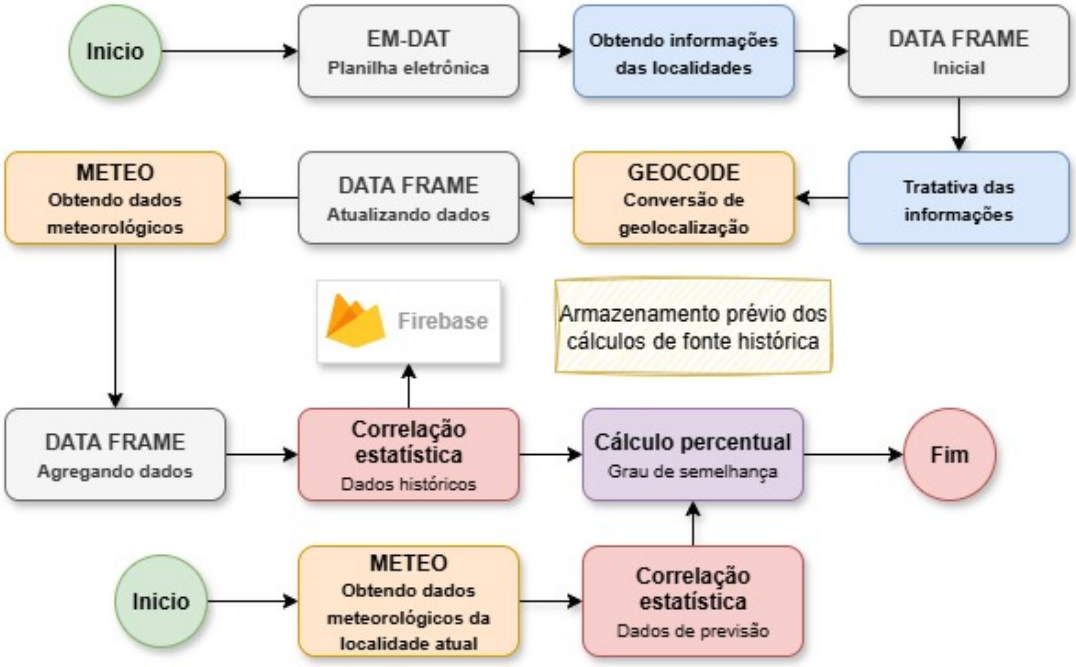


Figura 2. Pandora Analytics: Fluxo de execução

O modelo de armazenamento adotado, baseado em documentos no Firebase Firestore, oferece uma estrutura hierárquica que facilita a expansão do sistema para diferentes países e eventos, além de permitir o armazenamento eficiente de correlações pré-calculadas, reduzindo a necessidade de processamento em tempo real. Essa arquitetura possibilita a execução de consultas que recuperam dados históricos, calculam similaridades percentuais entre correlações históricas e atuais e suportam visualizações probabilísticas de riscos climáticos.

Entretanto, a plataforma enfrenta limitações relacionadas ao volume de dados: embora o Firebase Firestore seja eficiente para consultas rápidas em documentos específicos, não é otimizado para processamentos estatísticos complexos ou análise em tempo real diretamente no banco. Assim, a estratégia adotada prioriza o pré-processamento e armazenamento das correlações, garantindo agilidade na entrega dos resultados aos usuários. Dessa forma, o PANDORA equilibra a complexidade do processamento estatístico com a flexibilidade e escalabilidade do modelo NoSQL, assegurando desempenho satisfatório na previsão de eventos ex-

tremos [Google 2024a, Diouf et al. 2018, pandas development team 2020, Zippenfenig 2023, CRED 2024, Almeida et al. 2024, Soares and Matos 2017].

6. Resultados preliminares

Os resultados preliminares do projeto PANDORA demonstram a aplicação da estatística para identificar padrões climáticos associados a desastres naturais. A Figura 3 apresenta a comparação entre os coeficientes de correlação estatística extraídos de dados históricos e os dados climáticos atuais da localidade do usuário.

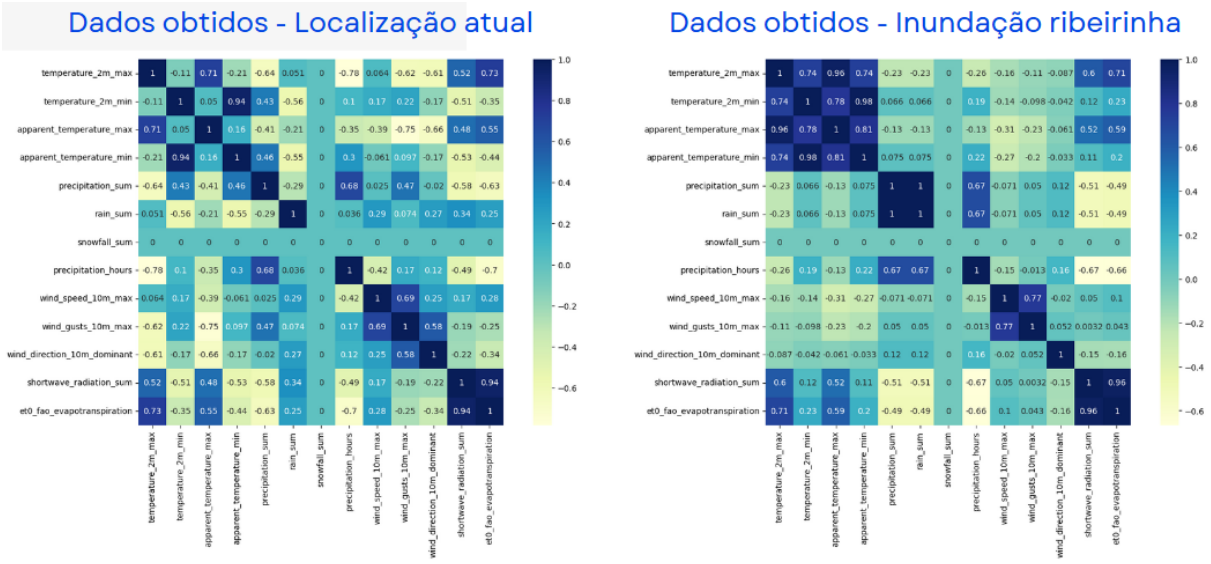


Figura 3. Grau de semelhança: Comparação de correlação estatística

Para cada evento registrado na base EM-DAT (como inundações, secas e ciclones), o sistema coleta os dados climáticos históricos da região e período correspondente. Com essas informações, é gerada uma matriz de correlação entre variáveis como temperatura, precipitação, pressão e umidade. Em seguida, os dados climáticos atuais da localidade informada pelo usuário são coletados via API Open-Meteo. Uma nova matriz de correlação é gerada com base nessas condições recentes. O sistema então compara essa matriz com os padrões históricos, utilizando uma métrica de similaridade percentual. O resultado indica o risco atual para cada tipo de evento. Por exemplo: 80% de similaridade com padrões de inundação ribeirinha, 10% com seca e 10% com outros eventos.

Essa abordagem vai além das previsões tradicionais, que apenas informam probabilidade de chuva ou temperatura. O PANDORA transforma dados históricos e atuais em um índice de risco real, contextualizado e baseado em estatística.

7. Conclusão

O projeto PANDORA destaca-se como uma solução tecnológica relevante frente ao aumento dos desastres naturais associados às mudanças climáticas. Ao integrar múltiplas fontes de dados e aplicar técnicas robustas de ETL, a plataforma possibilita análises preditivas eficazes, apoiando ações de prevenção e resposta. Sua arquitetura escalável, baseada em pré-processamento estatístico e armazenamento NoSQL, garante desempenho e acessibilidade. Com isso, o PANDORA reafirma o papel essencial da ciência de dados e da engenharia de software no enfrentamento de desafios ambientais contemporâneos.

Referências

Almeida, B., Frota, Y., and de Oliveira, D. (2024). Otimização de parâmetros em aplicações de big data baseadas em múltiplos frameworks. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 418–430, Porto Alegre, RS, Brasil. SBC.

Climatempo (2024). Previsão do tempo para o rio de janeiro. `https://www.climatempo.com.br/previsao-do-tempo-para-rio-de-janeiro`. Acesso em: 27 maio 2024.

CRED, p. o. t. U. o. L. U. (2024). Em-dat documentation.

Diouf, P. S., Boly, A., and Ndiaye, S. (2018). Variety of data in the etl processes in the cloud: State of the art. In *2018 IEEE International Conference on Innovative Research and Development (ICIRD)*, pages 1–5.

Google (2024a). Firebase realtime database: Managed nosql database service. `https://firebase.google.com/products/realtime-database`. É um serviço de banco de dados NoSQL totalmente gerenciado oferecido pela plataforma Firebase do Google. Ele fornece uma estrutura flexível e escalável para armazenar, sincronizar e consultar dados para aplicativos da web, móveis e em tempo real.

Google (2024b). Google cloud run: Managed container services. `https://cloud.google.com/run`. Serviço de computação em contêiner totalmente gerenciado oferecido pela Google Cloud Platform. Ele permite aos desenvolvedores implantar e executar facilmente aplicativos em contêineres sem se preocupar com a infraestrutura subjacente. O Cloud Run suporta contêineres do Docker e oferece integração perfeita com outras ferramentas e serviços do Google Cloud.

Hansen, J., Sato, M., and Ruedy, R. (2012). Perception of climate change. *Proceedings of the National Academy of Sciences*, 109(37):E2415–E2423.

Intergovernmental Panel on Climate Change (IPCC) (2021). Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.

Meteored (2024). Previsão do tempo para são paulo. `https://www.meteored.com.br/previsao-do-tempo-para-sao-paulo`. Acesso em: 27 maio 2024.

Montgomery, D. C. and Runger, G. C. (2021). *Applied Statistics and Probability for Engineers*. Wiley, Hoboken, NJ, 7th edition. O cálculo de correlação entre variáveis é uma técnica estatística fundamental que permite avaliar a relação linear entre duas ou mais variáveis.

pandas development team, T. (2020). pandas-dev/pandas: Pandas.

Soares, A. and Matos, P. (2017). Uma análise comparativa entre sistemas gerenciadores de bancos de dados nosql no contexto de internet das coisas. In *Anais do XXXII Simpósio Brasileiro de Bancos de Dados*, pages 306–311, Porto Alegre, RS, Brasil. SBC.

Zippenfenig, P. (2023). Open-meteo.com weather api.