Algoritmos de Ensemble Learning para Análise Preditiva acerca de Desastres Naturais

Melissa Frigi Mendes
Bacharelado em Ciência e Tecnologia
Universidade Federal de São Paulo
São José dos Campos, Brasil
melissa.frigi@unifesp.br

***Resumo* — O objetivo deste trabalho foi utilizar de algoritmos de ensemble learning para tarefas preditivas e análise do desempenho dos algoritmos com relação ao número de mortes relacionado a desastres naturais utilizando a base 'The International Disaster Database (EM-DAT). Foram realizadas tratativas de pré-processamento e mineração da base de dados, extração de informações com gráficos estatísticos e aplicação e avaliação dos algoritmos.**

Palavras-chave — Ensemble learning, tarefas preditivas, mineração de dados.

I. Introdução e Motivação

O Brasil é um país que se encontra cada vez mais suscetível a diversos tipos de desastres naturais, incluindo enchentes, deslizamentos de terra, secas, incêndios florestais, entre outros. Nos últimos anos, o país tem enfrentado uma série de catástrofes ambientais que resultaram em graves prejuízos e impactos significativos tanto para a população quanto para o ecossistema. Pose-se observar que as enchentes e deslizamentos, frequentemente associados a períodos de chuvas intensas, causam destruição de moradias, infraestruturas e até perda de vidas, enquanto as secas prolongadas afetam a agricultura e o abastecimento de água da população, além de acentuar problemas como a desigualdade social.

Um exemplo emblemático desse cenário foi a catastrófica seca ocorrida no Pantanal, que ocasionou múltiplos focos de enormes incêndios florestais entre 2019 e 2020, provocando a destruição de milhares de hectares de vegetação e afetando severamente o bioma [Embrapa, 2022]. Além disso, a fumaça resultante dos incêndios afetou a saúde de comunidades locais e contribuiu para o aumento das emissões de gases de efeito estufa. Tal ocorrência é apenas um dos vários incidentes já registrados no Brasil acerca de desastres naturais que evidenciam a grande importância da prevenção e do planejamento adequado para minimizar as catastróficas consequências dos desastres naturais no país.

Diante de tais desafios, as metodologias pertencentes ao campo da ciência de dados têm se mostrado cada vez mais eficazes. Estudos de validação de performance dos métodos, como o realizado por Pedreira et al. (2022), demonstram que a coleta e tratamento de dados, aliados a métodos de probabilidade e algoritmos de machine learning adequados, podem fornecer resultados satisfatórios para a previsão de desastres naturais.

II. Conceitos Fundamentais

A seguir, será apresentado de forma resumida alguns conceitos fundamentais necessários para compreender os tópicos discutidos neste relatório.

1) Ciência de Dados

Denomina-se como ciência de dados o campo interdisciplinar que, através da coleta, limpeza e organização de um grande volume de dados brutos e complexos, emprega técnicas estatísticas, matemáticas e computacionais, como algoritmos de aprendizado de máquina, para analisar e interpretar as informações contidas nos conjuntos de dados. Tais processos permitem a identificação de tendências, previsões, relacionamentos e insights que podem ser utilizados para otimizar processos, melhorar a tomada de decisões, desenvolver estratégias e realizar previsões eficazes.

2) Pré-processamento de Dados

Na primeira etapa do estudo, também chamada de fase de caracterização de dados, é realizado o processo de caracterização, exploração e pré-processamento do conjunto de dados. Os tipos de atributos são definidos como qualitativo/categórico (variáveis nominais e ordinais), e quantitativos/numéricos (variáveis intervalares e racionais).

3) Applicações de Estatística

Engloba a exploração preliminar e transformação dos dados univariados ou multivariados, que auxilia na compreensão das características e padrões do conjunto de dados ao fazer uso de estatísticas descritivas e ferramentas de visualização.

Com relação a técnicas de estatísticas descritivas, as quais permitem encontrar informações como frequência dos dados, tendência central, dispersão e distribuição, são utilizados conceitos como de média, moda, mediana, quartis e percentis, histogramas e gráficos boxplot, assim como medidas de dispersão como intervalo, variância, obliquidade, curtose e desvio padrão. Há ainda técnicas para dados multivariados (vários atributos), como covariância de dois atributos e correlação de Pearson.

4) Machine Learning

Após os dados serem devidamente categorizados e tratados, é dado início à fase da aplicação de modelos de Machine Learning, que é uma subárea da Inteligência Artificial cujo objetivo é automatizar a aprendizagem de algoritmos nas máquinas de computador, utilizando-se dos dados tratados para tarefas diferentes como classificação, associação e previsão, sendo este o foco deste trabalho. Portanto, é importante descrever dois paradigmas de aprendizado de máquina: supervisionado e não supervisionado.

O aprendizado supervisionado refere-se a um tipo em que um modelo é treinado usando dados previamente rotulados

por humanos. Esses dados são utilizados para realizar tarefas específicas, como classificação, que envolve categorizar um conjunto de dados em classes distintas, ou regressão, que se concentra em fazer previsões de valores numéricos contínuos. Já no aprendizado não supervisionado não há tais dados rotulados por humanos, e o trabalho dos algoritmos envolve a busca por padrões ocultos que agrupam um conjunto de informações, e o modelo envolve técnicas para agrupamento, associação e sumarização (Faceli et. al, 2021).

A partir do conceito de árvores em estrutura de dados, o modelo de aprendizado supervisionado de árvores de decisão e regressão utilizam o método dividir e conquistar para solucionar recursivamente problemas de classificação e regressão, no qual cada subconjunto de dados (nós da árvore) fornecerá alguma informação significativa com relação às variáveis de entrada, e uma vez construída, poderá classificar novos exemplos ou prever valores de regressão.

5) *Algoritmos de Ensemble Learning*

Algoritmos de ensemble learning (métodos de agupamento) são técnicas de aprendizado de máquina que combinam vários modelos diferentes ditos "mais fracos" para aumentar o desempenho, precisão e robustez nas previsões comparado ao desempenho único desses algoritmos em si. Existem alguns tipos de métodos de agrupamento, mas o mais utilizados são *Bagging*, *Boosting* e *Stacking*.

O método Bagging utiliza da técninca de *Bootstrapping*, que consiste em subdividir a base de dados em conjuntos menores, com elementos aleatórios e com reposição, e constrói um preditor para cada subconjunto. A previsão final combina os preditores gerados anteriormente.

O Boosting também segue a premissa de produzir um modelo mais forte a partir de modelos "mais fracos", mas diferentemente do algoritmo anterior, os modelos são treinados de forma sequencial onde cada novo modelo é gerado a fim de corrigir erros do modelo anterior. Ao final, os modelos são combinados para compor a solução final.

O Stacking tem como princípio combinar previsões de algoritmos base utilizando um algoritmo chamado de "meta", que irá retornar a previsão final.

Neste projeto, serão utilizados o 3 tipos de algoritmos, descritos a seguir.

6) *Algoritmo Random Forest*

Random Forest (floresta aleatória) é um algoritmo de aprendizado de máquina do tipo *Bagging* que utiliza várias árvores de decisão através do conceito de ensemble learning para realizar tarefas de classificação e regressão, combinando múltiplas árvores de decisão individuais para obter um modelo final mais robusto e preciso.

É um modelo capaz de lidar com uma carga de dados muito alta, é resistente ao sobreajuste (overfitting) dos dados e possui alta precisão.

7) *Algoritmo XGBoost*

O XGBoost (*Extreme Gradient Boosting*) é um algoritmo do tipo *Boosting* que também se baseia na estrutura de árvores de decisão, se destacando por sua alta performance e capacidade de lidar com um alto volume de dados. Pode ser descrita como um aperfeiçoamento do algoritmo *Gradient Boosting*, um algoritmo do tipo Boosting que otimiza os modelos sequencialmente para diminuir erros prévios.

Da mesma forma, a cada nova ávore gerada pelo XGBoosting, o erro pode ser reduzido através das iterações da árvore anterior gerada. Por fim, o modelo gera o resultado final com base na combinação ponderada das árvores.

8) *Algoritmo AdaBoost*

O AdaBoosting (Adaptive Boosting) representa mais um algoritmo da categoria *Boosting*, o qual combina modelos de machine learning mais simples e de baixa complexidade treinados sequencialmente para criar um modelo final mais robusto. Durante o processo de iteração do algoritmo, os pesos dos dados de treinamento e do desempenho de cada modelo são atualizados, e o resultado final consiste em uma combinação dos modelos testados com base no peso atribuído em cada um.

9) *Algoritmo de Stacking*

O algoritmo *Stacking* utiliza uma técnica hieráquica que combina as previsões dos modelos testados através de um "meta-modelo", ao contrário dos modelos de técnica de *Bagging e Boosting*, o quais realizam os treinos de vários modelos de forma independente.

Após o treinamento dos modelos bases, é formado então um novo conjunto de dados contendo os resultados das previsões obtidos, e então este é utlizado como teste e o conjunto do alvo continua sendo o original, resultando na previsão final do algoritmo.

III. TRABALHOS RELACIONADOS

Serão apresentados, a seguir, alguns trabalhos que já propuseram soluções para predição de desastres, assim como análises das performances dos algoritmos de aprendizado de máquina.

No estudo em [Pedreira et al. 2022], foi desenvolvido uma base de dados com integração e pré processamento aplicados, e quatro modelos de machine learning para serem estudados: a Árvore de Decisão (AD), um algoritmo de aprendizado de máquina supervisionado que cria um modelo em forma de árvore para prever a variável-alvo com base em várias variáveis de entrada [Breiman et al. 1984]; o Random Forest (RF), um método de aprendizado de máquina que utiliza a técnica de ensemble learning (aprendizado em conjunto) para realizar tarefas de classificação e regressão, combinando múltiplas árvores de decisão individuais para obter uma predição fina [Breiman 2001]; Redes Neurais Artificiais (RNA) do tipo Multilayer Perceptron (MLP), um algoritmo de aprendizado de máquina que simula o funcionamento do cérebro humano por meio de uma rede neural artificial com cada camada da rede conectada à camada seguinte por meio de pesos sinápticos, utilizando-se de uma técnica de treinamento supervisionado chamada back-propagation para ajustar os pesos sinápticos e minimizar o erro de predição [Haykin 2009]; por fim, o Light Gradient Boosting Machine (LGBM), um algoritmo de aprendizado de máquina de boosting que utiliza múltiplas árvores de decisão para realizar previsões precisas, é amplamente utilizado em várias aplicações de aprendizado de máquina, incluindo

classificação, regressão e ranking de dados [Ke et al. 2017]. Após as medidas de eficácias escolhidas serem aplicadas, o estudo apontou o LGBM e o RF como sendo os melhores algoritmos de aprendizado de máquina para predições, com o RF tendo as melhores avaliações na maioria de todos os casos de avaliação de performance.

Em [Lee et. Al 2017], o objetivo do estudo é a criação de mapas de suscetibilidade a inundações para a cidade metropolitana de Seul, Coréia do Sul, usando um modelo de floresta aleatória e um modelo de árvore impulsionada, a fim de se comparar ambos os modelos. Primeiramente, os fatores hidrológicos foram coletados de mapas topográficos, de solo, de uso de terras e geológicos, e o fatores relacionados como um modelo de elevação digital (DEM) foram extraídos, e dados como declives, curvatura de planos e erosão média foram devidamente calculados. Tais informações foram extraídos para servirem como dados de treinamento e validação e, em seguida, tais conjuntos de dados foram convertidos para um formato de grade com interpolação bilinear utilizando-se do modelo de ArcGIS. Na etapa de aplicação dos modelos de aprendizado de máquina com o programa STATISTICA (desenvolvida por StatSoft), as variáveis foram categorizadas em variáveis dependente e independentes, e realizou-se o cálculo da importância preditora cada fator, assim os resultados foram comparados utilizando-se as áreas alagadas conhecidas. Os resultados por comparação em gráficos retornaram uma acurácia de 78,78% para o modelo de regressão Random Forest e uma acurácia de 77,26% para o modelo de classificação Árvore Impulsionada.

Diante das técnicas utilizadas e nos algoritmos escolhidos para cada trabalho, é possível notar que algoritmos de regressão como Random Forest e Redes Neurais demonstraram retornar uma acurácia de resultados de previsão maiores, sendo de grande interesse de estudo.

IV. OBJETIVOS

O objetivo deste trabalho envolve aplicar e analizar o desempenho da tarefa de regressão dos algoritmos selecionados Random Forest, XGBoost, Ada Boost e Stacking com relação à previsão da variável alvo escolhida, o número de óbitos relacionado aos desastres do tipo naturais utilizando a base de dados extraída do repositório público EM - DAT: The International Disaster Database.

V. METODOLOGIA EXPERIMENTAL

Uma série de metodologias foram aplicadas para o correto tratamento da base de dados, para que os algoritmos pudessem ser aplicados e analisados posteriormente.

1) Pré-processamento dos Dados

A base de dados foi extraída a partir do site official da EM – DAT, e foi feito seu carregamento na ferramenta de desenvolvimento em formato Jupyter Notebook, Google Colab, utilizando as bilbiotecas ‘pandas’ e ‘os’ os quais permitiu a importação para a ferramenta através do Google Drive e a modificação de seu formato para ‘.csv’.

Inicialmente, a base possuía 46 colunas (atributos) e 26.631 linhas (entradas), com dados do tipo numérico (inteiros e racionais) e tipo categórico (strings). Algumas colunas possuíam também bastante dados não preenchido/vazios e muitas destas colunas não agregavam valor nem sentido para o algoritmos, como códigos de identificação genéricos e colunas apenas com siglas de uma outra coluna com nomes completos.

Com base nessas análises, a limpeza da base de dados começou pela remoção dos tipos de desastres que não são alvo deste projeto, como os desastres tecnológicos e biológicos, resultando em uma redução de linhas para 15.518.

Em seguida, as colunas que não agregam valor, ou que envolvem métodos que não são abordados neste projeto, como análises de series temporais para tratamento periódico das datas também foram excluídas (colunas ‘End Year’, ‘End Month’ e ‘End Day’), assim como colunas com mais de 60% de dados faltantes. As linhas com menos de 10% de dados faltantes e as correlacionadas com ‘Total Deaths’ também foram removidas, assim como as linhas faltantes de ‘Total Deaths’ pois é a variável alvo, resultando em uma quantidade de linhas de 5032 e 19 colunas.

2) Exploração da Base de Dados

Foram gerados alguns gráficos que permitiram a visualização estatística e entendimento do comportamento da base de dados, a começar com o gráfico boxplot da distribuição geral de mortes.

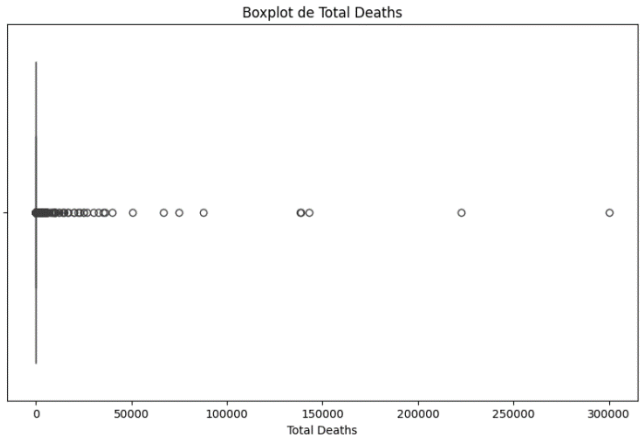

Figura 1 - Boxplot geral da variável 'Total Deaths'

Nota-se uma distribuição com muitos outliers, o que indica que a variável alvo possui comportamento desordenado.

A fim de entender a distribuição com um tipo de desastre específico, foi gerado um gráfico boxplot da variável alvo com relação ao desastre do tipo ‘Flood’ (enchentes).

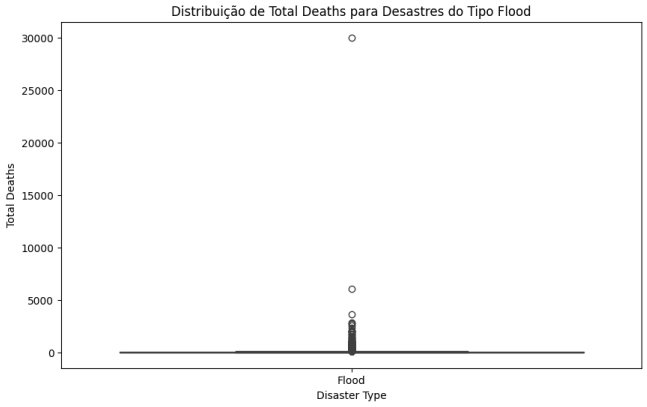

Figura 2 - Boxplot de 'Total Deaths' com relação ao desastre do tipo 'Flood'

Mesmo especificando para o desastre tipo ‘Flood’, a variável alvo ainda possui um número alto de outliers.

Ainda sobre o comportamento estatístico da variável, foram gerados gráficos boxplots para os subcontinents ‘Southern Asia’, ‘Eastern Asia’, ‘Central Asia’ e ‘Northern Europe’ como exemplos aleatórios.

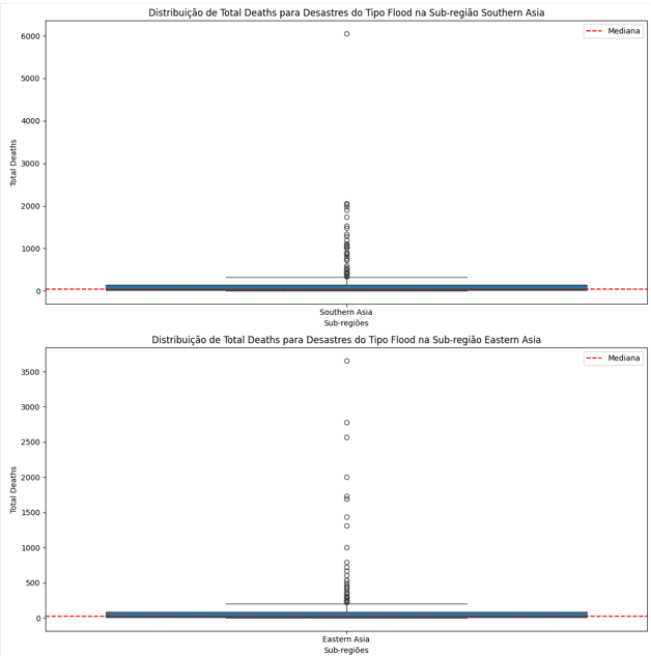

Figura 3 - Boxplots gerados para os subcontinentes 'Southern Asia' e 'Eastern Asia'

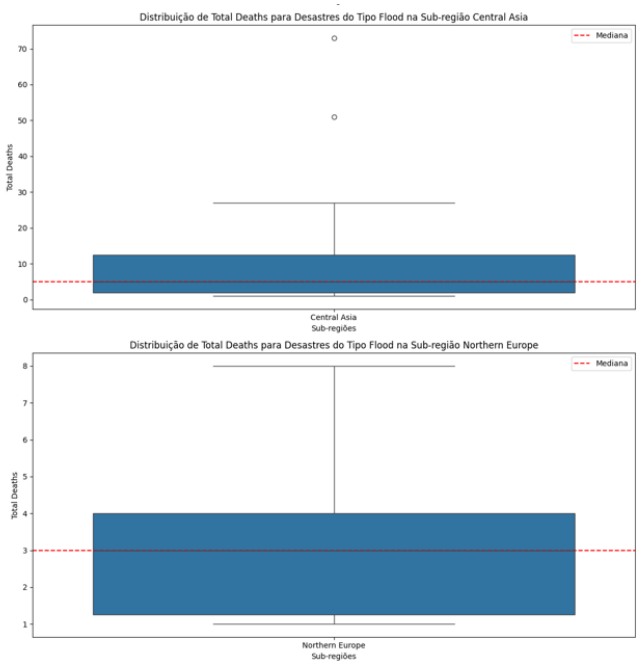

Figura 4 - Boxplots gerados para os subcontinentes 'Southern Asia' e 'Eastern Asia'4

Em seguida, foi gerado um gráfico histograma com o total de mortes por continente mas não foi possível extrair nenhuma informação relevante sobre isso pois ficou muito generalizado. Portanto, uma análise mais detalhada foi feita, analisando as mortes por subcontinentes e subtipos de desastres naturais. Alguns exemplos gerados mostram o padrão do comportamento variar bastante com relação a cada tipo de desatre natural.

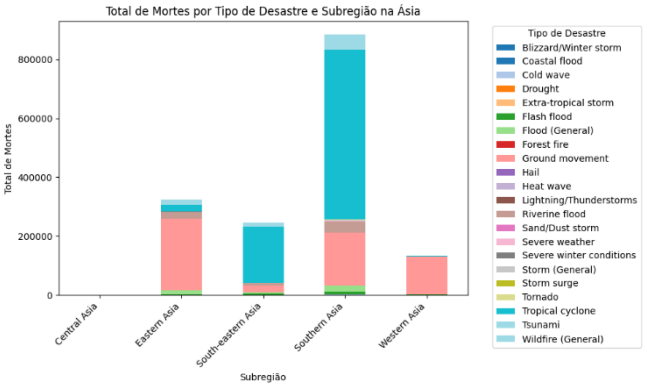

Figura 5 - Histograma de mortes por tipo de desastre natural para os subcontinentes da Ásia

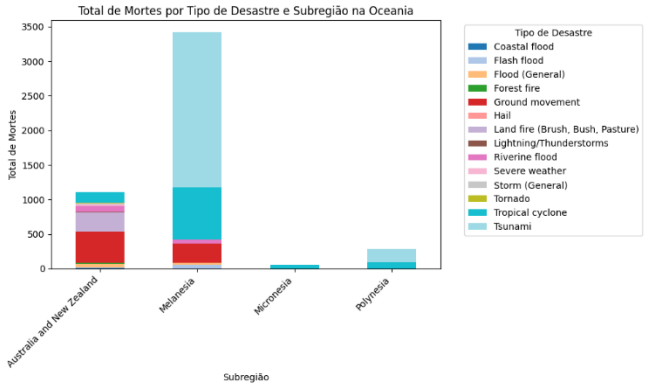

Figura 6 - Histograma de mortes por tipo de desastre natural para os subcontinentes da Oceania

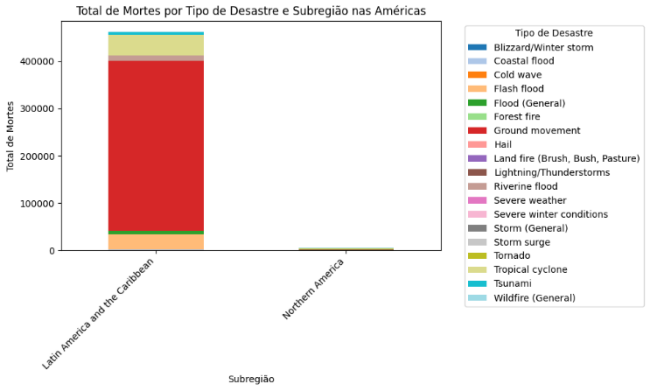

Figura 7 - Histograma de mortes por tipo de desastre natural para os subcontinentes das Américas

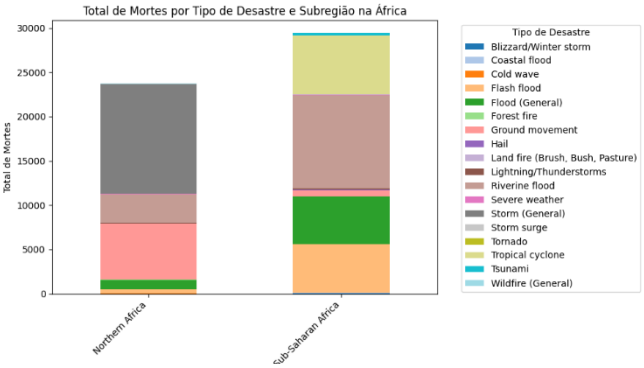

Figura 8 - - Histograma de mortes por tipo de desastre natural para os subcontinentes da África

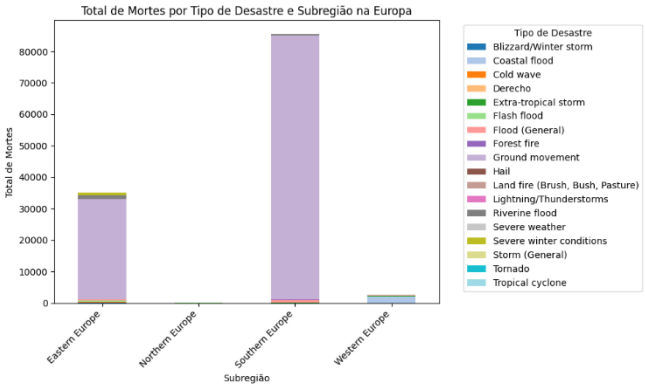

Figura 9 – Total de mortes por subtipo de desastre natural na Europa

Pode-se observer que o padrão de mortes varia muito conforme o tipo de desastre e a localidade, o que se deve também por fatores geográficos. Porém, sem a informação de como as mortes se comportam ao longo dos anos, então foram gerados gráficos do tipo 'Scatter Plot' que leva esse fator em consideração para alguns subcontinentes.

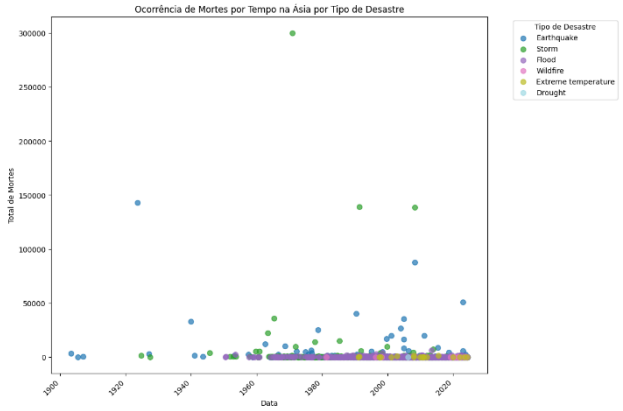

Figura 10 – Scatter plot da ocorrência de mortes por subtipo de desastre natural na Ásia

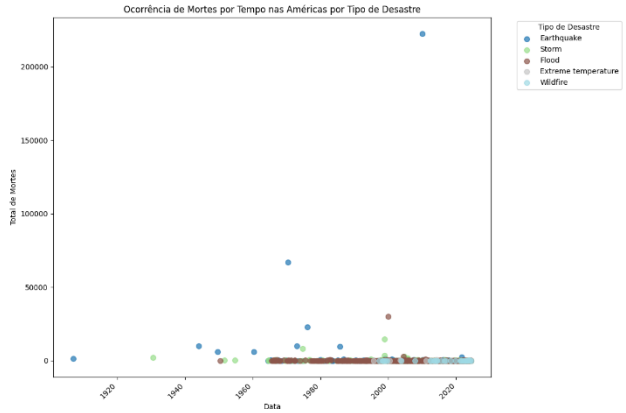

Figura 11 – Scatter plot da ocorrência de mortes por subtipo de desastre natural nas Américas

Note que é possível perceber um aumento da ocorrência dos tipos de desastres 'Flood' para ambos os continents da Ásia e nas Américas com o passer das décadas.

A percepção de tais padrões pode ser ainda mais específica caso seja analisada por subcontinetes, e para tanto, foi gerado os seguintes gráficos para analisar a ocorrência de mortes por tipo de desastres na América Latina.

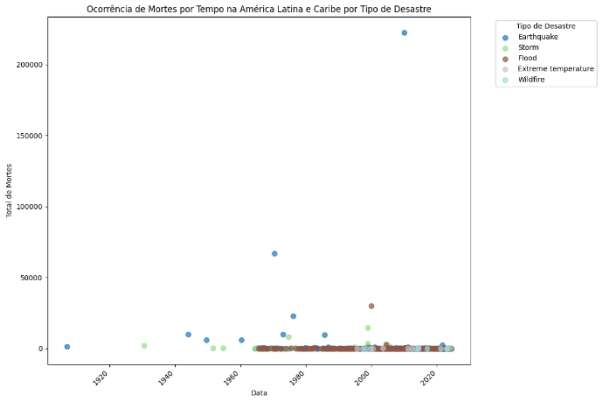

Figura 12 - Sactter Plot para a América Latina

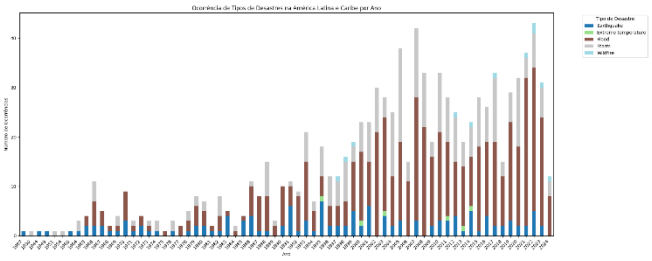

Figura 13 - Histograma para os tipos de desastres naturais na América Latina

É possível notar o aumento significativo de enchentes ao longo das décadas na região, portanto o Brasil foi analisado mais especificamente.

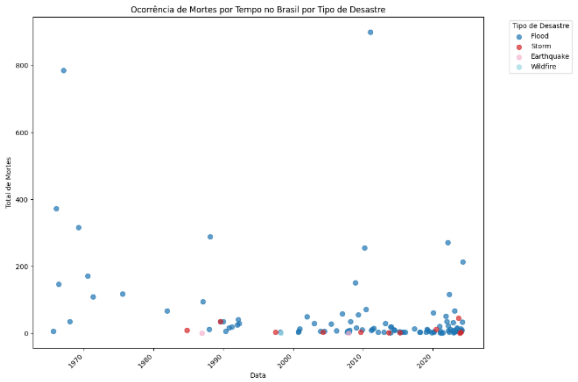

Figura 14 – Scatter plor para os tipos de desastres no Brasil

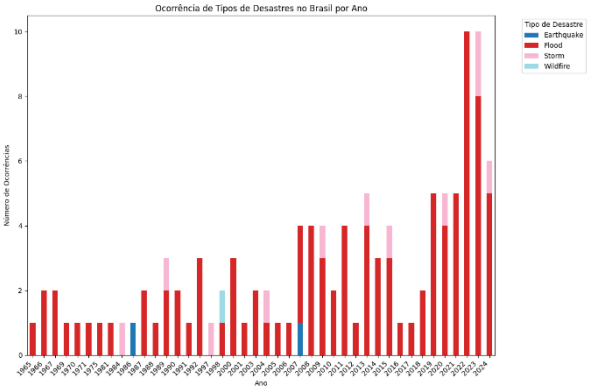

Figura 15 – Histograma para a ocorrência de desastres no Brasil

Como o aumento de ocorrências de desastres do tipo enchentes está crescendo bastante na América Latina e no Brasil, o desastre se tornou alvo de interesse a ser estudado com relação a outras regiões, e para isso alguns gráficos adicionais foram gerados.

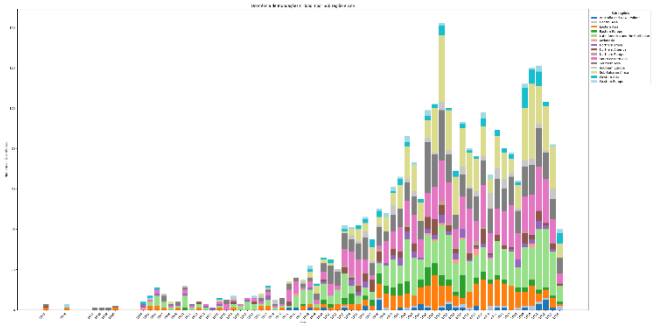

Figura 16 – Histograma para a ocorrência de enchentes nos subcontinentes

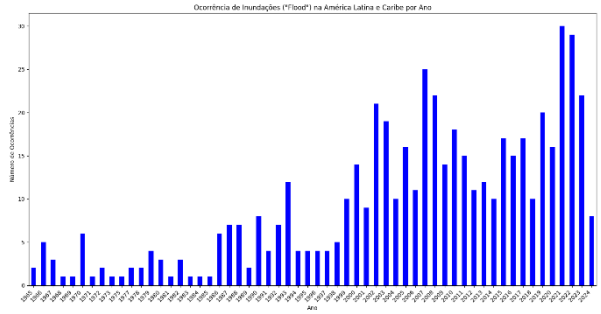

Figura 17 – Histograma para a ocorrência de enchentes no Brasil

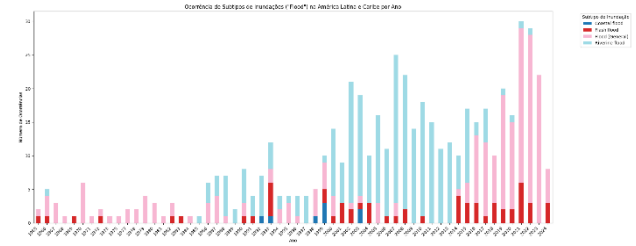

Figura 18 – Histograma para a ocorrência de subtipos de enchentes no Brasil

É possível notar um crescimento ao longo dos anos nas ocorrências de desastres naturais no geral, mas as enchentes no Brasil se tornam visivelmente crescentes nas últimas décadas, portanto este projeto tomou como foco analisar o as subregiões com base no tipo de desastre enchente (atributo 'Flood' da coluna 'Disaster Type').

Para melhor análise dos atributos numéricos, um grráfico heatmap foi gerado.

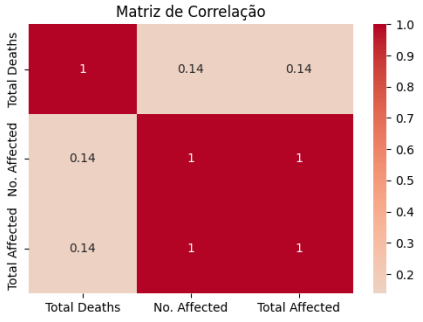

Figura 19 – Matriz de correlação entre as variáveis numéricas

As variáveis 'No. Affected' (número de afetados) e 'Total Affected' (total de afetados) se correlacionam em certo grau com o número de mortes.

VI. Resultados e Discussões

Após a fase de exploração, os dados ainda precisavam ser corretamente formatados, como a conversão de valores reais para inteiros quando necessário. Em seguida, os valores numéricos foram normalizados com o método 'MinMaxScaler' e os valores categóricos foram convertidos em vetores de números binários utilizando-se o método 'OneHotEncoder', ambos da biblioteca Scikit Learning, para o tratamento de regressão dos algoritmos.

Algumas traravitas foram levadas em consideração, devido à complexidade do problema, sendo a primeira delas tratando a base de modo geral, sem especificar um tipo de desastre, subregião ou ordenando a base por data, o qual é uma etapa fundamental para que dados do future não sejam usados para prever o passado, além de preencher dados numéricos faltantes com o algoritmo k-NN.

Como especificado para a realização do projeto, a base foi separada em 20% teste e 80% treino através do método 'train_test_split' da bilbioteca Scikit Learning.

Ainda assim, após aplicar o algoritmo utilizando as bibliotecas RandomForestRegressor, 'mean_squared_error' (erro quadrático médio) e 'r2_score' (coeficiente de determinação), os resultados obtidos foram MSE = 7780413481.8554 e R^2 = 0.0851. Após aplicar o modelo de otimizador de parâmetro 'RandomizedSearchCV' para encontrar os melhores hiperparâmetros, os resultados foram MSE = 7602348400.8124 e R^2 = 0.1060. Para o algoritmo XGBoost, após instalar o pacote "xgb", os resultados foram MSE = 8262394592.8608 e R^2 = 0.0284. A performance do algoritmo AdaBoost, que foi implementado pelo pacote "AdaBoostRegressor" e "DecisionTreeRegressor" do Scikit Learning para o modelo base, retornou os resultados MSE = 8438869282.6684 e R^2 = 0.0077. Com relação ao algoritmo de Stacking, os modelos bases incluem o 'RandomForestRegressor' e o 'GradientBoostingRegressor', sendo o modelo-meta 'LinearRegression', retornando MSE = 8487781364.7079 e R^2 = 0.0019.

A segunda tratativa envolveu o tratamento da ordenação da base, o tipo espefífico de desastre a ser analisado 'Flood' e a utilização do algoritmo K-Means para utilizer os cluster como 'features adicionais' nas amostras do algoritmo a fim de tentar faze-lo captar padrões adicionais.

Além do K-Means, foi utilizado o método de 'Elbow' para encontrar o número ótimo de 'k', que foi k = 3, para o algoritmo e a geração da imagem contou com o método estatístico 'PCA' para redução de dimensionalidade.

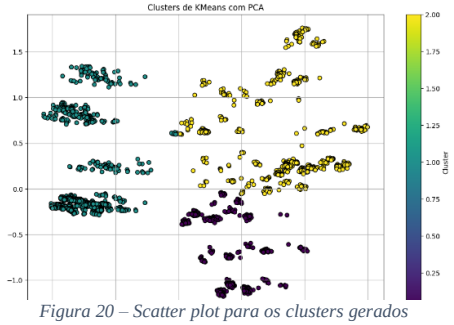

Figura 20 – Scatter plot para os clusters gerados

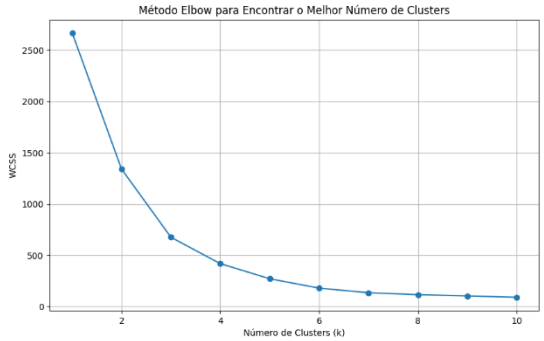

Figura 21 – Gráfico do algoritmo de Elbow

Após a geração dos clusters, separando a base a ser utilizada em 70% treino e 30% teste, e adicionado como coluna extra na base de amostras para o algoritmo, o RandomForestRegressor retornou os seguintes resultados: MSE = 248145.7455 e R^2 = -7.0009, o modelo de Random Forest aplicado ao RandomSearchCV foi MSE = 101651.3880 e R^2 = -2.2775, o modelo de XGBoost retornou MSE = 604398.5734 e R^2 = -18.4873, o modelo AdaBoost resultou em MSE = 30208.9317 e R^2 = R2 = 0.0260, e por fim o modelo Stacking retornou MSE = 129258.0002 e R^2 = R2 = -3.1676.

A terceira tratativa envolveu escolher dois subcontinentes que se comportam de forma semelhante com relação ao tipo 'Flood' (Subregion_Southern Asia e Subregion_Eastern Asia) com base nas análises estatísticas, com train_test_split de 70%/30%, considerando apenas o ano e com a base ordenada. Os resultados foram os seguintes: RandomForestRegressor com MSE = 867723.3444, R2 = -2.4820, o mesmo com RandomizedSearchCV: MSE = 686056.5865, R2 = -1.7530, o modelo XGBoost retornou MSE = 3677921.4440, R2 = -13.7589, o modelo AdaBoost teve um resultado de MSE = 1393368.3037, R2 = -4.5913, e o Stacking retornou MSE = 924061.8792, R2 = -2.7081.

VII. CONCLUSÕES E TRABALHOS FUTUROS

Os resultados obtivos não foram muito satisfatórios mas algumas condições devem ser analisadas primeiro, assim como sugestões para implementações futuras.

Os resultados da primeira tratativa foram bastante insatisfatórios, mas isso de deve à falta do tratamento adequado da base de dados, não levando em consideração a ordenação dos dados, ou seja, dados do future estavam prevendo o passado, também não levou em conta que os desastres devem ser avaliados de forma independente, pois não se comportam da mesma forma para todas as regiões, levando a um problema de generalização para os algoritmos.

A segunda tratativa também possui problemas, o K-Means não deve ser usado para variáveis categóricas pois após as mesmas serem convertidas com o One Hot Encoder, esses dados não podem ser tratados com distância euclidiana. Ao invés disso, há a alternativa de usar o K-Prototype, que lida com ambos dados categóricos e numéricos usando também a distância de Hamming, assim como a opção de verificar homogeneidade dos clusters para utilizar o mais homogêneo gerado como base de exemplos para o algoritmo, ou ainda integrar os clusters analisados com base em alguma subregião de interesse para aumentar a base de exemplo. Nesse sentido, na validação dos algoritmos pode ser interessante utilizar o erro médio absolute (MAE) pois os dados da coluna 'Total Deaths' não se comportam de forma ordenada, apresentando grande número de outliers devido à multifatoriedade que resulta no número de mortes por tipo de desastre natural.

Várias tratativas podem ser válidas, como a escolha temporal que se quer analisar, como apenas o ano. Mas isso pode acabar generalizando os dados e deixando de lado possíveis padrões temporais importantes que podem ser captados pelo algoritmo ao utilizar os meses na aplicação do algoritmo.

Ainda sobre a questão temporal, além da ordenação da base por data, pode ser conveniente escolher uma data de corte que faça sentido com as percepções derivadas das análises exploratória dos dados, observando padrões e estabelecendo uma data de corte específica ao invés de separar os exemplos em 70% treino e 30% teste, por exemplo. Além disso, a base possui datas de início e fim dos desastres, talvez fosse interessante analisar como a duração de eventos impacta sobre o número de mortes.

Sobre a especificidade das regiões, pode ser viável analisar as mortes por subcontinentes e excluir os países para evitar criar colunas demais após o One Hot Encoder, como também haver risco ao mesmo tempo de generalizer demais os dados.

Por fim, dado que para cada tipo de desastre há fatores distintos para serem analisados, como posição geográfica das regiões, questões sociais e tecnológicas que impactam na prevenção ou até o aumento do número de mortes, e eventos climáticos que influenciam em certos tipos de desastres naturais como o aquecimento global, faz-se necessário informações extras para maior entendimento dos dados e possivelmente uma melhor performance do algoritmo, ou avaliar se o método de ensemble learning é válido ou não para problemas do tipo.

REFERÊNCIAS

1. L. Breiman, Random forests. Machine learning, 2001, p.5–32.
2. C. N. Berlink, L. H. A. Lima, A. M. M. Pereira, E. A. R. Carvalho Jr, R. C. Paula, W. M. Thomas, R. G. Morato, "The Pantanal is on fire and only a sustainable agenda can save the largest wetland in the world", Brazilian Journal of Biology, 2022, vol. 82, e244200. https://doi.org/10.1590/1519-6984.244200.
3. S. S. Haykin, Neural networks and learning machines. 3rd ed. Pearson Education, 2009.K. Elissa, "Title of paper if known," aunpublished.
4. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T. Y. Liu, LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, Dez. 2017.
5. M. Kobiyama, M. Mendonça, D. A. Moreno, I. P. d. O. Maercelino, E. V. Marcelino, L. L. P. Brazetti, R. F. Goerl, M. G. S. F. Molleri, F. d. M. R. Rudorf, Prevenção de desastres naturais - Conceitos Básicos, 2006, p. 109.
6. L. S. Pereira, M. do S. C. S. Mateus, R.T. Calumby, Integração de sistemas para predição de deslizamentos de terra baseada em aprendizado de máquina, Anais Estendidos do XVIII Simpósio Brasileiro de Sistemas de Informação (SBSI), 2022.
7. K. Faceli, A. C. Lorena, J. Gama, AL, et. Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina. Grupo GEN, 2021.

8. J. Mueller, P. Massaron, Aprendizado de Máquina Para Leigos, Editora Alta Books, 2019.
9. L. Sunmin, K. Jeong-C, J. Hyung-S, J. L. Moung, L. Saro, Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. Geomatics, Natural Hazards and Risk, 8:2, 1185-1203, 2017. DOI: 10.1080/19475705.2017. 1308971.
10. F. T. Souza, T. C. Koerner, R. Chlad, A data-based model for predicting wildfires in Chapada das Mesas National Park in the State of Maranhão. Environ Earth Sci, 2015. DOI 10.1007/s12665-015-4421-8.