SCC-0275: Introdução à Ciência de Dados
Relatório do Trabalho:

FireWatch 2024: Wildfires and Drought in Brazil

Uma Análise Exploratória

Lélio Marcos Rangel Cunha
Rafael Comitre Garcia Conrado

1. Introdução

Neste trabalho, escolhemos o dataset público do *Kaggle* "*FireWatch 2024: Wildfires and Drought in Brazil*"[1], que contém dados relacionados às queimadas e secas no território brasileiro em 2024. É um tema extremamente atual: Apenas no ano de 2024, o Brasil registrou um aumento de 150% na área queimada em relação a 2023, sendo três em cada quatro hectares queimados de vegetação nativa. O mês de setembro marcou um número de 10,65 milhões de hectares queimados, um salto de 90% em relação ao mês anterior.

Além da questão da preservação ambiental, as secas e queimadas afetam diretamente a saúde da população: Entre janeiro e julho deste ano, quase 650 mil pessoas foram atendidas em hospitais públicos com doenças respiratórias relacionadas às queimadas.

Através do que foi exposto, é evidente que trata-se de uma questão de grande pertinência pública. Uma análise dos dados coletados é capaz de identificar e destacar padrões que causam secas e queimadas, áreas de maior risco - como biomas, estados e até mesmo cidades -, entre outros.

O nosso código - que faz uma análise prévia, limpeza e padronização dos dados; A análise exploratória em si, mostrando alguns dos resultados mais importantes; E posteriormente uma previsão com base em regressão - foi escrito inteiramente em Python por meio da ferramenta Google Colab. Para o agrupamento dos dados coletados e posterior manipulação deles, foi utilizada a biblioteca *Pandas.* Para a análise exploratória e posterior exibição dos resultados, optou-se por usar as bibliotecas *Seaborn* e *Matplotlib*. Por fim, para a previsão de risco de incêndio por bioma, foram utilizados modelos de regressão da biblioteca *Sklearn*.

Na seção **2** serão abordados alguns trabalhos de análise de dados também referentes às queimadas no Brasil, discutindo de forma breve quais foram as abordagens utilizadas e seus principais resultados, de forma a extrair conhecimento que será relevante para o nosso próprio trabalho.

Na seção **3** discutiremos como o dataset adotado é constituído: Sua origem, quantos e quais são seus atributos e como eles estão distribuídos; Descreveremos em detalhes como foi o pré processamento dos dados, quais foram nossas primeiras hipóteses ao iniciar a análise exploratória e como lidamos com algumas das dificuldades encontradas durante sua execução. Por fim, também detalharemos como foi implementada uma previsão de biomas com risco de incêndio por meio de técnicas e modelos de regressão.

Ao chegarmos na quarta (**4**) seção, abordaremos alguns dos experimentos realizados ao longo da análise exploratória e regressão, embora não tenhamos nos restringido apenas à segunda parte.

Por fim, na última seção (**5**), é feita uma síntese de todo o trabalho, com ênfase na discussão dos principais resultados obtidos e de como eles podem ser significativos para o contexto atual do país.

2. Trabalhos Relacionados

O trabalho de *Araújo T.B et al***[2]** detalha o desenvolvimento de um serviço capaz de lidar com alguns dos desafios do combate às queimadas no Brasil. Ele coleta e integra dados multimodais - alguns provenientes do Instituto Nacional de Pesquisas Espaciais (INPE) - e apresenta padrões observados em uma interface de usuário, facilitando o planejamento e apoio no combate às queimadas. É usado atualmente pela Brigada de Incêndio da Paraíba, na região da Caatinga, e recebeu feedback positivo de seus usuários.

Em [**3**], *Silva C. V. J et al* fazem um estudo sobre os impactos das queimadas na região amazônica ao longo das últimas décadas por meio da análise de um grande dataset que combinou diversos censos florestais. Entre suas conclusões, destaca-se a observação de que as queimadas em regiões úmidas de florestas tropicais reduzem de forma significativa a biomassa da floresta por décadas, aumentando as taxas de mortalidade de todas as árvores.

Por fim, *Souto-Oliveira C. E. et al* em [**4**] avaliam os impactos das queimadas nas florestas brasileiras e plantações de cana na qualidade do ar da cidade de São Paulo. O estudo foi feito por meio da análise de dados provenientes de diversas

estações de monitoramento do ar da CETESB, na região metropolitana de São Paulo. Eles constataram uma piora considerável na qualidade do ar, fato que pode gerar grandes impactos para a saúde da população.

3. Materiais e Métodos

Neste trabalho, fizemos uso do dataset do *Kaggle* "FireWatch 2024: Wildfires and Drought in Brazil" [4], que possui dados provenientes do Instituto Nacional de Pesquisas Espaciais (INPE) e atualizações diárias via Google Apps Script. De forma geral, o dataset inclui informações como focos de queimadas, média de dias sem precipitação por localidade, risco médio de fogo e outras, oferecendo uma visão mais precisa sobre a situação climática atual no país.

O dataset possui 8 atributos: *data*, *municipio, estado, bioma, avg_numero_dias_sem_chuva* - que indica a média de dias sem chuva em um dado município - , *avg_precipitacao* - média de precipitação -, *avg_risco_fogo* - uma média do risco de incêndio - e *avg_frp,* uma medida do risco do fogo se espalhar. São 228776 tuplas únicas.

Inicialmente, buscamos identificar alguns fatos importantes: Quais biomas, estados e municípios correm mais risco de queimada e quais são os principais fatores usados para a determinação. Além disso, também desejávamos trabalhar em um ou mais modelos que, através dos fatores detectados, possam prever regiões de maior risco de seca e queimadas.

Para isso, um tratamento nos dados precisou ser realizado. Em uma primeira análise, notou-se alguns valores faltantes: 1 em *bioma* e 93 em *avg_frp*. Como o primeiro era único, fizemos uma interpretação manual por meio dos outros atributos do registro e percebemos que é uma região predominantemente de pampa. Assim, fizemos a suposição de que esse era o bioma em questão. Em relação ao atributo *avg_frp*, fizemos imputação com estratégia de mediana.

Depois disso, procuramos determinar a existência de *outliers*, que foram detectados via plotagem do gráfico *boxplot* nos atributos numéricos *avg_numero_dias_sem_chuva, avg_precipitacao, avg_risco_fogo* e *avg_frp*. Esses outliers foram retirados por meio da estratégia dos limites superior e inferior (IQR).

Ao iniciar a Análise Exploratória propriamente dita, percebemos certas inconsistências. No gráfico plotado da Evolução da Média de Risco de Fogo por Dia, notou-se uma grande quantidade de informação faltante ou incoerente à partir do dia 2024-09-09, fato também notado pela usuária que disponibilizou o dataset no *Kagle*. Optamos por eliminar as linhas que foram determinadas após esse período.

Já em nossa tentativa de previsão - que será comentada mais adiante - , também foram necessárias certas transformações. Em uma função de pré-processamento, eliminamos os atributos *data, avg_frp, estado* e *municipio*, que não eram relevantes para o problema de regressão planejado. Os dados restantes passaram por pipeline de *StandardScaler*.

Em nossa tentativa de previsão, optamos por adotar os modelos de Regressão Linear e o Regressor Multi-Layer Perceptron (MLP Regressor), provenientes da biblioteca *Sklearn*.

Para a Regressão Linear usamos os argumentos padrões. É apenas um regressor linear de mínimos quadrados. No caso do MLP Regressor, obtivemos 3 camadas e 100 nós em cada camada oculta. Também definimos o número máximo de iterações como sendo 7000; A função de ativação adotada foi a *relu - rectified linear unit function*. *learning_rate* foi constante e o *momentum* foi deixado no default com valor de 0.9.

4. Experimentos

Nosso principal objetivo referente à previsão foi de avaliar a relação entre média de precipitação (*avg_precipitacao)*, média de dias sem chuva (*avg_dias_sem_chuva)* e o risco de incêndio/queimada relacionado (*avg_risco_fogo*) por bioma.

Tratando-se de um problema de regressão, não foi possível avaliar métricas como *F1-Score* ou *Recall*, aplicadas em contextos de classificação. Sendo assim, adotamos as métricas *MSE* (*Mean Squared Error*) - que calcula a média dos erros quadráticos entre valores previstos e os valores reais. Assim, quanto menor o seu valor, melhor é considerado o modelo - e R^2 (*R-Squared)* - que mede a proporção de variância nos dados de destino explicados pelo modelo. Quanto mais próximo de 1, melhor o modelo.

Bioma	**Modelo**	**MSE**	**R^2**
Amazônia	Regressor Linear	43.02	0.91
Amazônia	MLP Regressor	47.84	0.90
Cerrado	Regressor Linear	77.44	0.94
Cerrado	MLP Regressor	37.23	0.97
Caatinga	Regressor Linear	156.96	0.78

Caatinga	MLP Regressor	96.20	0.87
Mata Atlântica	Regressor Linear	64.74	0.83
Mata Atlântica	MLP Regressor	43.45	0.89
Pampa	Regressor Linear	551.13	-0.16
Pampa	MLP Regressor	304.77	0.36
Pantanal	Regressor Linear	246.93	0.76
Pantanal	MLP Regressor	111.84	0.89

Percebe-se pela tabela apresentada que o MLP Regressor apresentou melhores resultados para as métricas de *MSE* e R^2 em praticamente todos os biomas avaliados, exceto nos dados da Amazônia. De forma geral, podemos afirmar que os modelos aplicados tiveram excelentes resultados na previsão da relação dos atributos de precipitação com o risco de fogo.

5. Conclusões

Em suma, neste trabalho buscamos avaliar os principais fatores de risco de secas e queimadas para regiões do território brasileiro. Os dados, provenientes do Instituto Nacional de Pesquisas Espaciais (INPE), apesar de extremamente confiáveis, apresentaram certas inconsistências e valores faltantes e precisaram passar por uma etapa de reorganização e tratamento, por técnicas e ferramentas de manipulação presentes na biblioteca *Pandas*. Também optamos por eliminar as linhas do dataset referentes a datas posteriores ao dia 9 de setembro justamente por possuírem muitos valores incoerentes.

Em nossa análise exploratória, buscamos exibir visualmente alguns dos padrões encontrados por meio de gráficos que apontam os principais estados e biomas com maiores riscos de incêndio, além de suas relações com fatores como precipitação. Essa parte de visualização foi implementada com bibliotecas como *Seaborn* e *Matplotlib*.

Por fim, trabalhamos com técnicas e modelos de regressão para prever o risco de fogo nos biomas brasileiros por meio da média de precipitação nas regiões e média de dias sem chuva. Decidimos trabalhar com os modelos de Regressão Linear e MLP Regressor, obtendo bons resultados em duas diferentes métricas em ambos os casos, com destaque para o MLP Regressor.

O notebook com o código completo pode ser acessado por https://colab.research.google.com/drive/1ZNr9OJFSADVySsc7wOxNdzHr1ABIGfJ-?

usp=sharing, e o dataset do Kaggle pode ser obtido via https://www.kaggle.com/datasets/mayaravalliero/fire-watch-brazil-2024/data.

Referências

[1] FireWatch 2024: Wildfires and Drought in Brazil, Kaggle dataset, último acesso em 22/11/2024: https://www.kaggle.com/datasets/mayaravalliero/fire-watch-brazil-2024/data

[2] Araújo T. B, de Almeida D. R, Lopes Filho J. G, *et al*, (2024), *A Decision-support Service for Firefighting in Environments of Dry Tropical Forest*, 39th Brazilian Symposium on Data Bases, pages 820-826

[3] Silva C. V. J, Aragão L. E. O. C, Barlow J., *et al*, (2018), *Drought-induced Amazonian Wildfires instigate a decade-scale disruption of forest carbon dynamics*, The Royal Society, Vol. 373

[4] Souto-Oliveira C. E, Marques M. T. A, Nogueira T., *et al* (2023), *Impact of extreme wildfires from the Brazilian Forests and sugarcane burning on the air quality of the biggest megacity on South America*, Science of The Total Environment, Vol. 888