Análise Exploratória de Dados sobre incêndios florestais no Brasil

MBA em Data Science & Advanced Analytics

Bruno Henrique Freitas Paravela
Ra:2302785
Faculdade Impacta
São Paulo, Brasil

Leonardo Dias Damasceno
Ra:2303600
Faculdade Impacta
São Paulo, Brasil

Amanda Cristina Veloso
Ra:2302874
Faculdade Impacta
São Paulo, Brasil

Charles Muller
Ra: 2303526
Faculdade Impacta
São Paulo, Brasil

I. Introdução

A disciplina de Ciência de Dados está ganhando cada vez mais destaque e sendo aplicada em diversos campos das ciências sociais aplicadas, como administração, logística e gestão pública, bem como nas ciências naturais, incluindo física, química e astronomia. Isso se deve à sua capacidade de abordar uma ampla gama de problemas multidisciplinares, utilizando algoritmos, sistemas, processos e métodos científicos para extrair insights e informações de conjuntos de dados, independentemente de estarem estruturados ou não (DHAR, 2012).

Assim, observa-se um constante aumento na variedade de técnicas e métodos sendo desenvolvidos diariamente para aprimorar a qualidade das aplicações em ciência de dados. No entanto, é amplamente reconhecido na literatura que não há um processo padronizado, linear e sistemático para a condução de projetos em Ciência de Dados, especialmente durante a fase inicial de exploração e análise de dados. Em vez disso, o processo é mais semelhante a um ciclo, exigindo repetidas iterações e reavaliações das decisões tomadas em estágios anteriores, desde a coleta dos dados até a construção de modelos capazes de descrever com precisão os fenômenos em questão.

Inicialmente, os dados são armazenados em diferentes formatos, como arquivos, bancos de dados ou APIs web, e são posteriormente importados para um ambiente de desenvolvimento. É crucial adequar esses dados aos objetivos do projeto, organizando e formatando-os de modo a viabilizar sua utilização na construção de gráficos e modelos de aprendizado de máquina. Em seguida, ocorre a etapa de transformação dos dados, que engloba desde a aplicação de filtros até a criação de novas variáveis, derivadas de variáveis existentes, como exemplificado pelo cálculo do índice de massa corporal. Após essa fase, são elaboradas visualizações que elucidam causalidades e/ou correlações previamente desconhecidas ou incertas. Uma vez que as questões levantadas são suficientemente precisas para os objetivos do projeto, é natural a construção de modelos que tornem o processo escalável e facilmente reproduzível. Conforme destacado por Wickham e Grolemund (2017), tais modelos representam ferramentas matemáticas ou computacionais fundamentalmente escaláveis. esse processo é cíclico, pois a adição de novas descobertas e variáveis demanda retornar à fase de transformação dos dados e criar novos gráficos que incorporem as modificações introduzidas. Por fim, concluído o ciclo de exploração, a comunicação dos resultados é um aspecto crítico de um projeto de Ciência de Dados. Um projeto, por mais robusto que seja em termos de modelos e visualizações, carece de relevância se não for comunicado eficazmente. Neste estudo, apresenta-se uma análise exploratória de dados sobre incêndios florestais no Brasil.

II. Objetivo

O objetivo primordial deste estudo é compreender, identificar e documentar padrões, sejam eles geográficos, sazonais ou outros, relacionados aos incêndios florestais nos biomas brasileiros, com o intuito de formular hipóteses e, quando possível, verificar suposições. Para alcançar este objetivo, serão empregadas uma série de técnicas e práticas comuns durante a análise exploratória de dados. Dessa forma, são delineados os seguintes objetivos específicos:

(a) Coletar dados provenientes das bases do Instituto Nacional de Pesquisas Espaciais (INPE)

(b) Realizar o processamento, organização, limpeza e filtragem dos dados;

(c) Realizar uma análise crítica dos resultados obtidos.

III. Referencias Conceituais

A. Análise Exploratória de Dados

No âmbito das práticas em Ciência de Dados, destaca-se a Análise Exploratória de Dados (AED). Conforme descrito por Peng (2016), os objetivos da AED abrangem diversas áreas, incluindo a identificação de relações entre variáveis de interesse, a investigação de evidências a favor ou contra hipóteses declaradas, a detecção de problemas nos dados coletados (como dados faltantes ou erros de medição) e a identificação de lacunas que demandam mais coleta de dados. Neste contexto, a ênfase não recai necessariamente sobre a apresentação detalhada dos dados ou evidências, mas sim na extração de informações significativas a partir dos dados disponíveis. Em outras palavras, na AED, o foco do analista não reside primariamente na qualidade estética das visualizações produzidas, mas sim nas informações substanciais que podem ser derivadas dos dados. Portanto, trata-se de uma prática recorrente em projetos de Ciência de Dados, uma vez que possibilita a descoberta de insights até então desconhecidos por meio da observação e análise do relacionamento entre as variáveis.

B. Amazônia Legal (IBGE)

A Amazônia Legal representa uma região delimitada por razões sociopolíticas, em resposta aos desafios ambientais, políticos e de infraestrutura compartilhados pelos Estados brasileiros que a compõem. Abrangendo uma extensão territorial de 5.217.423 km², corresponde a 61% do território nacional e inclui 20% do bioma Cerrado, além de parte do Pantanal no estado do Mato Grosso. Os nove Estados que integram essa área abrigam aproximadamente 55% da população indígena do Brasil. De acordo com o Ministério do Meio Ambiente, conforme destacado no Caderno da Região Hidrográfica, a importância dessa região reside em sua rica diversidade ambiental e na abundância de recursos naturais. Destaca-se a presença da Bacia Amazônica, a maior bacia hidrográfica do mundo, que detém cerca de um quinto do volume total de água doce do planeta. Além disso, a região abriga aproximadamente 40 mil espécies de plantas, mais de 400 espécies de mamíferos, quase 1.300 espécies de aves, cerca de 3 mil espécies de peixes e milhões de insetos.

Um bioma representa uma unidade biológica ou uma região geográfica cujas características específicas são definidas pelo clima regional, fisionomia da vegetação, tipo de solo, altitude, entre outros critérios. Esses biomas constituem tipos distintos de ecossistemas, habitats ou comunidades biológicas que exibem certo grau de homogeneidade (Faria, 2019).

De acordo com informações do Ministério do Meio Ambiente, o Brasil é composto por seis biomas, cada um com características singulares: Amazônia, Caatinga, Cerrado, Mata Atlântica, Pampa e Pantanal. Cada um desses ambientes abriga diferentes tipos de vegetação e fauna. Embora os dados utilizados neste estudo englobem informações sobre os seis biomas brasileiros, o foco da Análise Exploratória de Dados (AED) está direcionado principalmente para o bioma Amazônico e o Cerrado, considerando que esses biomas ocupam mais da metade da área total do território brasileiro.

C. Queimadas no Brasil

O Brasil, com sua rica biodiversidade e vastas florestas, enfrenta um desafio persistente: o aumento das queimadas. As chamas que consomem paisagens e liberam gases poluentes não são apenas um problema ambiental, mas também uma questão social e econômica complexa, com raízes profundas na dinâmica populacional, nas mudanças climáticas e em diversos fatores socioeconômicos. O Brasil, com mais de 212 milhões de habitantes, concentra grande parte da sua população em áreas urbanas. Esse crescimento urbano, impulsionado por fatores como êxodo rural e oportunidades de trabalho, coloca pressão sobre os recursos naturais, especialmente nas regiões metropolitanas e suas periferias. A expansão urbana desordenada, muitas vezes sem infraestrutura adequada, leva à grilagem de terras e à ocupação irregular de áreas verdes, criando condições propícias para o surgimento de focos de queimadas.

As mudanças climáticas, com o aumento da temperatura média global e a intensificação de eventos climáticos extremos, como secas e ondas de calor, contribuem para o aumento da frequência e da intensidade das queimadas. O clima mais seco torna a vegetação mais propensa à combustão, e os eventos extremos podem dificultar o combate ao fogo.

Diversos fatores socioeconômicos também estão atrelados ao problema das queimadas. A pobreza, a falta de acesso à terra e a informalidade no trabalho rural impulsionam atividades como o desmatamento ilegal para a extração de madeira e o cultivo em áreas impróprias, frequentemente utilizando o fogo como ferramenta. Além disso, a falta de educação ambiental e de fiscalização adequada contribuem para a persistência das queimadas.

É importante ressaltar que a problemática das queimadas não se manifesta de forma homogênea em todo o Brasil. Cada região apresenta suas particularidades, com diferentes biomas, dinâmicas populacionais e atividades socioeconômicas.

A amazônia, a região amazônica, com sua rica floresta tropical úmida, enfrenta um dos maiores desafios em relação às queimadas. O desmatamento ilegal para pecuária, agricultura e exploração madeireira, muitas vezes impulsionado por grandes empresas e grupos de poder, é um dos principais fatores por trás da devastação.Cerrado**,** o Cerrado, o segundo maior bioma do Brasil, também tem sofrido com o aumento das queimadas. A expansão da fronteira agrícola, com o avanço da soja e do gado, e a falta de políticas públicas eficazes de proteção ambiental contribuem para a destruição desse bioma crucial para a biodiversidade e o clima do país, a região nordeste, no Nordeste, as secas frequentes e a pobreza rural são fatores importantes para o surgimento de queimadas. A população, muitas vezes dependente da agricultura familiar de subsistência, utiliza o fogo como forma de limpar a terra para o plantio, mesmo com os riscos que isso representa.

IV. Procedimentos para a Aquisição e Identificação dos Dados

A. Categorização do Estudo

Uma pesquisa pode ser categorizada a partir de três perspectivas distintas: sua natureza, abordagem de dados, objetivos e procedimentos técnicos. No que diz respeito à natureza da pesquisa, Silva e Menezes (2005) propõem as classificações de pesquisa básica e aplicada. A pesquisa básica tem como objetivo a produção de novos conhecimentos que contribuam para o avanço do estado da arte científica em uma determinada área, sem necessariamente visar uma aplicação imediata. Por outro lado, a pesquisa aplicada busca a construção de conhecimento por meio de aplicações práticas, lidando com questões e interesses específicos. Portanto, este estudo, de caráter exploratório, pode ser considerado como uma pesquisa aplicada. Além disso, quanto aos tipos de abordagem utilizados na coleta e análise dos dados, uma pesquisa pode ser classificada como quantitativa ou qualitativa. Uma pesquisa é considerada quantitativa quando pode ser expressa em termos numéricos por meio de modelos matemáticos, estatísticos e classificatórios. Por outro lado, uma pesquisa é qualitativa quando busca compreender ou transmitir aspectos subjetivos sobre o tema em questão, sem priorizar a representação numérica dos resultados e procedimentos. Assim, o presente estudo engloba elementos de ambas as categorias, indicando que se trata de uma pesquisa com abordagens qualitativas e quantitativas. Dependendo dos objetivos da pesquisa, ela pode ser classificada como exploratória, descritiva ou explicativa. Segundo Gil (2007), uma pesquisa que visa identificar padrões e fatores que influenciam ou contribuem para a ocorrência de um fenômeno tem caráter explicativo. Portanto, este estudo apresenta aspectos descritivos e explicativos, pois busca elucidar padrões e características relacionadas aos incêndios florestais no Brasil e promover uma compreensão mais aprofundada sobre o tema.

B. Instrumentação

Nesta seção, são detalhadas as ferramentas empregadas na elaboração deste estudo. Incluem-se a linguagem de programação Python e uma seleção de seus pacotes de código-fonte aberto, juntamente com o ambiente de desenvolvimento integrado utilizado.

C. Linguagem de Programação Python

A linguagem de programação Python foi selecionada para a condução deste estudo devido à sua natureza de código aberto, disponibilidade em diversos sistemas operacionais e ampla utilização na comunidade de desenvolvimento. Como resultado, ao realizar uma análise em Python, torna-se acessível a qualquer interessado a reprodução dos procedimentos adotados. Além disso, diversos motivos para essa escolha são ressaltados por autores como Jake VanderPlas (2016) em seu livro " Python Data Science Handbook ":

a) Python oferece um vasto conjunto de bibliotecas para modelagem estatística de dados, aprendizado de máquina, visualização, importação e manipulação de dados.

b) Dispõe de ferramentas poderosas para a comunicação de resultados;

c) Conta com um Ambiente de Desenvolvimento Integrado (IDE) especialmente projetado para análise de dados e programação voltada para as práticas estatísticas;

d) Oferece facilidade de meta-programação, permitindo a criação de funções de forma concisa e sucinta através de recursos de meta-programação.

D. Ambiente de Desenvolvimento Integrado

O Ambiente de Desenvolvimento Integrado (IDE) desempenha um papel crucial no processo de desenvolvimento de software e análise de dados em Python. Um exemplo proeminente é o Jupyter Notebook, que se tornou uma ferramenta essencial para cientistas de dados, pesquisadores e desenvolvedores em todo o mundo.

O Jupyter Notebook oferece uma interface interativa baseada na web que permite criar e compartilhar documentos que contêm código Python executável, visualizações de dados, texto explicativo e equações matemáticas. Esses documentos são organizados em células, onde cada célula pode conter código Python ou texto formatado usando a linguagem Markdown.

O Jupyter Notebook possui várias vantagens que o tornam ideal para ciência de dados:

*a) **Interatividade:** Os usuários podem executar código Python diretamente no notebook, visualizar os resultados imediataemnte e iterar rapidamente em análises e visualizações;*

*b) **Reprodutibilidade:** O notebook registra todas as etapas de análise de dados, desde a preparação dos dados até a visualização final. Isso facilita a reprodução dos resultados e compartilhamento com outras possoas.*

*c) **Visualização de dados**: O Jupyter suporta a exibição de gráficos e visualização diretamente no notebook, o que é fundamental para explorar e comunicar os insigths obtidos a partir dos dados.*

E. Banco de Dados

Foi utilizada uma base de dados retiradas do kaggle onde se detalha os focos de queimada, estados, ano e mês do evento, o formato do arquivo se encontra em CSV.

F. Pacotes e bibliotecas

Python é amplamente reconhecido como uma das linguagens mais poderosas para ciência de dados, em grande parte devido à sua vasta coleção de pacotes e bibliotecas especializados. Aqui estão as principais bibliotecas e pacotes Python utilizadas no projeto:

*a) **Pandas:** Uma das bibliotecas mais essenciais para manipulação e análise de dados em Python. O Pandas oferece estruturas de dados flexíveis e eficientes, como o DataFrame que facilita a manipulação de conjuntos de dados tabulares;*

*b) **Scikit-Learn:** Também conhecida como Sklearn, é uma biblioteca de aprendizado de máquina em Python que oferece uma ampla variedade de algoritmos de aprendizado supervisionado e não supervisionado, bem como ferramentas para pré-processamento de dados, avaliação de modelos e seleção de características. No código fornecido, está sendo utilizado o KMeans, que é um algoritmo de clusterização, e o LabelEncoder e StandardScaler, que são utilizados para pré-processamento de dados;*

*c) **PyECLAT:** Esta é uma biblioteca menos conhecida, mas muito útil, que implementa o algoritmo ECLAT (Equivalence Class Clustering and Bottom-Up Lattice Traversal) para mineração de itens frequentes em conjuntos de dados. É especialmente útil em tarefas de mineração de dados.*

*d) **Seaborn:** Esta é uma biblioteca de visualização de dados para Python. Ela é construída em cima de outra biblioteca famosa chamada Matplotlib, mas oferece uma interface de alto nível, o que significa que é mais fácil de usar para criar gráficos estatísticos informativos e visualmente agradáveis.*

*e) **Matplotlib** é uma biblioteca Python conhecida por ser uma ferramenta de visualização de dados. Ela é usada para criar gráficos estáticos, animações e visualizações 2D e 3D.*

*f) **Requests a** biblioteca Requests em Python é uma ferramenta poderosa para desenvolvedores que desejam simplificar a interação com a web. Ela permite enviar e receber informações através de requisições HTTP, tornando a comunicação com serviços online mais fácil e eficiente.*

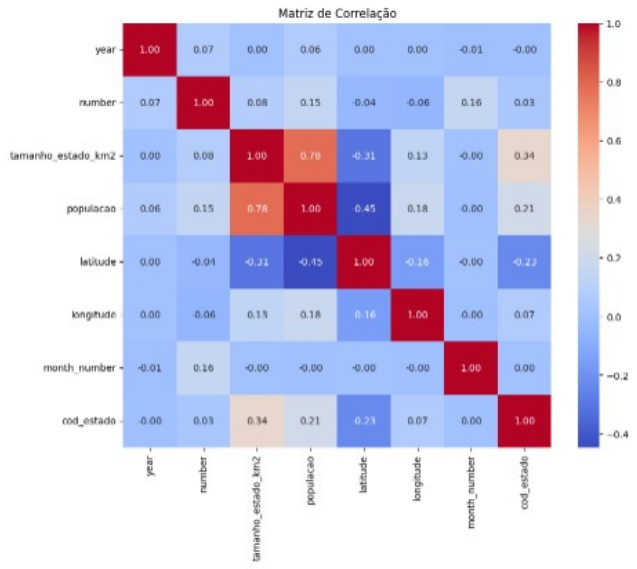
V. Análise de Dados

Uma pergunta pertinente que surge após uma pequena análise diz respeito aos meses que apresentam uma maior incidência de focos de incêndio, assim apresentando uma correlação entre quantidade de queimadas, estados e população. Usando a análise de dados podemos realizar vários tipos de ralações entre tamanho do estado com o número de queimadas, atingindo sua infraestrutura ou qual o tipo de bioma existe naquela região do pais.

Tabela de Relacionamento entre Estados, queimadas e população

	year	state	month	number	date	tamanho_estado_km2	pupulacao	latitude	longitude	month_number
0	1998	AC	JAN	0.0	1998-01-01	164123.964	514.050	-8.77	-70.55	1
1	1999	AC	JAN	0.0	1999-01-01	164123.964	527.937	-8.77	-70.55	1
2	2000	AC	JAN	0.0	2000-01-01	164123.964	541.873	-8.77	-70.55	1
3	2001	AC	JAN	0.0	2001-01-01	164123.964	574.355	-8.77	-70.55	1
4	2002	AC	JAN	0.0	2002-01-01	164123.964	586.942	-8.77	-70.55	1
5	2003	AC	JAN	10.0	2003-01-01	164123.964	600.595	-8.77	-70.55	1
6	2004	AC	JAN	0.0	2004-01-01	164123.964	630.328	-8.77	-70.55	1
7	2005	AC	JAN	12.0	2005-01-01	164123.964	669.736	-8.77	-70.55	1
8	2006	AC	JAN	4.0	2006-01-01	164123.964	686.652	-8.77	-70.55	1
9	2007	AC	JAN	0.0	2007-01-01	164123.964	68665.2	-8.77	-70.55	1

Gráfico de Correlação



Um dos principais desafios enfrentados durante o desenvolvimento deste estudo está relacionado à complexidade inerente à análise de uma grande quantidade de observações sobre incêndios florestais. Em muitas ocasiões, a magnitude dos dados em si pode se tornar um obstáculo para a viabilidade do projeto.

VI. Link para consulta do código

O link para consulta do código está:
https://github.com/CharlesMuller007/Trabalho

VII. Conclusão

A Análise Exploratória de Dados (AED) é uma abordagem metodológica amplamente utilizada que se vale de conhecimentos multidisciplinares com o propósito de obter insights e informações relevantes sobre um tema específico. No entanto, não existe uma metodologia linear universalmente aplicável para conduzir uma análise exploratória de dados. Este estudo buscou esclarecer alguns procedimentos comuns frequentemente encontrados no dia a dia de um profissional de ciência de dados. Ao aplicar essa abordagem ao contexto dos incêndios florestais no Brasil, questões pertinentes e recorrentes foram abordadas, e padrões de ocorrências foram identificados ao longo do período analisado. Através da análise exploratória, podemos desvendar padrões, identificar tendências e direcionar ações para a prevenção e o combate a esse problema complexo. A busca por soluções multidisciplinares e a colaboração entre diferentes áreas do conhecimento são fundamentais para construirmos um futuro mais sustentável, onde os incêndios florestais não representem uma ameaça à nossa rica biodiversidade e ao bem-estar da população.

Referências

Ocorrência de incêndios florestais no Parque Nacional da Chapada dos Veadeiros, Goiás. Ciência Florestal, Santa Maria, v. 16, n. 2, p. 153-161, 2006a.

GIL, A. C. Métodos e técnicas de pesquisa social. 5. ed. São Paulo. ATLAS, 1999. 206 p.

ALVES, K. M. A. da S.; NÓBREGA, R. S. Uso de dados climáticos para análise espacial de risco de incêndio florestal. Mercator-Revista de Geografia da UFC, Universidade Federal do Ceará, v. 10, n. 22, p. 209–219, 2011. 20, 21.

PINHEIRO, Ismael, D. P.; CUNHA, Sonia, B. da.; CARVAJAL, Santiago, R; GOMES, Gastão, C. Estatística básica – arte de trabalhar com dados. Rio de Janeiro: Elsevier, 2009.

ELLISON, A. M. Exploratory data analysis and graphic display: design and analysis of ecological experiments. New York: Chapman & Hall, 1993. p. 14-41.

NSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (INPE). SGI 2.5 – Introdução ao Sistema de Informações Geográficas – SGI. magem Geosistemas São José dos Campos: Instituto Nacional de Pesquisas Espaciais, 1995.