Análise Exploratória de Dados sobre incêndios florestais no Brasil

MBA em Data Science & Advanced Analytics

Leonardo Dias Damasceno Ra:2303600
Bruno Henrique Paravela Ra:2302785
Charles Muller Ra: 2303526
Amanda Cristina Veloso Ra:2302874

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 do natural park Montesinho em Portugal

II. Objetivo

Este artigo investiga a aplicação de modelos de regressão linear para prever a probabilidade de incêndios florestais na área do parque de Montesinho, examinando a relação entre a área queimada e vários fatores ambientais. O estudo emprega técnicas de aprendizado de máquina para analisar dados históricos, com o objetivo de aprimorar a compreensão da dinâmica do fogo e contribuir para estratégias eficazes de gerenciamento de incêndios. Os aspectos técnicos do código, incluindo a identificação e a remoção de outliers, são discutidos, e o desempenho do modelo é avaliado. Os resultados sugerem que determinadas variáveis, como temperatura e umidade, são preditores significativos da ocorrência de incêndios. O código usado para a análise é disponibilizado publicamente para pesquisa e validação adicionais

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. Parque Natural de Montesinho

O Parque Natural de Montesinho, localizado nos municípios de Vinhais e Bragança, no nordeste de Portugal, é uma área protegida de elevada importância ecológica e cultural. Este parque caracteriza-se por uma grande diversidade morfológica, geológica e climática, que se reflete na rica biodiversidade de espécies animais e vegetais adaptadas ao meio físico, contribuindo para uma paisagem única e diversificada.

A área abriga mais de 120 espécies de aves reprodutoras, 70% das espécies terrestres de animais que ocorrem em Portugal, incluindo uma das mais importantes populações de lobo ibérico da Península Ibérica, e é o único local em Portugal onde se encontra naturalmente a espécie Euonymus europaeus.

A luta contra os incêndios florestais nesta área é de extrema importância, não só pela preservação da biodiversidade, mas também pela proteção das comunidades locais e pela manutenção dos serviços ecossistêmicos que o parque fornece. Incêndios florestais podem causar a degradação dos ecossistemas, afetando a biodiversidade e o equilíbrio na provisão de bens e serviços ambientais, econômicos e sociais.

A área ardida em Portugal tem apresentado uma tendência crescente desde meados da década de 80, com anos de máximos em 2017, 2003 e 2005, e os incêndios rurais constituem um dos principais obstáculos à sustentabilidade da floresta e dos ecossistemas associados.

Em 2022, Portugal registrou 10.583 fogos florestais, que resultaram numa área ardida de 110.183 hectares, evidenciando a gravidade e a frequência dos incêndios no país.

A prevenção e o combate eficaz aos incêndios são essenciais para a conservação do Parque Natural de Montesinho e para a proteção das espécies e habitats que ele suporta. A gestão adequada do território, a educação e a sensibilização das comunidades locais, e a implementação de políticas públicas efetivas são medidas fundamentais para mitigar o risco de incêndios e garantir a resiliência dos ecossistemas.

Além disso, a existência de planos de defesa da floresta contra incêndios (PMDFCI) atualizados é crucial para a prevenção e o combate aos incêndios em áreas protegidas como o Parque Natural de Montesinho. No entanto, tem sido reportado que nem todos os concelhos abrangidos por áreas protegidas possuem PMDFCI em vigor, o que representa um desafio adicional para a proteção dessas áreas.

Em resumo, o Parque Natural de Montesinho é um patrimônio natural de Portugal que requer medidas efetivas de prevenção e combate a incêndios florestais para assegurar a sua conservação e a sustentabilidade dos recursos que oferece

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.

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. 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) **Numpy (numpy):** Uma biblioteca fundamental para computação científica com Python, oferece suporte para arrays e matrizes multidimensionais, juntamente com uma coleção de funções matemáticas para operar com estas estruturas.*

*c) **Matplotlib (matplotlib.pyplot):** Biblioteca de plotagem para Python e sua extensão matemática NumPy, usada para criar uma ampla variedade de gráficos e plots estáticos, animados e interativos.*

*d) **Scikit-learn (sklearn):** Uma das principais bibliotecas de aprendizado de máquina para Python, oferece ferramentas simples e eficientes para análise preditiva de dados, incluindo regressão, classificação, redução de dimensionalidade, entre outros.*

*e) **Scipy (scipy):** Biblioteca usada para realizar tarefas de computação científica e técnica, abrangendo módulos para otimização, álgebra linear, integração, entre outros.*

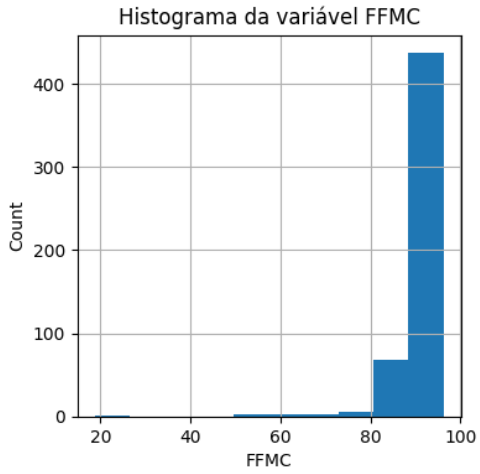
*f) **Warnings (warnings):** Utilizada para gerenciar avisos em Python, permitindo ao usuário controlar como os avisos são apresentados, ignorados ou tratados, útil para suprimir avisos em bibliotecas como o Scikit-learn.*

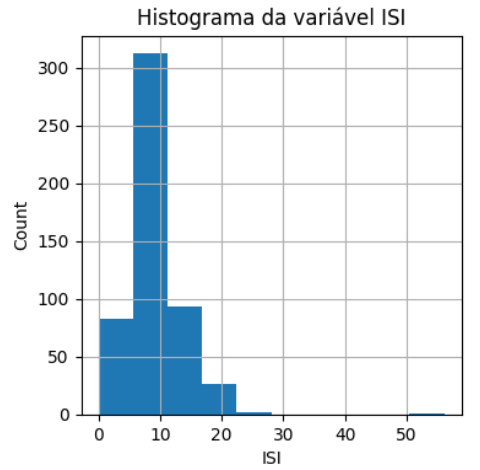
V. DATASET

O dataset é composto pelas seguintes colunas:
X: Coordenada espacial x dentro do mapa do Parque Montesinho, variando de 1 a 9.
Y: Coordenada espacial y dentro do mapa do Parque Montesinho, variando de 2 a 9.
mês: Mês do ano, de "jan" a "dez".
dia: Dia da semana, de "seg" a "dom".
FFMC: Índice FFMC do sistema FWI, de 18,7 a 96,20.
DMC: Índice DMC do sistema FWI, de 1,1 a 291,3.
DC: Índice DC do sistema FWI, de 7,9 a 860,6.
ISI: Índice ISI do sistema FWI, de 0,0 a 56,10.
temp: Temperatura em graus Celsius, de 2,2 a 33,30.
RH: Umidade relativa em %, de 15,0 a 100.
wind: Velocidade do vento.
rain: Precipitação.
area: Área queimada em quilômetros quadrados.

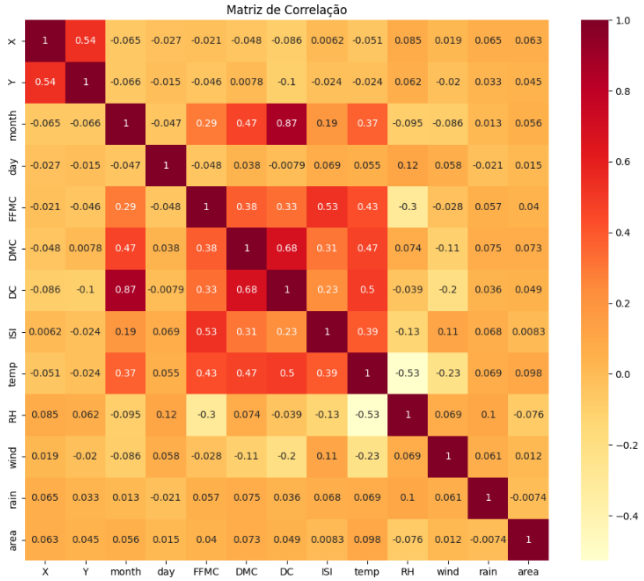
VI. ANÁLISE DE DADOS

O código começa com a importação das bibliotecas necessárias e a preparação do conjunto de dados para análise. Depois de carregar os dados, é gerado um histograma para visualizar a distribuição das variáveis. O histograma revela dois outliers importantes que são as variáveis FFMC e ISI, que são valores extremos que se desviam significativamente do restante dos dados. Os outliers podem ter um efeito desproporcional no desempenho do modelo, podendo levar a resultados enganosos. Portanto, eles são identificados e removidos para melhorar a precisão do modelo.





Também foi feito a matriz correlação entre as variáveis para identitificar possiveis correlações e guiar o modelo, para isso foi preciso trasnformar a coluna month e day em números que futuramente também serão usadas no modelo de regressão.



VII. Introdução ao Modelo de Regressão

O modelo de regressão linear implementado no código do anexo tem como objetivo prever a variável dependente "area", que representa a área queimada por incêndios florestais, com base em várias outras variáveis independentes disponíveis no conjunto de dados "forest fire.csv". Este conjunto de dados inclui variáveis como o índice de umidade do combustível (FFMC), o índice de umidade do solo (DMC), o índice de seca (DC), o índice de propagação inicial do fogo (ISI), temperatura, umidade relativa (RH), velocidade do vento e precipitação, além de variáveis categóricas como mês e dia da semana.

VIII. Etapas do Modelo

*a) **Preparação dos Dados:** O código começa importando as bibliotecas necessárias e carregando o conjunto de dados. Uma transformação é aplicada à variável "month" para convertê-la de categórica para numérica, facilitando o processamento matemático.*

*b) **Divisão dos Dados:** Os dados são divididos em conjuntos de treinamento e teste usando a função train_test_split. Isso é essencial para validar a eficácia do modelo em dados não vistos durante o treinamento.*

*c) **Padronização dos Dados:** Os dados de treinamento são padronizados usando StandardScaler, o que é uma prática comum para modelos de regressão linear, pois ajuda a evitar que variáveis com maior magnitude dominem o modelo.*

*d) **Treinamento do Modelo:** Um modelo de regressão linear é treinado com os dados de treinamento. O treinamento envolve ajustar uma linha que minimiza a soma dos quadrados das diferenças entre os valores observados e os valores previstos pela linha de regressão.*

*e) **Avaliação do Modelo:** O modelo é avaliado usando o erro quadrático médio (MSE) para quantificar o desempenho do modelo. O MSE é calculado como a média dos quadrados das diferenças entre os valores observados e os valores previstos.*

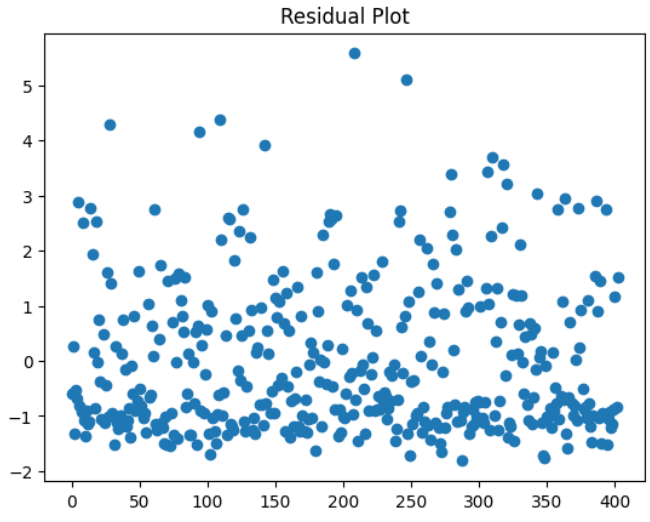
IX. Resultados e Interpretação

O código não fornece detalhes explícitos sobre os resultados numéricos finais, como os valores dos coeficientes ou o desempenho do modelo no conjunto de teste. No entanto, a função model_results é projetada para calcular o MSE, que é uma métrica chave para avaliar a precisão das previsões do modelo. Um MSE menor indica que o modelo tem um bom desempenho, enquanto um valor maior sugere que o modelo pode não estar capturando todas as nuances dos dados.

```
RMSE: 1.371

*****Shapiro-Wilks test for Normality*****

Residuals do no follow normal distribution!
```



X. Conclusão

Após a conclusão da análise, o desempenho do modelo é avaliado usando métricas como o erro quadrático médio, o erro absoluto médio e a pontuação R-quadrado. Os resultados indicam que, embora o modelo possa identificar algumas relações entre os fatores ambientais e a área queimada, a presença de outliers e a complexidade da dinâmica do fogo significam que o poder preditivo do modelo é limitado. A conclusão também sugere áreas para pesquisas futuras, como a incorporação de variáveis adicionais, o uso de algoritmos de aprendizado de máquina mais complexos e a aplicação do modelo a diferentes regiões geográficas para melhorar sua precisão de previsão. O estudo demonstra o potencial do aprendizado de máquina na ciência ambiental, especialmente no contexto da previsão e do gerenciamento de desastres naturais.

Referências

Dataset: https://www.kaggle.com/datasets/sumitm004/forest-fire-area/data

PARQUE NACIONAL DE MONTESINHO https://www.rotaterrafria.com/pages/221/?geo_article_id=7248

FUNCAÇÃO FRANCISCO MANUEL DOS SANTOS https://ffms.pt/pt-pt/livraria/os-incendios-florestais-em-portugal

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.