Relatório do Projeto: Desenvolvimento de Dataset e Modelo para Previsão de Risco de Incêndios Florestais no Estado de São Paulo (2022)

Lancelot Chagas Rodrigues / 554707

Ana Carolina Martins da Silva / 555762

Kauan Alves Batista / 555082

---

1. Motivação do Projeto

Os incêndios florestais representam uma ameaça considerável ao meio ambiente, à biodiversidade, à saúde pública e à economia. Reconhecendo a criticidade desses eventos, nosso projeto foi motivado pela necessidade de desenvolver uma solução baseada em dados para aprimorar a capacidade de previsão do risco de ocorrência de incêndios no estado de São Paulo. A escolha do ano de 2022 para o desenvolvimento inicial deu-se pela disponibilidade de dados anuais completos, permitindo um ciclo de desenvolvimento mais focado. Embora não tenhamos utilizado dados de múltiplos anos nesta fase primariamente por questões de performance e tempo de processamento, o modelo e o pipeline de dados foram pensados para acomodar futuras atualizações e a inclusão de um histórico mais extenso.

2. Objetivo Principal

O objetivo central de nosso trabalho foi a criação de um dataset abrangente e o desenvolvimento de um modelo de Machine Learning capaz de prever a ocorrência de incêndios florestais (ocorreu_incendio: 1 para sim, 0 para não) em nível municipal e diário para o estado de São Paulo, referente ao ano de 2022. Um objetivo secundário, de grande importância, foi otimizar o modelo para maximizar a detecção de incêndios reais (alto recall para a classe "Incêndio"), dada a severidade das consequências de um evento não previsto. Adicionalmente, projetamos que o modelo e o dataset pudessem, em fases futuras, ser integrados a uma aplicação web interativa para análises e, potencialmente, dar suporte a funcionalidades de Q&A com Modelos de Linguagem Grandes.

3. Estratégias Adotadas e Desenvolvimento do Dataset/Modelo

A construção da solução seguiu um processo iterativo, englobando coleta, integração, limpeza de dados e engenharia de features, culminando na modelagem e otimização.

- **Fontes de Dados Primárias Utilizadas:**
    - Focos de Queimada (INPE/TerraBrasilis): Utilização do dataset focos_br_todos-sats_2022.csv.
    - Cobertura do Solo (MapBiomas): Dados de área por classe (mapbiomas_areas_municipios_sp_2022_long_v3.csv), processados para proporções.
    - Dados Meteorológicos Detalhados (INMET): Coleta de dados horários de 38 estações em São Paulo para 2022, agregados para valores diários.
    - Cadastro de Municípios: Arquivo (municipios.csv) com informações cadastrais dos municípios brasileiros.
- **Processamento e Integração de Dados:**
    - Filtragem dos dados de focos para São Paulo, 2022.
    - Criação da variável alvo binária ocorreu_incendio.
    - Construção de um dataset base (dataset_SP_2022_focos_meteo_mapbiomas_norm.csv).
    - Desenvolvimento de um script (preparar_dados_meteo.py) para processar e integrar dados do INMET, resultando em dataset_SP_2022_completo_com_inmet.csv.
- **Engenharia de Features Avançada:**
    - Criação de features temporais (Lags e Médias Móveis) para variáveis meteorológicas, FRP e numero_dias_sem_chuva_first.
    - Implementação de dias_secos_consecutivos (baseado em dados INMET).
    - Desenvolvimento de FWI-like Proxies (FFMC, DMC, DC, ISI, BUI, FWI) usando dados INMET defasados.
    - Criação de features de interação como dias_secos_X_umidade_baixa_lag1.
- **Pré-processamento para Modelagem:**
    - Imputação de NaNs (mediana).
    - Target Encoding para municipio.

    - Escalonamento com StandardScaler.
- **Abordagem de Modelagem:**
    - Divisão temporal dos dados.
    - Tratamento de desbalanceamento (SMOTE e scale_pos_weight no LightGBM).
    - Otimização de hiperparâmetros (RandomizedSearchCV) focando no F2-score.
    - Ajuste do limiar de decisão.

4. Ferramentas Utilizadas

- Linguagem: Python.
- Manipulação de Dados: Pandas, NumPy.
- Machine Learning: Scikit-learn, Category Encoders, Imbalanced-learn.
- Modelo: LightGBM.
- Visualização: Matplotlib, Seaborn.
- Serialização: Joblib.

5. Principais Desafios e Soluções Adotadas

- Data Leakage: Identificado e corrigido através da remoção de features problemáticas e revisão da ordem de criação e imputação de features.
- Integração de Dados Externos (INMET): Superados desafios na leitura de formatos e na associação espacial município-estação.
- Qualidade das FWI-like Proxies: Melhorada significativamente com a integração dos dados do INMET.
- Desbalanceamento de Classes: Abordado com SMOTE e scale_pos_weight.
- Tempo de Execução da Otimização: Gerenciado por ajustes na grade de busca e n_iter.

6. Resultados Esperados e Alcançados

- Modelo Final: LightGBM otimizado para F2-score.

- Métricas (com limiar otimizado ~0.140):
    - Recall (Incêndio): 0.639 (63.9%)
    - Precisão (Incêndio): 0.102 (10.2%)
    - Falsos Negativos (Incêndios Perdidos): 761
    - AUC-ROC: 0.791
    - AUC-PR: 0.221
    - F2-Score (Incêndio): 0.361
- Importância das Features: O modelo utiliza uma combinação de municipio_encoded, dados de MapBiomas, condições de seca, FWI-like proxies e dados meteorológicos defasados do INMET.

7. Impacto Esperado e Conclusões

Nossa equipe conseguiu desenvolver um pipeline de dados robusto e um modelo preditivo com capacidade promissora na detecção de incêndios florestais, priorizando o recall. A integração de diversas fontes de dados e a depuração rigorosa foram cruciais.

Embora a precisão para a classe minoritária indique espaço para refinamentos futuros visando reduzir falsos alarmes, o recall de quase 64% representa um avanço importante para sistemas de alerta precoce. O modelo e o dataset servem como uma base sólida para a futura aplicação web e para explorações futuras, incluindo a expansão do histórico de dados para outros anos, o que foi inicialmente limitado por questões de performance e tempo.