Global Solution
Data Science and Statistical Computing

Solução “SmokeSignal”

Integrantes:

RM 554070 - Lucas Garcia

RM 554272 - Enzo Barbeli

RM 554259 - Felipe Santana

Faculdade de Informática e Administração Paulista, 2025

Sumário

Objetivo

Este projeto tem como objetivo prever futuros incêndios florestais com base nas variáveis **causa**, **estado** e **mês do ano**, utilizando dados históricos registrados nos Estados Unidos.

O estudo parte do contexto crescente de queimadas, intensificadas por fatores como mudanças climáticas, aumento das temperaturas globais e longos períodos de seca. Esses incêndios, que podem ser provocados tanto por causas naturais (como raios) quanto humanas (como queimadas descontroladas ou atos criminosos), geram impactos ambientais severos e riscos à saúde pública.

A análise busca identificar padrões temporais e geográficos nos registros de incêndios, investigando as causas mais comuns em diferentes regiões e épocas do ano. Por meio de ferramentas estatísticas e modelos preditivos, pretende-se não apenas entender o comportamento histórico dos incêndios, mas também propor estratégias de prevenção mais eficazes.

A abordagem inclui a aplicação de técnicas como **inferência estatística**, **testes de hipóteses**, **modelos de regressão** e **classificação**, com o objetivo de verificar a relevância estatística dos padrões encontrados e contribuir para ações de monitoramento e controle ambiental.

1. Amostragem

Para garantir a viabilidade computacional da análise e ao mesmo tempo manter a representatividade da população original, foi realizada uma amostragem aleatória simples (AAS) de 500.000 registros a partir de um conjunto total com aproximadamente 2 milhões de observações.

Essa técnica estatística atribui igual probabilidade a todas as unidades da população, sendo, portanto, isenta de vieses sistemáticos. Para assegurar a reprodutibilidade da seleção, foi definida uma semente aleatória fixa (`seed = 11`). A amostra foi extraída sem reposição, evitando duplicações e reforçando a integridade do conjunto amostral.

O processo consistiu em três etapas:

1. Contagem total das linhas do arquivo original (desconsiderando o cabeçalho);
2. Geração dos índices das linhas amostradas via `numpy.random.choice`;
3. Escrita das linhas selecionadas em um novo arquivo `.csv`.

O arquivo final, contendo os 500 mil registros, foi salvo em `../data/amostra_500k.csv`.

---

2. Pré-processamento dos Dados

Seleção de Colunas

A partir do conjunto amostral, foram selecionadas as seguintes colunas, conforme documentação fornecida:

- `FOD_ID`: Identificador único do incêndio.
- `FIRE_NAME`: Nome atribuído ao incêndio.
- `FIRE_YEAR`: Ano de descoberta do incêndio.
- `DISCOVERY_DATE`: Data exata da descoberta.
- `DISCOVERY_DOY`: Dia do ano correspondente à descoberta.
- `NWCG_CAUSE_CLASSIFICATION`: Classificação ampla da causa (ex.: Humana, Natural).
- `NWCG_GENERAL_CAUSE`: Causa geral do incêndio (ex.: Raio, Ato criminoso).
- `CONT_DATE`: Data de controle do incêndio.
- `CONT_DOY`: Dia do ano correspondente ao controle.
- `FIRE_SIZE`: Área afetada (em acres).
- `FIRE_SIZE_CLASS`: Categoria do incêndio segundo o tamanho da área.
- `LATITUDE` e `LONGITUDE`: Coordenadas da localização.
- `STATE`: Sigla do estado norte-americano onde ocorreu o incêndio.

Tratamento de Dados

Tratamento de Valores Ausentes

Foi identificado que as colunas `FIRE_NAME`, `CONT_DATE` e `CONT_DOY` continham valores nulos. Para garantir a integridade das análises, as linhas com ausência de dados nessas variáveis foram removidas do conjunto.

Conversão de Tipos

As colunas `FIRE_YEAR`, `DISCOVERY_DATE` e `CONT_DATE` foram convertidas para o tipo `datetime`. Já as colunas `DISCOVERY_DOY` e `CONT_DOY` foram convertidas para valores numéricos (`int64`), permitindo análises temporais mais precisas.

Remoção de Duplicatas

Após o tratamento, foi verificado que não existiam registros duplicados no conjunto de dados.

---

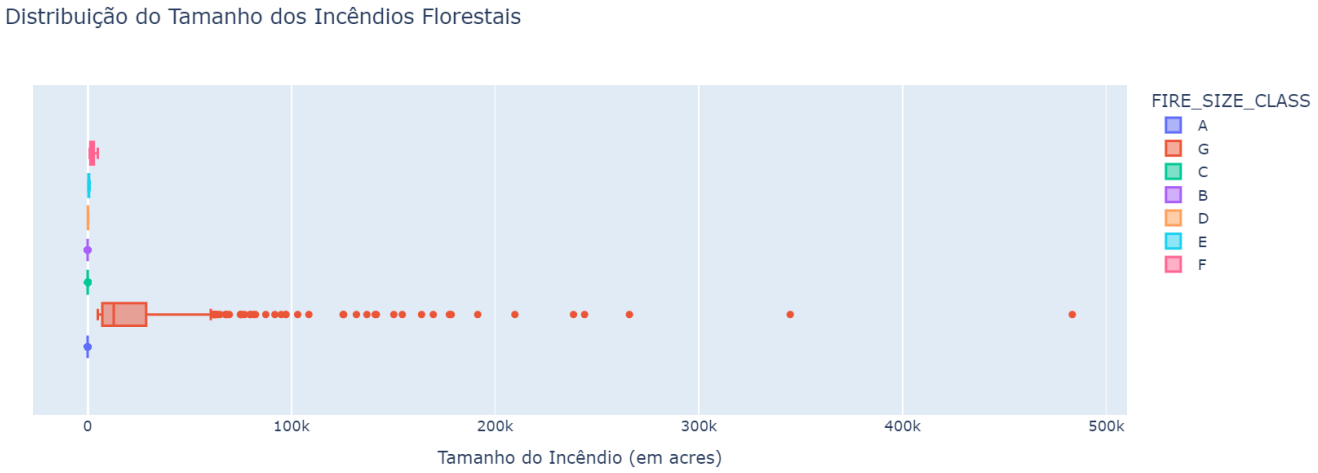
3. Análise Exploratória

Estatísticas Descritivas

A variável `FIRE_SIZE` (área afetada) apresentou grande variabilidade, o que é esperado dado o comportamento errático dos incêndios florestais. Ao agrupar por `NWCG_GENERAL_CAUSE` e `STATE`, observou-se que a distribuição do tamanho dos incêndios varia significativamente entre as causas e entre os estados.

Visualização Geral dos Incêndios

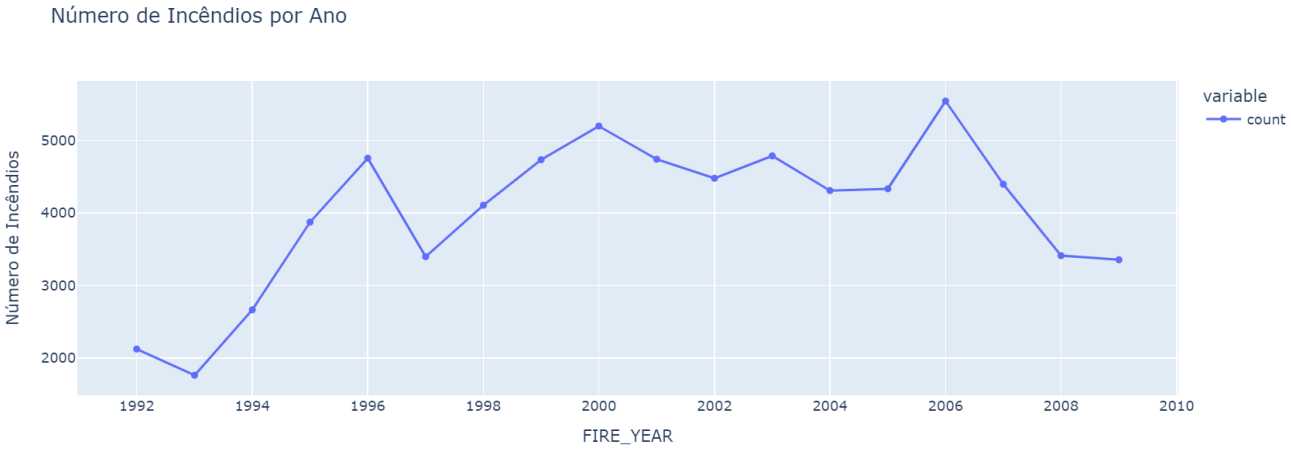
Foi construída uma visualização em caixa (boxplot) para representar a distribuição do tamanho dos incêndios, categorizados pela classe de tamanho (`FIRE_SIZE_CLASS`), demonstrando a predominância de pequenos focos de incêndio, com alguns outliers de grande escala.



Tendência Anual de Incêndios

A partir da contagem do número de incêndios por ano (`FIRE_YEAR`), identificou-se uma tendência de aumento nos registros ao longo do tempo, com variações pontuais que podem

estar associadas a eventos climáticos extremos ou mudanças nas políticas de monitoramento.



Análise Sazonal

Os incêndios foram agrupados por mês, revelando que os meses de verão (maio a setembro) concentram a maior parte dos casos, em consonância com as condições climáticas mais propícias (calor, seca e ventos intensos). Esse padrão sazonal reflete um risco acentuado entre os **meses 5 e 9.**

Gráfico de Barra:

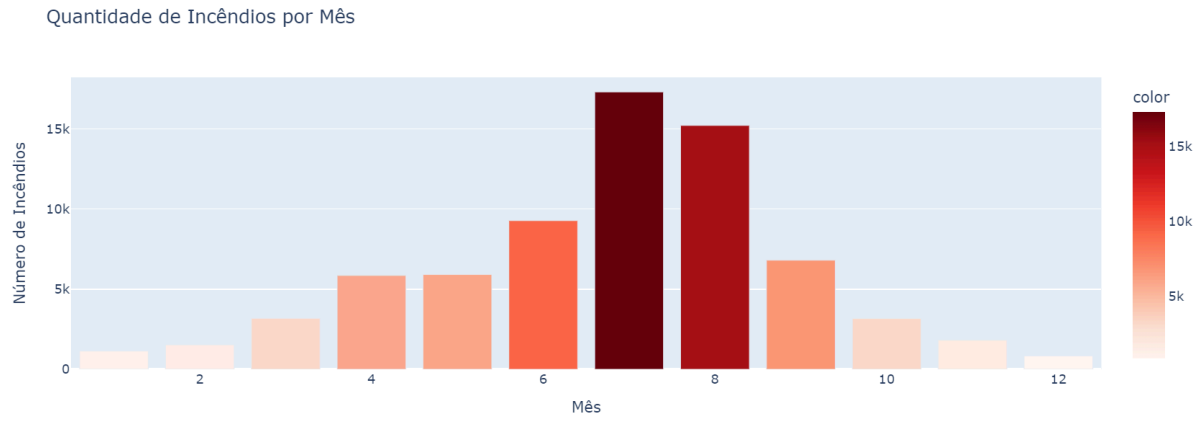
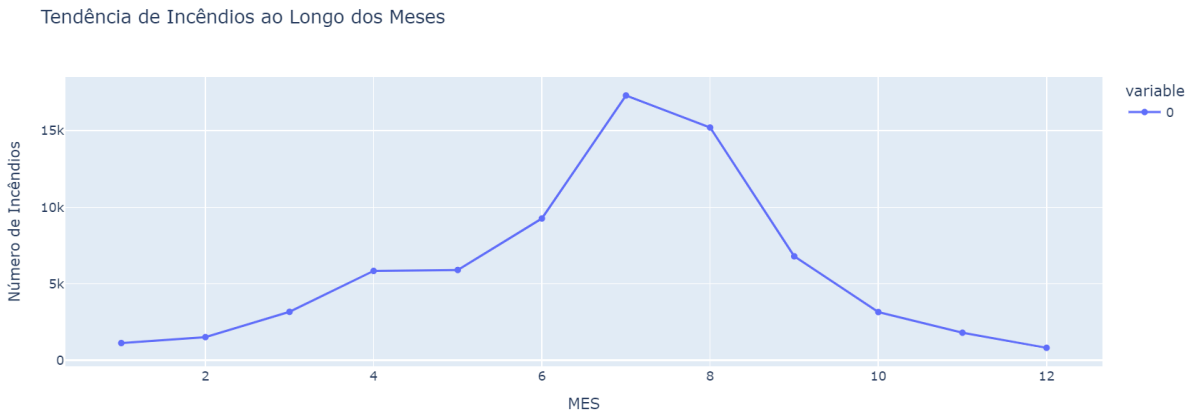
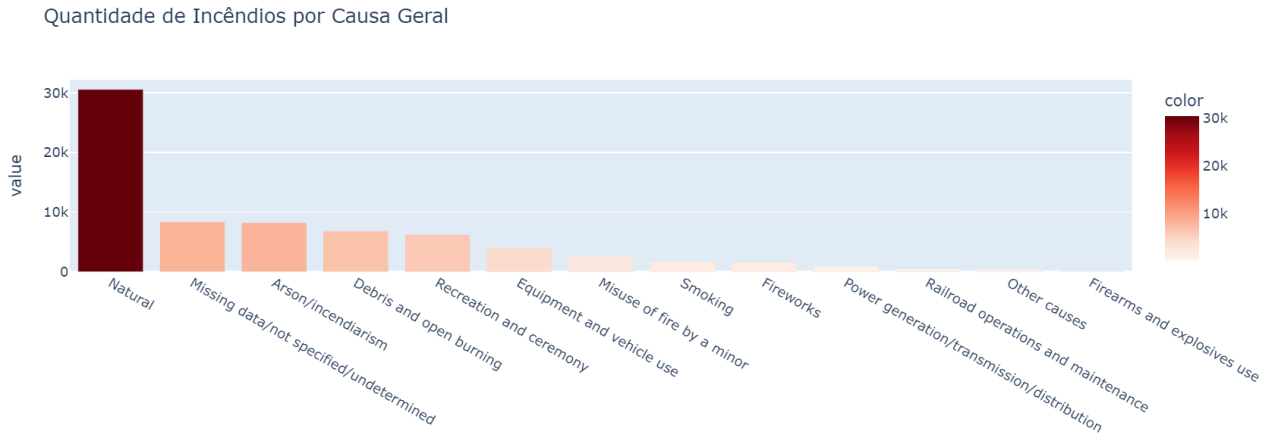


Gráfico de Linha:



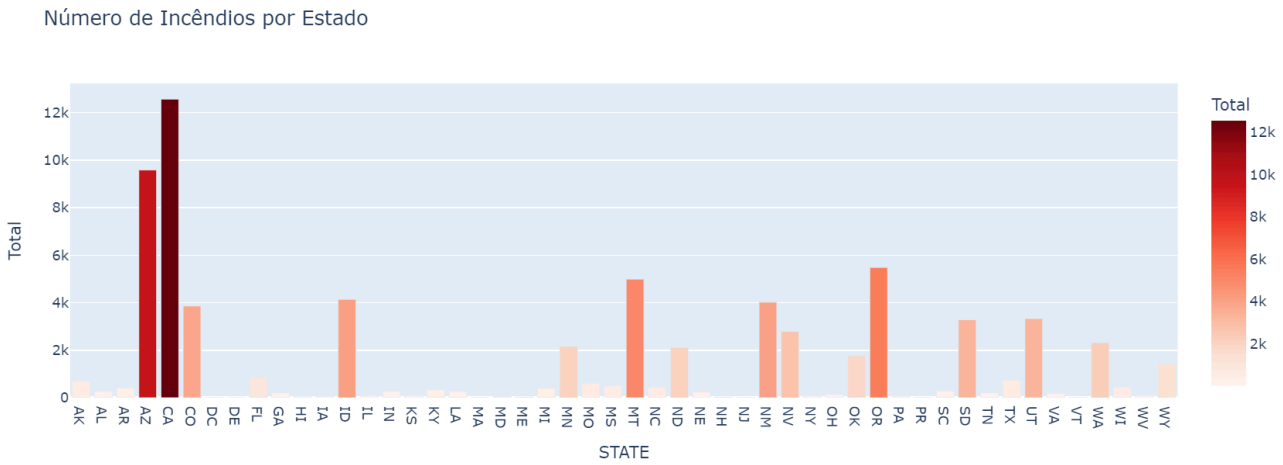
Causas dos Incêndios

A análise da variável **`NWCG_GENERAL_CAUSE`** revelou que a maioria dos incêndios tem origem **natural**, seguida por causas humanas, como queima de resíduos ou uso indevido de fogo. Visualizações de barras reforçaram essa distribuição.



Distribuição por Estado

A **Califórnia** destacou-se como o estado com **maior número** de incêndios registrados, enquanto **Delaware apresentou o menor número**. Ao agrupar a causa mais comum por estado, constatou-se que há variações regionais significativas.



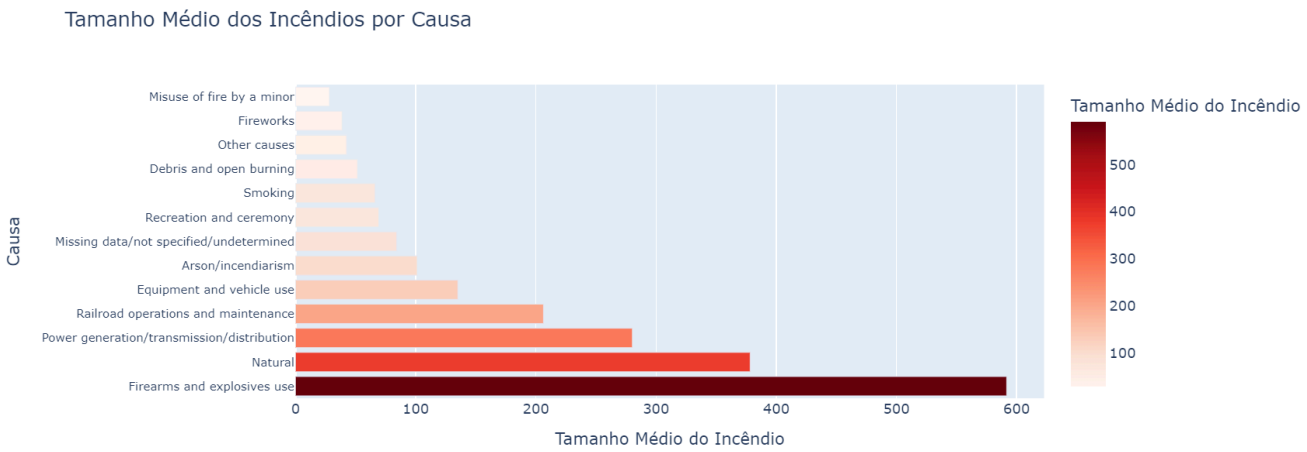
Causa Mais Frequente por Mês

A agregação mensal mostrou que as causas **naturais** dominam os meses com maior incidência de incêndios (meses de verão), evidenciando a forte correlação entre **clima** e **origem** do fogo.

MES	NWCG_GENERAL_CAUSE
1	Arson/incendiarism
2	Arson/incendiarism
3	Arson/incendiarism
4	Debris and open burning
5	Natural
6	Natural
7	Natural
8	Natural
9	Natural
10	Recreation and ceremony
11	Arson/incendiarism
12	Arson/incendiarism

Tamanho Médio dos Incêndios por Causa

Ao calcular a média de **`FIRE_SIZE`** por **`NWCG_GENERAL_CAUSE`**, foi possível observar que causas como "uso de armas de fogo e explosivos" apresentam, em média, as maiores áreas queimadas. Esse tipo de incidente, apesar de menos frequente, tende a ser mais devastador em extensão territorial.



---

4. Análise Inferencial

Amostragem

Foi utilizada **amostragem aleatória simples**, garantindo que cada linha do dataset tivesse a mesma chance de ser selecionada. Essa abordagem é apropriada para análises inferenciais, pois minimiza o viés de seleção e permite estimativas estatísticas confiáveis.

Intervalos de Confiança para a Média da Área Queimada

a) Por Causa do Incêndio

Foram calculados intervalos de confiança de 95% para a média da área queimada (em acres), segmentados por causa. Os resultados mostram:

- **Firearms and explosives use**:
  Média = 591,70 acres | IC95% = (-35,95 ; 1219,35)
  → Intervalo muito amplo e inclui valores negativos, indicando **alta variabilidade** e **baixa precisão**.

- **Natural**:
  Média = 378,22 acres | IC95% = (310,76 ; 445,68)
  → Intervalo bem definido. Incêndios naturais geram, **consistentemente**, áreas queimadas maiores.

- **Power generation/transmission/distribution**:
  Média = 280,24 acres | IC95% = (113,55 ; 446,93)

→ Intervalo largo, mas ainda indica impacto significativo.

- **Equipment and vehicle use**:
  Média = 134,97 acres | IC95% = (51,30 ; 218,64)
  → Média considerável, com intervalo de confiança moderado.

- **Arson/incendiarism**:
  Média = 101,16 acres | IC95% = (45,51 ; 156,80)
  → Causa relevante com intervalo relativamente estreito, mostrando consistência nos dados.

b) Por Estado

Os estados com maiores médias de área queimada incluem:

- **Alasca (AK)**:
  Média = 6788,10 acres | IC95% = (4405,84 ; 9170,36)
  → **Maior média disparada**. Incêndios de larga escala.

- **Nevada (NV)**:
  Média = 665,66 acres | IC95% = (402,12 ; 929,20)
  → Média alta com intervalo preciso.

- **Idaho (ID)**:
  Média = 369,41 acres | IC95% = (225,25 ; 513,58)

- **Wyoming (WY)**:
  Média = 387,45 acres | IC95% = (147,47 ; 627,43)

- **Oregon (OR)** e **Utah (UT)** também têm médias relativamente elevadas, com intervalos moderadamente estreitos.

Observação: Estados como **DC, NH, VT, ME, NY, MA, NJ** apresentam médias muito baixas, o que indica que incêndios nesses locais tendem a ser menores e mais controlados.

Teste de Hipótese: Causas Naturais vs. Causas Humanas

Hipóteses formuladas:

- **H_0 (nula)**: A média da área queimada é a mesma para causas naturais e humanas.

- **H_1 (alternativa)**: As médias são diferentes.

Resultados do teste t (com variâncias desiguais):

- Estatística t = **8,11**
- p-valor = **< 0.00001**

Interpretação:

- A diferença entre as médias é **estatisticamente significativa**.
- **Rejeita-se H_0** com alta confiança.
- Conclusão: Incêndios por causas **naturais queimam áreas significativamente maiores** do que os provocados por causas humanas.

Modelo de Regressão Linear Múltipla

Objetivo:

Estimar a variável dependente `FIRE_SIZE` (tamanho do incêndio) com base em:

- **Causa**
- **Estado**
- **Mês do ano**

Resultados do modelo:

- **$R^2 = 0.027$**
  → Apenas 2,7% da variabilidade da área queimada é explicada pelo modelo.
  → Indica que há muitos outros fatores influenciando os incêndios que não foram incluídos.

- **Significância global do modelo**:
  Apesar do baixo R^2, o modelo é **estatisticamente significativo** (F-statistic com $p < 0.001$), o que significa que é melhor do que um modelo nulo.

Variáveis com significância estatística (p < 0.05):

- **Causas**:
  - `Natural`: coeficiente positivo (~+162 acres), p = 0.007

    - `Firearms and explosives use`: coeficiente negativo (~-919 acres), p = 0.021

    - `Recreation and ceremony`: coeficiente negativo (~-149 acres), p = 0.045

- **Estados**:

    - Muitos estados apresentam **coeficientes negativos significativos**, sugerindo que o estado de referência (omitido no one-hot encoding) tem incêndios muito maiores.

- **Mês do Ano (MES)**:

    - Incluído no modelo como variável contínua. Pode ter impacto, mas possivelmente fraco ou não-linear.

Conclusão do Modelo:

- O modelo é limitado em capacidade preditiva.

- Ainda assim, **confirma algumas relações esperadas**, como o maior impacto de incêndios naturais.

- Sugere que **fatores adicionais** (climáticos, vegetação, políticas públicas, densidade populacional) deveriam ser incluídos para melhor modelagem.

Modelo de Classificação Probabilística — Previsão de Incêndios

Objetivo:

Prever a probabilidade de ocorrência de incêndio (valor entre 0 e 1) com base no estado e no mês.

Etapas resumidas:

1. Agrupamento dos dados: contamos os incêndios por estado e mês, marcando com 1 os casos onde houve ocorrência.

2. Geração de combinações: criamos todas as combinações possíveis de estado e mês, inclusive aquelas sem incêndio.

3. Junção e preenchimento: unimos os dados e preenchemos os casos sem incêndio com 0.

4. Preparação do modelo: aplicamos one-hot encoding nas variáveis categóricas (**STATE** e **MES**), dividimos em **X** e **y**, e treinamos uma regressão logística.

5. Avaliação:

    - ROC AUC: 0.888 — bom desempenho em prever probabilidades.
    - Recall 1.00 para incêndios — o modelo acerta todos os casos positivos.
    - Baixa performance para classe 0 — erra bastante ao prever ausência de incêndio, pois tende a prever que sempre haverá.

6. **Teste de exemplo:** ao prever para o estado da **Califórnia** em **julho**, o modelo retorna uma probabilidade de 91,10% de incêndio.

```
entrada = entrada[X.columns]

proba = model.predict_proba(entrada)[:, 1][0]
print(f"Probabilidade de incêndio: {proba:.2%}")
```

```
Probabilidade de incêndio: 91.10%
```

Conclusão do Modelo:

O modelo é eficiente para identificar onde pode haver incêndio, mas precisa de melhorias para não exagerar nos falsos positivos. Pode ser útil como ferramenta de alerta preventivo.

---

5. Recomendações Finais

Com base na análise exploratória, estatística inferencial e modelagem realizada sobre a amostra de 500 mil registros de incêndios florestais nos Estados Unidos entre 1992 e 2020, é possível formular recomendações relevantes para a prevenção, combate e monitoramento de incêndios florestais.

1. **Identificação de padrões estatisticamente significativos**

- Testes de hipóteses confirmaram que **incêndios de origem humana causam, em média, mais dano ambiental** do que os naturais. Isso reforça a necessidade de **fiscalização ativa e campanhas educativas** sobre práticas de risco, como queimadas agrícolas e descarte impróprio de resíduos.

- A sazonalidade demonstrou ser um fator crítico: **julho e agosto concentram os maiores picos de incêndios**, especialmente em estados com clima mais seco. Isso valida políticas sazonais e **reforço operacional nos meses críticos**.

---

2. **Modelagem preditiva e mapeamento de risco contínuo**

- A **regressão** indicou que **estado, causa e mês do ano** são bons preditores do **tamanho dos incêndios**.
- O **modelo probabilístico** permite estimar com precisão a **chance de ocorrência de incêndios em cada estado e mês**, com saídas interpretáveis como:

  "Em julho na Califórnia, há 91% de chance de ocorrer um incêndio."
  "Em dezembro no Texas, há 3% de chance de incêndio."

- Essa abordagem viabiliza a criação de **mapas de calor mensais probabilísticos**, permitindo que os gestores atuem **antes mesmo de novos focos surgirem**.

---

3. **Políticas públicas baseadas em predição de ocorrência**

- A introdução do modelo probabilístico amplia a capacidade de **antecipação do risco** em regiões críticas, direcionando:
  - Alocação de brigadas preventivas;
  - Posicionamento de aeronaves e torres de vigilância;
  - **Criação de alertas mensais por estado**, apoiados por dados históricos.
- Estados com **altas probabilidades de incêndio recorrente** devem receber **ações educativas localizadas**, especialmente em meses historicamente críticos, alinhando prevenção com conscientização comunitária.

---

4. **Recomendações práticas para órgãos ambientais e gestores públicos**

- **Gerar relatórios mensais com previsões de risco por estado**, cruzando variáveis históricas com o modelo probabilístico, para orientar políticas públicas sazonais.
- **Integrar o modelo preditivo de ocorrência a sistemas meteorológicos e satélites** para atualizar o risco em tempo real, priorizando regiões de alerta máximo.
- **Adotar modelos mistos (ocorrência + gravidade)** no combate a incêndios, estimando tanto a probabilidade de um foco surgir quanto sua potencial gravidade.
- Estabelecer um **painel interativo (dashboard)** acessível a gestores, mostrando os índices de risco por região e mês, permitindo **respostas ágeis e informadas**.

Conclusão

A união de análises estatísticas clássicas com modelagem preditiva baseada em aprendizado de máquina oferece uma base robusta para **políticas públicas de prevenção e resposta adaptativa**. A predição de **probabilidades mensais por estado** representa um avanço no combate proativo aos incêndios, permitindo **decisões baseadas em dados concretos e históricos**, com forte potencial de **salvar vidas, proteger ecossistemas e otimizar recursos operacionais**.