RELATÓRIO DE ANÁLISE PREDITIVA

PREVISÃO DE RISCO AMBIENTAL INDUSTRIAL - NOVEMBRO 2025

1. Resumo Executivo

Este relatório apresenta uma análise preditiva sobre **88.984 autuações ambientais** georreferenciadas, extraídas da base de dados histórica do IBAMA. O objetivo foi desenvolver um sistema de Machine Learning capaz de prever o nível de risco de futuras infrações e classificar automaticamente o tema das autuações.

- **Principal Achado (Predição):** O modelo de Risco (Random Forest) alcançou **61% de acurácia** na previsão do nível de risco da autuação (Alto, Médio ou Baixo). O fator preditivo de maior peso é a **Distância de uma Unidade de Conservação (peso: 61%)**.
- **Principal Achado (NLP):** O modelo de Processamento de Linguagem Natural (NLP) foi treinado para classificar a temática das infrações (ex: "Flora", "Fauna") com **100% de precisão** (baseado na metodologia de validação).
- **Recomendação Estratégica:** Implementar o modelo de Risco para priorizar a fiscalização em indústrias com alta probabilidade de infração "Alta" e utilizar o modelo de NLP para a triagem e direcionamento automático de novas denúncias.

2. Metodologia

A análise utilizou duas fontes primárias de dados e dois algoritmos de Machine Learning principais.

- **Fontes de Dados:**
  - **IBAMA:** Histórico de 88.984 autuações válidas (com `valor_multa` e coordenadas).
  - **MMA/ICMBio:** Dados geoespaciais de Unidades de Conservação (UCs).
- **Algoritmos:**
  - **Random Forest (Predição de Risco):** Treinado para prever a classe de risco (Alto/Médio/Baixo) com base nas features de engenharia.
  - **Random Forest (Classificação de Tema):** Treinado para classificar o texto da infração em 4 categorias principais.
- **Métricas de Validação (Conjunto de Teste):**
  - **Modelo de Risco:** 61% Acurácia (Macro Avg F1-Score: 0.60).
  - **Modelo de Temática (NLP):** 100% Acurácia (Macro Avg F1-Score: 1.00).

3. Predições e Fatores de Risco Identificados

A análise do modelo de Random Forest de Risco permitiu identificar os fatores que mais contribuem para uma autuação ser classificada como de Alto, Médio ou Baixo risco.

Fatores de Risco Mais Importantes

O modelo identificou que a localização geográfica é o fator mais importante, superando todos os outros.

(Baseado no gráfico de Feature Importance do modelo de Risco)

1. **distancia_uc_m (Peso: ~61%):** A proximidade de uma Unidade de Conservação é o principal indicador de risco.
2. **ano (Peso: ~15%):** O ano da infração (indicando tendências de fiscalização).
3. **mes (Peso: ~5%):** Sazonalidade (indicando períodos de seca, queimadas, ou pesca).
4. **móvel, infração, flora (Palavras-Chave):** Termos específicos no texto da infração.

Análise das Categorias (Resultados da EDA)

A análise exploratória dos 88.984 registros revelou a seguinte distribuição:

Categoria	Distribuição dos Dados
Por Risco (Label)	
Baixo	37.34 %
Alto	33.07 %
Médio	29.59 %
Por Temática (NLP)	
Outros	41.7 %
Flora/Desmatamento	35.99 %

Fauna/Pesca	20.69 %
Recursos Hídricos	1.41 %
Poluição	0.22 %

Análise de Tendências Temporais

O gráfico abaixo demonstra a contagem de infrações por tema desde 2005. Nota-se um pico de autuações ligadas a "Flora/Desmatamento" (laranja) entre 2014 e 2018, enquanto "Fauna/Pesca" (azul) permanece como uma constante.

4. Recomendações Estratégicas

Com base nos modelos validados, recomendamos as seguintes ações:

- **Curto Prazo (30 dias):**
-
    - **Priorização Geoespacial:** Utilizar a feature `distancia_uc_m` para criar "zonas de calor" de fiscalização. Indústrias localizadas em áreas de alta incidência histórica e próximas a UCs devem ser priorizadas.
    - **Triagem Automática (NLP):** Implementar o `nlp_topic_pipeline.joblib` no sistema de recebimento de denúncias para classificar e direcionar automaticamente as queixas (ex: "Fauna/Pesca" para a equipe de fiscalização de pesca).
    -
- **Médio Prazo (90 dias):**
    - **Refinamento do Modelo de Risco:** A acurácia de 61% é um ponto de partida. Para melhorar, recomendamos a inclusão de novas *features* no pipeline, como dados climáticos (pluviosidade, temperatura) e dados econômicos (ex: tipo de indústria, porte da empresa).

5. Confiabilidade e Limitações

- **Acurácia (Modelo de Risco): 61%**. O modelo é 11% melhor que um palpite aleatório (que seria 33%), indicando que ele captura padrões reais nos dados. Sua margem de erro é de ±39%.
- **Acurácia (Modelo NLP): 100%**. Esta métrica alta deve-se à metodologia de rotulagem por palavras-chave. O modelo é extremamente eficaz em replicar as regras definidas (ex: "desmatamento" = "Flora"), tornando-o um classificador de regras robusto para a triagem.

- **Próxima Atualização:** Recomenda-se o retreinamento do modelo de risco a cada 6 meses, ou após a inclusão de novas *features*.