Relatório Final: Análise e Predição de Apreensão de Armas no Rio de Janeiro

Autor: Heitor Brunini Araujo Barbosa
Data: 12 de agosto de 2025

1. Resumo da Proposta da Análise

O objetivo inicial deste projeto foi realizar uma análise exploratória sobre o conjunto de dados de apreensão de armas do Instituto de Segurança Pública (ISP) do Rio de Janeiro. A meta era investigar padrões, tendências e distribuições nos registros de apreensões entre os anos de 2003 e 2019, buscando responder a perguntas como: "Qual o tipo de arma mais apreendida?" e "Existem tendências temporais?".

Posteriormente, o projeto evoluiu para um desafio de mineração de dados mais complexo: **construir e avaliar modelos de Machine Learning capazes de prever a quantidade total de armas apreendidas** em um determinado mês e local (Circunscrição de Polícia - CISP). O objetivo final tornou-se, então, não apenas descrever o passado, mas criar uma ferramenta preditiva e comparar a eficácia de diferentes algoritmos para esta tarefa.

2. O que Deu Certo

O projeto foi bem-sucedido em executar um ciclo completo de mineração de dados, desde a exploração até a modelagem preditiva. Os principais sucessos foram:

- **Análise Exploratória Robusta:** Conseguimos analisar o dataset de forma aprofundada, identificando os tipos de armas mais comuns (pistolas e revólveres) e visualizando a correlação entre as variáveis através de um mapa de calor, o que nos deu uma base sólida para a etapa de modelagem.
- **Implementação Completa da Modelagem:** Todos os modelos exigidos foram construídos, treinados e avaliados com sucesso:
    1. **Regressão Linear (Baseline):** Serviu como um excelente ponto de partida para comparação.
    2. **Random Forest (Ensemble):** Demonstrou ser um modelo extremamente eficaz para este problema.
    3. **Rede Neural (MLP Regressor):** Foi implementada com sucesso, incluindo a etapa crucial de normalização dos dados.
- **Resultados Claros e Conclusivos:** A comparação final entre os modelos foi inequívoca. O **Random Forest Regressor se destacou como o modelo de melhor desempenho**, alcançando um coeficiente de determinação (R^2) de **0.88**, o que indica que ele foi capaz de explicar 88% da variabilidade no total de armas

apreendidas. Este resultado superou com folga tanto o modelo baseline quanto a configuração inicial da rede neural, provando sua adequação para dados tabulares complexos como este.

3. O que Deu Errado

O projeto não encontrou erros impeditivos, mas enfrentou limitações inerentes ao escopo e aos dados disponíveis:

- **Análise Causal Limitada:** Não conseguimos determinar as *causas* por trás das flutuações nas apreensões. Por exemplo, não foi possível responder "Por que as apreensões aumentaram em um determinado mês?". Isso ocorreu porque o dataset se limitava aos registros de apreensão, sem variáveis externas (socioeconômicas, operacionais, etc.).
- **Desempenho da Rede Neural:** Embora funcional, a Rede Neural não superou o Random Forest. Sem um processo aprofundado de *hyperparameter tuning* (ajuste fino de parâmetros), que demanda alto custo computacional e tempo, o modelo não atingiu seu potencial máximo. Isso evidencia que modelos mais complexos não são garantia de melhores resultados "prontos para uso".

4. O que Faria diferente?

Se o projeto fosse iniciado hoje, com o conhecimento adquirido, o foco principal seria em **enriquecer o conjunto de dados** antes mesmo de iniciar a modelagem.

1. **Correlação com Outros Datasets:** A principal mudança seria buscar e integrar dados de outras fontes públicas para criar um dataset mais rico e com maior poder explicativo. As fontes poderiam incluir:
    - **Dados de Criminalidade:** Cruzar as apreensões com taxas de homicídios, roubos e outros crimes violentos por região (RISP/AISP). Isso poderia ajudar a verificar se um aumento nas apreensões tem correlação com a diminuição de outros crimes.
    - **Dados Socioeconômicos:** Integrar dados do IBGE, como densidade populacional, renda per capita e taxa de desemprego por município, para investigar a relação entre o contexto social e a violência armada.
    - **Dados de Operações Policiais:** Se disponíveis, dados sobre grandes operações policiais poderiam explicar picos anormais de apreensões.
2. **Engenharia de Features Mais Sofisticada:** Com um dataset mais rico, criaríamos variáveis mais inteligentes, como "taxa de apreensão por habitante" ou variáveis de sazonalidade (feriados, períodos de férias), que poderiam melhorar a acurácia dos modelos.

5. Aprendizados sobre Mineração de Dados

Este projeto proporcionou aprendizados práticos fundamentais sobre o processo de mineração de dados:

- **A Análise Exploratória (EDA) Governa o Projeto:** Entendemos que a EDA não é apenas uma etapa inicial, mas a fundação sobre a qual todo o projeto se apoia. É nela que entendemos as limitações dos dados e formulamos as hipóteses que guiarão a modelagem.
- **Pré-processamento é 80% do Trabalho:** Ficou claro que os modelos não funcionam magicamente. A preparação dos dados, especialmente a transformação de variáveis categóricas (OneHotEncoder) e a normalização de dados numéricos (Standard Scaler), é essencial e impacta diretamente o desempenho final.
- **Não Existe "Bala de Prata":** O modelo mais complexo (Rede Neural) não foi o melhor. Aprendemos que é crucial testar diferentes abordagens e que modelos como o Random Forest frequentemente oferecem um excelente equilíbrio entre performance, robustez e facilidade de implementação para problemas com dados estruturados.
- **O Poder dos Pipelines:** A utilização da classe Pipeline da biblioteca scikit-learn se mostrou uma prática indispensável. Ela não apenas organiza o código, mas também garante que o pré-processamento seja aplicado de forma correta e sem vazamento de dados (*data leakage*) entre os conjuntos de treino e teste, tornando o processo mais robusto e profissional.