UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CIÊNCIA DA COMPUTAÇÃO

ARTUR PIGARI PRATA

PROJETO FINAL
Relatório Técnico

Campo Mourão
2025

Sumário

1. Principais descobertas

Este estudo analisou a letalidade dos incidentes de disparo registrados em Nova York entre 2006 e 2024, utilizando o NYPD Shooting Incident Data (Historic). A combinação de análise estatística, consultas SQL e exploração visual mostrou que, embora a taxa geral de letalidade permaneça relativamente estável ao longo dos anos, ela varia de forma importante quando observada em maior detalhe. O horário do dia revelou forte associação com o desfecho, com maior proporção de mortes pela manhã. No eixo geográfico, boroughs apresentam diferenças pequenas e não significativas, mas precincts exibem contrastes pronunciados, indicando padrões locais específicos. Entre as variáveis demográficas, a idade da vítima se destacou como o fator mais associado à letalidade, além de efeitos estatisticamente significativos relacionados à raça da vítima, raça do agressor e sexo do agressor.

A modelagem supervisionada confirmou a complexidade do fenômeno. Modelos baselines, Regressão Logística, Random Forest e XGBoost, tiveram baixo desempenho inicial na detecção de casos letais devido ao forte desbalanceamento da variável-alvo. Após o uso de undersampling e tuning, o F1-score da classe letal aumentou significativamente, alcançando aproximadamente 0.38 -- 0.39, com recall acima de 0.60. Esses resultados mostram que, apesar das limitações do dataset, é possível capturar padrões relevantes de letalidade por meio de algoritmos supervisionados. Em síntese, tanto a análise exploratória quanto a modelagem indicam que a letalidade de tiroteios em Nova York é influenciada por fatores temporais, territoriais e demográficos, reforçando o potencial de abordagens baseadas em dados para apoiar ações de prevenção, planejamento policial e formulação de políticas públicas.

2. Problema, perguntas de pesquisa e hipóteses

A violência armada permanece como um desafio relevante nas grandes metrópoles e, em Nova York, mesmo com a redução geral da criminalidade, os incidentes com disparos de arma de fogo seguem representando um risco significativo à segurança pública. O dataset NYPD Shooting Incident Data (Historic), que reúne registros de tiroteios entre 2006 e 2024, oferece informações temporais, espaciais e demográficas, incluindo dados de vítimas, agressores e o desfecho de

cada caso via *STATISTICAL_MURDER_FLAG*, permitindo investigar fatores associados à letalidade. Com base nesse material, o estudo busca compreender como características temporais, geográficas e demográficas influenciam a probabilidade de um tiro resultar em homicídio, e avaliar se tais informações são suficientes para estimar automaticamente esse risco no momento do incidente. Para isso, foram definidas perguntas de pesquisa e hipóteses que guiam toda a análise, abrangendo exploração de dados, testes estatísticos para verificar H1 e H2, e a construção de modelos supervisionados voltados à avaliação de H3, com foco especial no desempenho de algoritmos baseados em árvores de decisão na previsão da letalidade.

2.1. Perguntas de pesquisa

- Quais fatores estão associados à letalidade dos tiroteios na cidade de Nova York?
- É possível estimar, a partir das informações disponíveis no momento do crime, a chance de um tiroteio resultar em homicídio?

2.2. Hipóteses

- **H1:** A probabilidade de um tiroteio resultar em morte varia significativamente de acordo com o horário, bairro e distrito policial em que ocorre.
- **H2:** O perfil demográfico da vítima e do agressor (idade, sexo e raça) está associado à letalidade dos tiroteios.
- **H3:** Modelos de aprendizado de máquina baseados em árvores de decisão apresentam desempenho superior aos modelos lineares na previsão da letalidade de incidentes com disparos.

3. Metodologia e limitações

3.1. Fonte e estrutura dos dados

O estudo utiliza o dataset **NYPD Shooting Incident Data (Historic)**, contendo registros de tiroteios ocorridos entre 2006 e 2024 na cidade de Nova York. Cada ocorrência inclui informações temporais (ano, hora), espaciais (borough, precinct, coordenadas X/Y), dados demográficos de vítimas e agressores e o

desfecho do caso por meio da variável *STATISTICAL_MURDER_FLAG*. Após limpeza e filtragem, 28.491 incidentes válidos foram utilizados nas análises, de um total de 29.744.

3.2. Preparação dos dados

A preparação envolveu:

- Padronização e limpeza de categorias textuais e eliminação de registros inconsistentes.
- Criação de variáveis derivadas, como período do dia e grupos etários.
- One-hot encoding das variáveis categóricas para modelagem.
- Tratamento do desbalanceamento da variável-alvo (~19% de casos letais) com undersampling e, posteriormente, ajuste de threshold no XGBoost.
- Separação dos dados em treino e teste, mantendo representatividade das classes.

3.3. Estratégia analítica

A metodologia combinou três abordagens principais:

A. Análise Exploratória (EDA)

Incluiu inspeções de distribuição temporal, espacial e demográfica, bem como visualizações da taxa de letalidade por horário, borough, precinct, faixa etária e perfil do agressor. Matrizes de correlação foram utilizadas para avaliar a relação entre variáveis numéricas.

B. Testes Estatísticos de Associação

Para verificar H1 e H2, foram aplicados testes de Qui-quadrado, acompanhados do V de Cramér para medir força da associação. As análises avaliaram relações entre letalidade e fatores como período do dia, localização, idade, sexo e raça de vítimas e agressores.

C. Modelagem Supervisionada

Para responder à segunda Pergunta de Pesquisa, foram treinados modelos de classificação binária: Regressão Logística, Random Forest e XGBoost, em versões baseline e tunadas. A avaliação utilizou F1-score da classe letal, recall, ROC-AUC e PR-AUC, com tuning via GridSearchCV.

3.4. Limitações do estudo

O estudo apresenta algumas limitações:

- Forte desbalanceamento da variável-alvo, que reduz desempenho preditivo.
- Presença de categorias “UNKNOWN” em atributos do agressor.
- Ausência de informações contextuais importantes (ex.: tipo de arma, motivação, distância do disparo).
- Correlações numéricas baixas, dificultando modelos lineares.
- Os resultados indicam associação, não causalidade.

4. Resultados das análises

As análises exploratórias revelaram padrões claros relacionados à letalidade de tiroteios em Nova York. No eixo temporal, observou-se que a taxa de letalidade permaneceu relativamente estável ao longo do período analisado, variando entre 17% e 22%. Entretanto, ao analisar a distribuição por horário, identificou-se um comportamento distinto: incidentes ocorridos pela manhã apresentaram proporção significativamente maior de mortes em comparação aos demais períodos do dia. A análise por hora do dia confirmou um pico expressivo entre 6h e 8h, sugerindo que fatores contextuais relacionados ao horário podem influenciar o desfecho.

No aspecto espacial, as diferenças entre boroughs (Bronx, Brooklyn, Queens, Manhattan e Staten Island) foram pequenas e não estatisticamente significativas, indicando que a divisão administrativa não explica variações na letalidade. Entretanto, quando a análise é refinada para o nível de precinct, surgem contrastes marcantes: alguns precincts apresentam taxas superiores a 35%, enquanto outros se mantêm próximos de 25%. Isso indica que microterritórios possuem dinâmicas próprias de violência e que a granularidade espacial é determinante para entender padrões locais.

A análise demográfica apontou que a idade da vítima é um dos fatores mais associados à letalidade: vítimas com 65 anos ou mais apresentam taxas superiores a 30%, enquanto menores de 18 anos ficam abaixo de 15%. A raça da vítima e a raça do agressor também apresentaram diferenças significativas, embora com tamanhos de efeito pequenos. Em relação ao agressor, o sexo mostrou associação relevante: incidentes com agressoras mulheres apresentaram letalidade maior que aqueles envolvendo homens.

Os testes estatísticos confirmaram esses achados. O Qui-quadrado indicou associações significativas entre letalidade e: período do dia (H1-a), precinct (H1-c), idade da vítima (H2-a), raça da vítima (H2-c), sexo do agressor (H2-d) e raça do agressor (H2-e). As associações com borough (H1-b) e sexo da vítima (H2-b) não foram estatisticamente significativas. Os valores de V de Cramér foram baixos, sugerindo efeitos pequenos, mas consistentes.

Na modelagem supervisionada, os modelos baseline (Regressão Logística, Random Forest e XGBoost) demonstraram dificuldade inicial em identificar casos letais devido ao desbalanceamento da variável-alvo. Após a aplicação de undersampling e tuning, os modelos apresentaram avanços expressivos: o F1-score da classe letal aumentou de aproximadamente 0.10 -- 0.15 (baselines) para 0.38 -- 0.39, com recall acima de 0.60 nos modelos tunados. Entre eles, o XGBoost tunado (com ajuste de threshold) apresentou o melhor equilíbrio entre F1, recall, ROC-AUC e PR-AUC, tornando-se o modelo mais eficaz para identificar casos letais. Embora o desempenho ainda seja insuficiente para uso operacional, os resultados mostram que a letalidade não é aleatória e que padrões relevantes podem ser aprendidos por modelos supervisionados.

De forma geral, as análises mostram que a letalidade é influenciada por fatores temporais, características locais de precincts e atributos demográficos de vítimas e agressores, confirmando parcialmente H1 e amplamente H2, além de demonstrar que modelos baseados em árvores de decisão (como previsto em H3) apresentam melhor desempenho após tuning e tratamento do desbalanceamento.

5. Discussão dos resultados

Os resultados obtidos apontam para um fenômeno de letalidade multifatorial, no qual elementos temporais, territoriais e demográficos têm participação relevante, embora com efeitos geralmente modestos. A análise estatística mostrou que o momento do incidente (especialmente o período do dia) apresenta associação significativa com a letalidade, reforçando parcialmente H1. A ausência de significância no nível dos boroughs indica que grandes regiões da cidade mascaram variações internas; por outro lado, a forte associação identificada no nível dos precincts confirma que as dinâmicas de violência são espacialmente concentradas

em microterritórios, onde condições locais, infraestrutura urbana e padrões de policiamento provavelmente desempenham papel determinante.

No que diz respeito ao perfil das pessoas envolvidas, os achados dão suporte robusto a H2. A idade da vítima foi o fator mais claramente associado ao aumento da letalidade, confirmando padrões amplamente observados na literatura, indivíduos mais velhos tendem a apresentar maior vulnerabilidade fisiológica. Raça da vítima, sexo do agressor e raça do agressor também apresentaram associações estatisticamente significativas, mas com tamanhos de efeito pequenos, sugerindo que influenciam o risco de letalidade de maneira real, porém limitada. A não significância do sexo da vítima indica que sua contribuição é pequena ou inexistente quando comparada às demais variáveis demográficas.

Na modelagem supervisionada, os resultados iniciais mostraram que o desbalanceamento do dataset impõe barreiras substanciais ao desempenho dos modelos, afetando principalmente o recall da classe letal. Após a aplicação de undersampling e tuning, modelos lineares e baseados em árvores, especialmente o Random Forest e o XGBoost, alcançaram desempenho muito superior ao baseline, confirmando parcialmente H3. Ainda que o F1-score tenha se estabilizado próximo de 0.38 -- 0.39 para os melhores modelos, esses valores demonstram que há padrões estatísticos relevantes capazes de sustentar uma estimação automatizada do risco de morte. No entanto, os resultados também revelam os limites práticos dessa abordagem: a precisão continua baixa, indicando que os modelos produzem muitos falsos positivos, enquanto o recall permanece insuficiente para aplicações operacionais sensíveis, como decisões táticas de policiamento.

Em conjunto, os achados reforçam que a letalidade não é aleatória; ela emerge de uma combinação de fatores contextuais, territoriais e individuais. Entretanto, os efeitos são relativamente fracos quando analisados isoladamente, o que explica tanto os baixos valores de V de Cramér nos testes estatísticos quanto o desempenho moderado da modelagem supervisionada. Isso sugere que a letalidade de incidentes com disparo envolve interações complexas entre variáveis, possivelmente influenciadas por elementos não capturados no dataset, como distância até hospitais, tipo da arma, número de disparos, tempo de resposta policial ou comportamentos situacionais. Assim, embora os modelos construídos indiquem tendência e estrutura, eles ainda estão longe de capturar a totalidade do processo real.

6. Recomendações práticas

Os resultados deste estudo sugerem que políticas públicas mais eficazes no enfrentamento da letalidade de tiroteios devem priorizar intervenções direcionadas a contextos específicos, em vez de abordagens amplas aplicadas de maneira homogênea na cidade. A significativa variação entre precincts indica que microterritórios constituem unidades-chave para a prevenção: esforços como reforço de patrulhamento orientado por dados, programas de intervenção comunitária e ações coordenadas com serviços de saúde e assistência social podem ser alocados de forma mais precisa nesses locais de maior risco. Tais estratégias podem reduzir tanto a incidência quanto a gravidade dos incidentes.

No âmbito demográfico, a maior vulnerabilidade de vítimas mais velhas e os padrões associados ao perfil racial e ao sexo do agressor sugerem a necessidade de políticas complementares de proteção e mediação de conflitos em grupos populacionais específicos. Campanhas comunitárias, iniciativas de prevenção à violência familiar e programas voltados para homens jovens em contextos de maior risco podem contribuir para a redução indireta da letalidade. Por fim, embora os modelos preditivos não estejam prontos para uso operacional, eles demonstram potencial como ferramentas auxiliares. Futuramente, modelos mais robustos, treinados com variáveis adicionais e melhor balanceamento, podem apoiar sistemas de alerta situacional, priorização de recursos e monitoramento de áreas críticas.

7. Trabalhos futuros

Os resultados deste estudo evidenciam que ainda há amplo espaço para aprofundamento na análise da letalidade de tiroteios em Nova York. Uma primeira direção promissora envolve o enriquecimento do dataset com variáveis externas, como histórico criminal, indicadores socioeconômicos de microáreas, presença de gangues, dados meteorológicos ou características do armamento utilizado. Tais informações podem aumentar substancialmente a capacidade explicativa dos modelos e revelar dimensões estruturais da violência que não estão capturadas pelos registros originais do NYPD.

Outra frente importante é o desenvolvimento de técnicas de modelagem mais robustas. Abordagens como SMOTE, ensembles especializados em classes raras, modelos probabilísticos e arquiteturas baseadas em aprendizado profundo podem

ser exploradas para lidar melhor com o desbalanceamento severo da variável-alvo. Além disso, estudos futuros podem avaliar o uso dos modelos preditivos em cenários reais, como apoio à priorização de recursos ou monitoramento de áreas críticas, sempre com atenção às implicações éticas, aos riscos de viés e à necessidade de validação transparente. Por fim, análises longitudinais e comparativas, seja entre diferentes períodos, seja entre cidades, podem contribuir para compreender como padrões de letalidade evoluem no tempo e quais fatores permanecem estáveis ou se transformam.

8. Referências

CITY OF NEW YORK. NYPD Shooting Incident Data (Historic). Disponível em: https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8. Acesso em: 15 out. 2025.

CHEN, Tianqi; HE, Tong; BENESTY, Michael; KHOTILOVICH, Vadim; TANG, Yuan; CHO, Hyunsu; CHEN, Kailong; MITCHELL, Rory; CANO, Ignacio; ZHOU, Tianyi; LI, Mu; XIE, Junyuan; LIN, Min; GENG, Yifeng; LI, Yutian; YUAN, Jiaming; CORTES, David; et al. xgboost: Extreme Gradient Boosting — Python package. versão 3.2.0.0, 2025. Disponível em: https://xgboost.readthedocs.io/en/stable/python/python_intro.html. Acesso em: 26 nov. 2025.

SCIKIT-LEARN developers. *RandomForestClassifier — scikit-learn 1.7.2 documentation*. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. Acesso em: 25 nov. 2025.

SCIKIT-LEARN developers. *LogisticRegression — scikit-learn documentation*. Disponível em: https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression. Acesso em: 25 nov. 2025.