Leonardo Naime Lima
Pablo da Silva Sena

Relatório Técnico

Relatório técnico apresentado para a entrega do projeto final disponibilizado pela disciplina de Ciência de Dados da UTFPR-CM.

Universidade Tecnológica Federal do Paraná - UTFPR

Departamento Acadêmico de Computação - DACOM

Bacharelado em Ciência da Computação - BCC

Campo Mourão - PR

2025

Sumário

1 Resumo

Este trabalho analisou o conjunto de dados *NYPD Arrest Data (Year to Date)*, contendo registros de prisões na cidade de Nova Iorque. O estudo focou na limpeza de dados, análise exploratória via SQL e aplicação de modelos de *Machine Learning* para responder questões sobre perfis demográficos e tipologia criminal. Foram testadas hipóteses sobre a incidência criminal em jovens (18-24 anos), padrões raciais e sazonalidade. A modelagem preditiva utilizou algoritmos como *Random Forest* e *Gradient Boosting*. Um desafio crítico identificado foi a ausência de registro de idade em mais de 50% dos dados, o que introduziu um viés sistemático (MNAR) analisado em profundidade. O melhor modelo para prever a faixa etária jovem obteve acurácia de 86,78%, enquanto a classificação do tipo de crime atingiu 54,99%.

2 Definição do Problema

O dataset escolhido foi o *NYPD Arrest Data*, disponibilizado pelo portal de dados abertos de Nova Iorque. Ele contém informações sobre prisões efetuadas pela polícia de NY, incluindo data, local (distrito/precinct), tipo de crime e dados demográficos do suspeito. O tamanho da amostra analisada foi de 100.000 registros.

Os dados cobrem os 5 distritos (*Boroughs*): Bronx, Brooklyn, Manhattan, Queens e Staten Island.

Foram formuladas as seguintes perguntas de pesquisa e hipóteses:

- **P1:** Pessoas mais jovens (18-24 anos) cometem mais crimes proporcionalmente?
- **P2:** A raça da pessoa influencia no tipo de crime (grave/leve) cometido?
- **H1:** Haverá uma concentração maior de crimes entre pessoas mais jovens em comparação com faixas etárias mais altas.

- **H2:** Grupos demográficos específicos (ex: Asiáticos) apresentam taxas de prisão significativamente menores.
- **H3:** Existe uma sazonalidade positiva, com mais crimes ocorrendo no verão do que no inverno.

3 Metodologia

3.1 Limpeza e Padronização de Dados

Utilizamos estratégias robustas para garantir a qualidade dos dados:

- **Dados Faltantes:** Identificamos 2 valores nulos na coluna `law_cat_cd` e os tratamos como "UNKNOWN". O problema mais grave foi a coluna `age_group`, com mais de 50% de dados nulos (tratado na seção 2.4).
- **Inconsistências:** Identificamos que a delegacia (*Precinct*) 114 estava associada a múltiplos distritos. Corrigimos para associá-la exclusivamente ao Queens (Q).
- **Outliers:** O método IQR não detectou outliers significativos nos códigos de delegacia.
- **Padronização:** Conversão de datas, extração de componentes temporais (ano, mês, dia, hora) e conversão de coordenadas geográficas.

3.2 Análise Exploratória via SQL (DuckDB)

Utilizamos consultas SQL avançadas para extrair insights:

1. **Tendência:** Aplicação de média móvel de 7 dias para suavizar a curva de prisões diárias.
2. **Ranking:** Identificação dos 3 principais crimes por distrito.
3. **Concentração (Pareto):** Verificamos quais delegacias concentram 50% das prisões.

4. **Padrão Temporal:** Análise cruzada entre hora do dia e dia da semana para crimes de Agressão Grave (*Felony Assault*).

3.3 Aplicação das Hipóteses

3.3.1 H1: Faixa Etária e Criminalidade

A análise dos dados **com idade conhecida** mostrou que a faixa 25-44 anos é a mais predominante (59,27%), seguida pela 45-64 anos. A faixa jovem (18-24) representou cerca de 13,27%. *Observação:* Este resultado é fortemente impactado pelos 57% de dados sem idade registrada.

3.3.2 H2: Criminalidade por Etnia

A hipótese H2 foi analisada comparando os totais absolutos. Asiáticos representaram aproximadamente 5,91% das prisões, um número significativamente menor que grupos como Black (46,56%) e White Hispanic (26,00%).

3.3.3 H3: Sazonalidade (Verão vs. Inverno)

A consulta SQL de sazonalidade confirmou a hipótese H3. Observou-se um pico de atividades criminais nos meses de verão (Junho-Agosto) em comparação com o inverno (Dezembro-Fevereiro), corroborando a teoria de que temperaturas mais altas correlacionam com maior incidência de crimes de rua.

3.4 Modelagem com Aprendizado de Máquina

3.4.1 O Problema dos Dados Faltantes (Age Group)

A análise detalhada revelou que a falta de dados na coluna de idade **não é aleatória** (Teste Chi-quadrado: $p < 0.05$). A taxa de dados faltantes varia drasticamente por distrito (ex: Bronx 65% vs Staten Island 33%). Adotamos uma **Abordagem Híbrida**:

- Para modelagem preditiva, removemos registros sem idade para evitar ruído (dataset reduzido para 42.000 linhas).

- Para análise exploratória, mantivemos a categoria 'UNKNOWN' para visualizar o viés.

3.4.2 Pergunta 1: Previsão de Jovens (18-24 anos)

O objetivo foi prever se um detido pertence à faixa etária de 18-24 anos com base em raça, sexo, distrito e gravidade do crime. Foram testados: Regressão Logística, Random Forest, Gradient Boosting e MLP.

Tabela 1 – Comparação de Modelos - Previsão Faixa Etária (18-24)

Modelo	**Acurácia**
Random Forest (Otimizado)	**0.8678**
Random Forest (Baseline)	0.8674
Logistic Regression	0.8673
Gradient Boosting	0.8673
Neural Network	0.8673

O modelo **Random Forest Otimizado** foi o vencedor. A *Feature Importance* indicou que a **Raça** e o **Distrito** foram as variáveis mais determinantes para o modelo.

3.4.3 Pergunta 2: Previsão do Tipo de Crime

O objetivo foi prever a categoria da lei (*Felony, Misdemeanor, Violation*) com base nas características do suspeito.

Tabela 2 – Comparação de Modelos - Previsão Tipo de Crime

Modelo	**Acurácia**
Gradient Boosting (Otimizado)	**0.5499**
Random Forest	0.5464
Logistic Regression	0.5452
Gradient Boosting (Baseline)	0.5401

A acurácia próxima de 55% indica que as características demográficas sozinhas **não são fortes preditoras** do tipo de crime, sugerindo que fatores situacionais ou econômicos (não presentes no dataset) são mais relevantes.

4 Discussão dos Resultados

A análise revelou um grave problema de qualidade nos dados de idade, classificado como *Missing Not At Random (MNAR)*. Isso significa que a ausência da idade está correlacionada com o local da prisão, o que pode introduzir viés nas conclusões sobre a criminalidade juvenil.

Apesar disso, os modelos conseguiram identificar padrões. O *Random Forest* mostrou-se robusto para dados categóricos. A sazonalidade do crime foi confirmada, e a distribuição geográfica mostrou concentrações específicas (Lei de Pareto) em certas delegacias.

5 Recomendações Práticas

Com base nos achados:

- **Melhoria na Coleta de Dados:** É imperativo padronizar o registro de idade em todos os distritos, especialmente no Bronx e Brooklyn, onde a perda de dados supera 60%.
- **Alocação de Recursos:** O policiamento deve considerar a sazonalidade (reforço no verão) e focar nas delegacias identificadas como "Hotspots"na análise de Pareto.

6 Trabalhos Futuros

Sugerimos cruzar estes dados com o Censo Demográfico para calcular taxas de crime *per capita*, eliminando o viés populacional. Também seria valioso aplicar técnicas de imputação avançada (como KNN Imputer) para tentar recuperar parte da informação de idade perdida, validando se o padrão MNAR se mantém.

7 Referências Bibliográficas

1. NYC Open Data - NYPD Arrest Data (Year to Date).

2. Documentação Scikit-Learn e Pandas.