

# Análise de Crimes em Chicago durante os últimos seis anos

José Alberto Martins Fernandes

Mestrado em Ciência de Computadores

Faculdade de Ciências da Universidade do Porto

Porto, Portugal

up201707227

**Abstract**—Este relatório explora os padrões de crimes em Chicago desde 2017 até 2023, e aplica quatro algoritmos de *machine learning* para a previsão de crimes. Através da análise exploratória de dados, são identificadas *trends* temporais e espaciais. Os modelos preditivos, incluindo Random Forest, Support Vector Machines, Naive Bayes e Gradient Boosting, demonstram uma precisão promissora. Eles auxiliam na alocação de recursos, prevenção pró-ativa de crimes e identificação de fatores que influenciam atividades criminosas. Esta pesquisa contribui para a compreensão da dinâmica dos crimes e apoia estratégias de aplicação da lei baseadas em dados.

## I. INTRODUÇÃO

A análise e previsão de crimes desempenham um papel crucial na melhoria da segurança pública e na adoção de estratégias pró-ativas de aplicação da lei. Com a disponibilidade cada vez maior de *datasets* em larga escala e o avanço dos algoritmos de *machine learning*, há um interesse crescente em utilizar a análise de dados automática para previsão de crimes.

Este estudo concentra-se na análise exploratória de um conjunto abrangente de dados sobre crimes em Chicago de 2017 a 2023, com o objetivo de identificar padrões temporais e espaciais nos dados. Através da análise exploratória, procura-se obter informações sobre a natureza dos crimes, as localizações mais frequentes e os períodos de maior incidência de crimes. Além disso, o estudo aplica quatro modelos amplamente utilizados de *machine learning* para a previsão de crimes.

## II. ANÁLISE EXPLORATÓRIA

### A. Análise Inicial dos Dados Fornecidos

Antes de efetuar quaisquer modificações aos dados fornecidos, procedi à análise preliminar do conjunto de dados base. Ao examinar o *dataset*, encontrou-se colunas de identificação do caso, identificação do crime, localização e registos temporais. Com base nessa análise, foram obtidas as seguintes conclusões:

- Para este estudo, iremos remover a coluna *ID*, uma vez que se trata de um identificador único que não influencia os crimes analisados. Da mesma forma, a coluna *Year* pode ser removida, uma vez que essa informação já está contida na coluna *Date*.
- Na análise exploratória, podemos manter as múltiplas colunas de localização, no entanto, para a aplicação dos algoritmos de aprendizado de máquina, utilizaremos apenas a coluna *Block*, por ser a mais específica.

- É necessário realizar uma limpeza às colunas. As colunas *Case Number* e *Primary Type* apresentam casos com formatos inconsistentes, e as colunas *Date* e *Updated On* estão em formatos que, embora estejam corretos, são diferentes.
- Também é necessário efetuar uma limpeza às linhas. Embora alguns algoritmos de *machine learning* possam lidar com a presença de "não aplicáveis" (NA's), neste caso optou-se por desconsiderá-los. Dos um milhão e meio de casos, apenas trinta mil contêm NA's. Portanto, decidiu-se dar preferência aos casos completos.

### B. Análise Aprofundada dos Dados Fornecidos

No começo do projeto, o código fornecido já nos fornecia a primeira análise visual do *dataset*. Este código gerava um gráfico com a quantidade de crimes mensais, representado aqui:

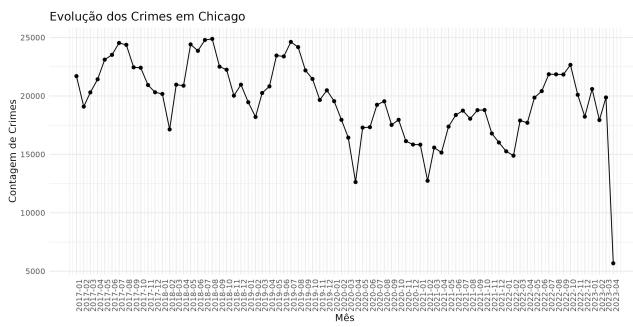


Fig. 1. Evolução dos Crimes em Chicago.

Podemos verificar que Abril de 2023 é um *outlier* neste estudo pois não contém casos semelhantes ao padrão demonstrado. Logo, na situação de estudos mensais deverá ser desconsiderado.

Após esta primeira análise visual, decidi construir um mapa com as coordenadas fornecidas por cada caso com o objetivo de verificar as zonas mais comuns na prática de crimes (Fig. 2). Apesar de ter presente os diferentes anos, fica demasiado complicado ver com detalhe devido à enorme quantidade de casos. O resultado foi um mapa abstrato de Chicago, sendo que não ajudou na pesquisa pois detalhou basicamente o mapa todo sendo imperceptível a diferença de anos. Decidi então fazer uma pesquisa semelhante mas separada por anos (Fig. 3).



Fig. 2. Localização dos Crimes em Chicago.

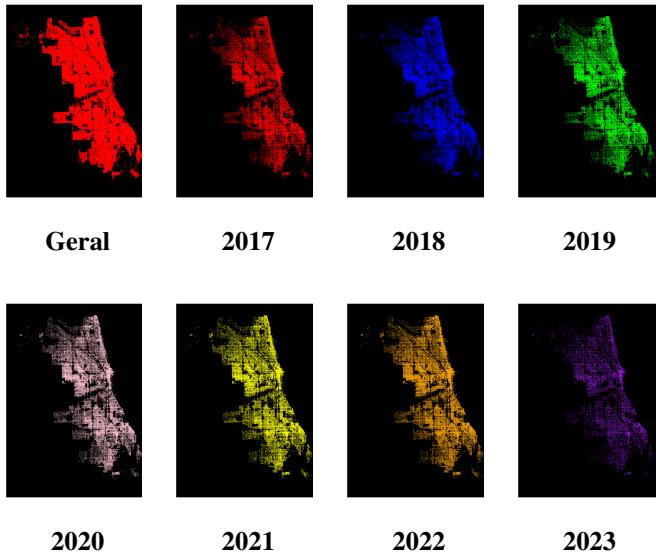


Fig. 3. Localização dos Crimes em Chicago por Ano

Podemos analisar que, como o ano de 2023 tem menos casos presentes no *dataset* (pois apenas tem quatro meses representados) a densidade da dos pontos é bastante menor, algo que a Fig. 2 não demonstra.

Para analisar melhor o tipo de crimes que ocorre em Chicago, podemos também fazer uma tabela de frequências e analisar quais os predominantes. No entanto, os códigos do FBI por si próprios não dizem nada sem um pouco de pesquisa. Como estou apenas a trabalhar com este *dataset* decidi então fazer associação dos códigos aos *Primary Types* possíveis.

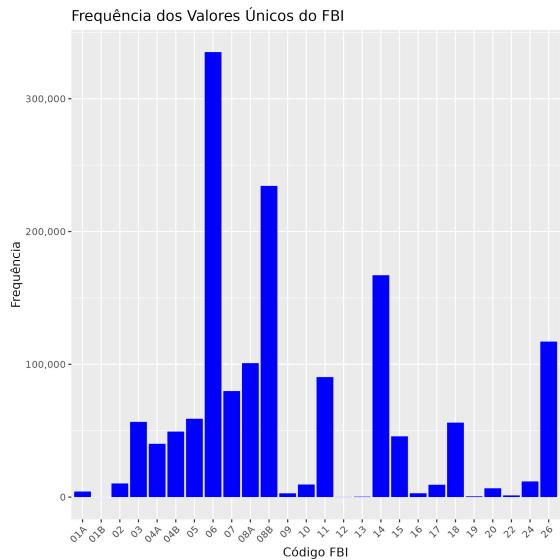


Fig. 4. Frequência dos Códigos do FBI.

Código FBI	Primary Type
0A	RECEIVED
1B	RECEIVED PROPERTY
1C	RECEIVED PROPERTY, SUSPECT UNKNOWN
1D	RECEIVED BAND
1E	RECEIVED PROPERTY, SUSPECT IDENTIFIED
1F	TYPE OFFICER, OFFICE, THREATEN, THREATENED, THREATENING, OFFICE, INVOLVED, RECENTLY, INVOLVED, PUBLIC POLICE VIOLATION, RISK
1G	TYPE OFFICER, OFFICE, THREATEN, THREATENED, THREATENING, OFFICE, INVOLVED, RECENTLY, INVOLVED, PUBLIC POLICE VIOLATION, RISK
1H	TYPE OFFICER, OFFICE, THREATEN, THREATENED, THREATENING, OFFICE, INVOLVED, RECENTLY, INVOLVED, PUBLIC POLICE VIOLATION, RISK
1I	KIDNAP, KIDNAP, CONCEALED CARRY LICENSE VIOLATION, MN, OREGON
1J	ARMED, ARMED, ARMED
1K	ARMED, ARMED, ARMED
1L	INTERFERENCE WITH PUBLIC OFFICER, PUBLIC POLICE VIOLATION, OTHER OFFICER
1M	ARMED, ARMED, ARMED
1N	ARMED, ARMED, ARMED
1O	EXPOSURE OF VULNERABLE PERSON, PUBLIC POLICE VIOLATION
1P	ARMED, ARMED, ARMED
1Q	ARMED, ARMED, ARMED
1R	ARMED, ARMED, ARMED
1S	ARMED, ARMED, ARMED
1T	ARMED, ARMED, ARMED
1U	ARMED, ARMED, ARMED
1V	ARMED, ARMED, ARMED
1W	ARMED, ARMED, ARMED
1X	ARMED, ARMED, ARMED
1Y	ARMED, ARMED, ARMED
1Z	ARMED, ARMED, ARMED
1A	RECEIVED PROPERTY

Fig. 5. Primary Types associados aos códigos do FBI.

### III. CONCLUSÃO

Neste projeto tive a oportunidade de continuar a explorar o mundo de Data Mining, de análise de dados e de algoritmos de previsão.

A maior dificuldade que encontrei na realização deste trabalho foi sem dúvida a falta de tempo e concentração num semestre tão desgastante. Sinto que tinha capacidades para produzir mais, mas as condições limitaram-me a este resultado. Uma outra dificuldade (que é habitual nesta área) deriva da quantidade de tempo necessário para correr o *dataset* de cada vez. Com um pouco mais de um milhão e meio de casos, o processamento necessário e, consequentemente, o tempo necessário não ajuda na parte de manter o fluxo de produção.