

# 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:

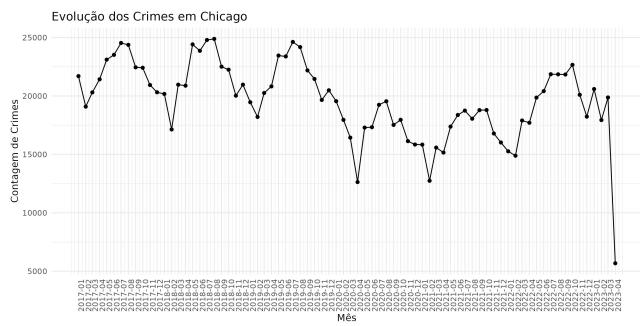


Fig. 1: Evolução dos Crimes em Chicago.

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.



Fig. 2: Localização dos Crimes em Chicago.

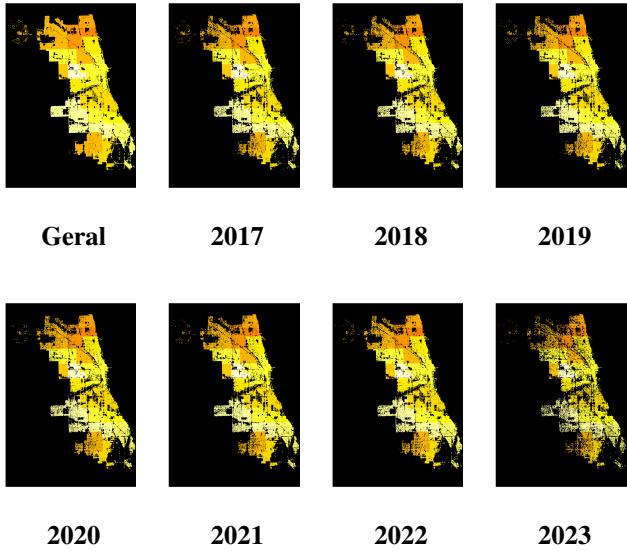


Fig. 3: Localização dos Crimes em Chicago por Ano

Decidi então fazer uma pesquisa semelhante mas separada por anos (Fig. 3). Para além disso, atribuí um esquema de cores estilo *heatmap* aos distritos, isto é, os distritos com maior incidência de crimes terão cores mais escuras e os distritos com menor incidência de crimes terão cores mais claras.

Podemos analisar na Fig. 3 que, como o ano de 2023 tem menos casos presentes no *dataset* (pois apenas tem quatro meses representados) a densidade dos pontos é bastante menor, algo que a Fig. 2 não demonstra. No entanto, esta densidade mais baixa consegue apresentar-nos algo que

Para analisar melhor o tipo de crimes que ocorreu em Chicago, podemos também fazer uma tabela de frequências e analisar quais são os crimes predominantes (Fig. 4). No entanto, os códigos do FBI por si próprios não dizem nada sem

um pouco de pesquisa externa. Como estou apenas a trabalhar com este *dataset* decidi então fazer associação dos códigos aos *Primary Types* possíveis. Podemos analisar então na Fig. 5 que alguns dos códigos estão associados apenas a um tipo de crime, mas outros estão associados até catorze tipos.

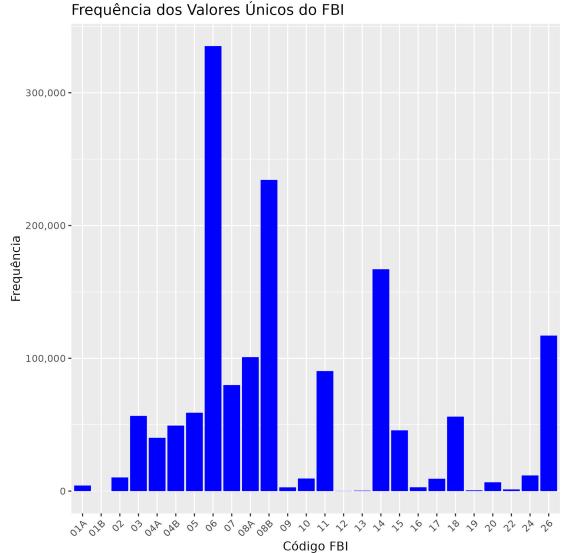


Fig. 4: Frequência dos Códigos do FBI.

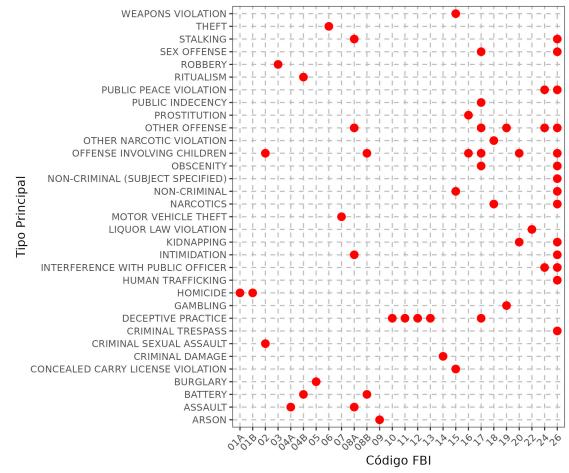
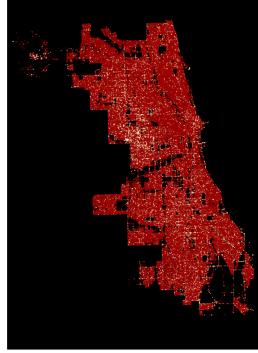


Fig. 5: Primary Types associados aos códigos do FBI.

Para finalizar, quis também fazer um mapa das localizações baseadas nos valores *Arrest* e *Domestic* (Fig. 6). Utilizei neste caso a cor amarela para identificar os casos de detenção e domésticos, e usei a cor vermelha para todos os outros casos. Conseguimos verificar que, na maioria, os crimes não exigem detenções e não são domésticos. Para concluir, embora o primeiro mapa seja um pouco mais inconclusivo, o segundo mapa demonstra toda uma área na costa norte onde há pouca incidência de crimes domésticos.



Arrest



Domestic

Fig. 6: Localização de Crimes com Detenção e Domésticos

### III. APLICAÇÃO DE MODELOS DE PREVISÃO

Nesta segunda parte do projeto, foi pedido a implementação de algoritmos de *machine learning* para previsão de crimes ou de qualquer um dos atributos. Optei pela previsão de crimes baseado na localização espacial e temporal. Numa situação ideal, seria aplicado o modelo de previsão a um *dataset* de todas as datas futuras possíveis para identificar que zonas necessitam de um patrulhamento mais elevado e que zonas podem economizar na presença policial. No entanto, o meu objetivo aqui será treinar um modelo preciso o suficiente para o problema e testar com qualquer tipo de situação (neste caso, situações predefinidas).

#### A. Escolha dos Atributos

Após a análise exploratória seria necessária uma escolha dos atributos para aplicação de algoritmos de *machine learning*. Decidi ficar com este tipo de colunas que me demonstraram maior importância para a previsão dos crimes:

- Colunas Temporais: Colunas indicativas de quando aconteceu o crime. Neste caso, temos as colunas CrimeYear, CrimeMonth, CrimeDay e CrimeTime.
- Colunas Espaciais: Colunas indicativas de onde aconteceu o crime. Neste caso, temos as colunas Block, District, Latitude, Longitude, Beat, Ward e Community Area.

Não incluí qualquer coluna relacionada a tipo de crimes ou *updates* porque queremos prever cegamente se uma zona deve ser mais policiada ou não. Não importa que tipo de crime será, importa apenas se houve ou não algum crime.

#### B. Escolha dos Algoritmos

Devido à falta de tempo, não me foi possível implementar os quatro algoritmos pedidos. No entanto, irei falar do que consegui implementar a tempo e dos seus resultados.

1) *Naive Bayes*: O primeiro algoritmo que decidi implementar foi o Naive Bayes. É um modelo que possibilita o uso de *dataframes* com colunas categóricas. No entanto, é um algoritmo que necessita de bastante cuidado no balanceamento. Isto demonstrou-se com a minha primeira tentativa.

O *dataset* inicial era composto por mil linhas verdadeiras misturadas com nove mil linhas falsas geradas aleatoriamente (dentro de limitações para gerar casos realísticos).

Num primeiro olhar, podemos analisar uma *Accuracy* bastante elevada, à volta dos 87%. No entanto, obtive uma *Recall* de 27%, o que indica uma alta taxa de acerto nos verdadeiros positivos.

Numa segunda tentativa, tentei mudar o *dataset* inicial para uma mistura de dez mil linhas verdadeiras misturadas com dez mil linhas falsas que, novamente, foram geradas aleatoriamente. Neste caso, descemos a *Accuracy* para os 60%. No entanto, subimos o *Recall* para 56%. Isto é, 56% dos casos verdadeiros foram determinados corretamente. Continua a não ser um valor aceitável, mas derivado da falta de tempo e dos constantes erros de código, foi o melhor que consegui obter.

De realçar que esta implementação foi feita com uma divisão treino/teste de 70/30.

### IV. 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.