# Aprendizado de Máquina aplicado ao cenário de Criminalidade na cidade de Chicago

**Eric Azevedo de Oliveira**[1], **Gabriel Luciano Gomes**[2], **Felipe Domingos da Cunha**[1]

[1]Instituto de Ciências Exatas e Informática
Pontifícia Universidade Católica de Minas Gerais
R. Dom José Gaspar, 500 – Belo Horizonte – MG – Brazil

eric.azevedo@sga.pucminas.br, felipe@pucminas.br

[2]Instituto de Computação
Universidade Estadual de Campinas (Unicamp)
SP – Brasil

g265673@dac.unicamp.br

***Abstract.*** *This research explored patterns of sexual crimes in Chicago, using a comprehensive method that combines spatial and social analyses with cutting-edge machine learning algorithms, such as Self-Organizing Maps (SOM), K-means, and DBSCAN. By examining public databases, it identified spatial patterns linking the frequency of these crimes to geographical and demographic factors. The findings uncover the complex nature of sexual crime patterns, providing valuable insights for developing prevention and intervention strategies. This research is crucial for leveraging smart city technologies and artificial intelligence to enhance urban safety and guide the formulation of effective public policies.*

***Resumo.*** *Este estudo investiga os padrões de crimes sexuais na cidade de Chicago, empregando uma abordagem multidisciplinar que integra análises espaciais e sociais com técnicas avançadas de aprendizado de máquina, incluindo Self-Organizing Maps (SOM), K-means e DBSCAN. Por meio da análise de bases de dados públicas, foram identificados padrões espaciais significativos que correlacionam a ocorrência desses crimes com fatores geográficos e demográficos. Os resultados revelam uma complexidades nos padrões de criminalidade sexual, oferecendo resultados favoráveis para estratégias de prevenção e intervenção. Este estudo se destaca pela sua importância da aplicação de tecnologias de cidades inteligentes e inteligência artificial para melhorar a segurança urbana e definir políticas públicas eficazes.*

## 1. Introdução

À medida que as metrópoles modernas crescem rapidamente, observa-se um paradoxo preocupante: o aumento dos índices de criminalidade, especialmente dos crimes de natureza sexual. Esses crimes deixam marcas profundas nas vítimas e têm um impacto significativo na estrutura social das cidades, abalando a sensação de segurança e coesão entre os cidadãos [Moreira de Carvalho 2006].

A cidade de Chicago, impulsionada por um robusto desenvolvimento na indústria automobilística, tornou-se um exemplo emblemático desse fenômeno. O rápido crescimento de Chicago ao longo dos anos colocou-a em destaque como um importante campo de estudo para entender a dinâmica criminal. A complexidade de seus fatores socioeconômicos e culturais oferece uma oportunidade única para investigar e encontrar soluções eficazes para combater os crimes sexuais, visando uma melhora significativa na qualidade de vida urbana.

Pesquisas recentes destacam a seriedade da situação em Chicago, onde menos de 20% das ocorrências de crimes sexuais reportadas levam a prisões. Ademais, apenas uma pequena parcela das vítimas escolhe denunciar o incidente às autoridades [News 2020]. Esta alarmante discrepância não apenas sublinha a extensão do problema, mas também realça os desafios enfrentados pelas forças policiais para responder de maneira eficaz a esses crimes e garantir a segurança da população.

Neste contexto, o presente estudo propõe-se a empregar uma nova abordagem, aplicando os princípios das cidades inteligentes e técnicas de inteligência artificial para ajudar na segurança urbana em Chicago. Por meio do desenvolvimento de modelos analíticos independentes, cada um utilizando uma técnica específica de clusterização — DBSCAN, Self-Organizing Maps (SOM) e K-Means —, este estudo visa identificar padrões associados a crimes sexuais. Analisando variáveis como: a localização dos bairros dos agressores, a proximidade de parques com esses bairros, a ocorrência de crimes anteriores, tendo como objetivo identificar os possíveis padrões espaciais e sociais que caracterizam esses tipos de crimes. Essa análise tem o potencial não apenas de acelerar o processo investigativo na cidade de Chicago, mas também de aumentar a eficácia das decisões estratégicas tomadas pelas autoridades policiais ao combate desses crimes.

A estrutura deste documento é organizada da seguinte maneira: a Seção 2 apresenta uma revisão da literatura, abordando estudos similares em diferentes contextos e trabalhos que influenciaram as diretrizes e parâmetros deste estudo. A Seção 3 explica detalhadamente como os dados foram coletados, a seleção de características relevantes e a escolha dos hiperparâmetros usados. A análise dos clusters formados e as principais conclusões são discutidas na Seção 4. Por fim, a Seção 5 recapitula os achados mais importantes do estudo e propõe caminhos para pesquisas futuras.

## 2. Trabalhos Relacionados

Esta seção apresenta um levantamento da literatura relevante, destacando técnicas e abordagens no campo da análise de crimes utilizando algoritmos de agrupamento. Analisam-se as semelhanças e inovações deste estudo em comparação com pesquisas anteriores.

O estudo de Harous et al. [Harous et al. 2019] realizou uma comparação de vários algoritmos de agrupamento em um conjunto misto de dados, tendo o destaque a eficiência do DBSCAN. Entretanto, esse estudo se diferencia ao explorar a análise de crimes sexuais em Chicago, com um foco particular nas variáveis urbanas e sociais, com intuito de enriquecer a compreensão dos padrões criminais específicos com base na região.

Em [Walczak 2021], foi empregado o uso de redes neurais para prever crimes baseando-se em dados temporais e geográficos, oferecendo um contraponto metodológico

ao presente estudo em questão. Ele adota uma abordagem de clusterização para identificar padrões de crimes sexuais, considerando um espectro mais amplo de fatores socioespaciais. Isso destaca a abordagem mais detalhada adotada neste estudo em comparação com a predição baseada em modelos temporais.

| **Autor(es)** | **Método Utilizado** | **Contribuições e Observações** |
|---|---|---|
| Harous et al. (2019) | DBSCAN e outros algoritmos de clusterização | Avaliação da eficácia de vários algoritmos de agrupamento com foco na utilidade do DBSCAN para conjuntos de dados mistos. |
| Walczak (2021) | Redes neurais | Uso de redes neurais para predição de crimes com base em dados temporais e geográficos. |
| Gomez (2021) | K-means e redes neurais | Previsão de uma variedade de crimes em Buenos Aires com uma combinação de K-means e aprendizado de máquina. |
| Groff e McCord (2010) | Análise estatística | Investigação sobre a relação entre a presença de parques e taxas de criminalidade, ressaltando o impacto dos espaços públicos nos padrões criminais. |
| Kumar (2021) | Técnicas de aprendizado de máquina para detecção de crimes | Desenvolvimento de uma metodologia para análise de relações temporais entre tipos de crimes. |
| Estudo Atual | DBSCAN, K-Means, SOM | Análise de crimes sexuais em Chicago, incorporando variáveis urbanas e sociais e utilizando uma combinação de técnicas de clusterização para revelar padrões e regiões específicas. |

**Tabela 1. Comparação das Abordagens de Análise de Crimes Sexuais.**

Gomez [Gómez et al. 2021] aplicou uma combinação do algoritmo K-means com redes neurais para prever diversos tipos de crimes em Buenos Aires, Argentina. Em contraste, o uso do K-means neste estudo, que foca na análise de crimes sexuais em Chicago, destaca uma aplicação mais específica. O algoritmo é utilizado para identificar padrões espaciais e analisar a interação com variáveis socioambientais, diferenciando-se da abordagem mais ampla de Gomez, que visa à previsão geral de crimes.

Groff e McCord [Groff and McCord 2010] investigaram o impacto da presença de parques na incidência criminal em Filadélfia. Inspirado por essa pesquisa, o presente estudo incorpora a análise da localização de parques como um fator significativo, destacando a influência de variáveis urbanas na ocorrência de crimes sexuais em Chicago, e adicionando uma nova dimensão à compreensão do contexto urbano na análise criminal.

Em contraste, o estudo descrito em [Kumar et al. 2021] introduz uma nova metodologia para a detecção de crimes, focando nas relações temporais entre diferentes tipos de delitos. Este trabalho, no entanto, expande a abordagem ao incorporar dimensões espaciais e uma ampla gama de variáveis socioambientais. Utilizando técnicas avançadas de clusterização, o estudo destaca padrões específicos relacionados a crimes sexuais em Chicago, oferecendo uma perspectiva mais profunda e abrangente sobre o tema.

Este trabalho se destaca pela abordagem integrada que emprega algoritmos de K-means, DBSCAN e Self-Organizing Maps. Na análise de crimes sexuais em Chicago, considerou-se uma ampla variedade de variáveis, incluindo dados relacionados a parques, ocorrências criminais gerais e registros de criminosos sexuais. Tal abordagem multidimensional visa elucidar os padrões subjacentes à criminalidade na região. Uma síntese comparativa das pesquisas relacionadas, os algoritmos aplicados e as principais observações podem ser encontradas na Tabela 1.

## 3. Metodologia

Esta seção descreve a metodologia adotada neste estudo. Inicialmente, a Figura 1 ilustra o fluxograma das etapas metodológicas, abrangendo desde a seleção das bases de dados até a aplicação dos modelos de agrupamento. Em seguida, detalham-se as bases de dados selecionadas como suas justificativas e relevâncias, e descrevem-se os procedimentos de pré-processamento utilizados. Sucedendo, discorre os modelos de agrupamento e os hiperparâmetros utilizados.

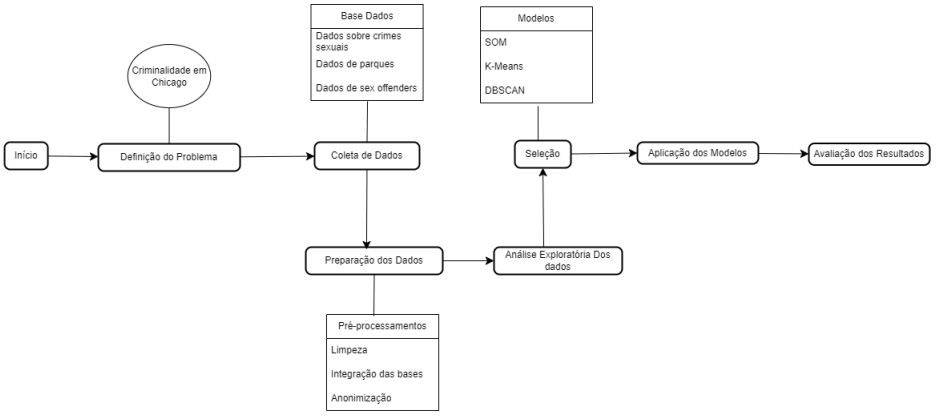


**Figura 1. Fluxograma apresentando as etapas de desenvolvimento deste estudo.**

### 3.1. Bases de Dados Utilizadas

Foram escolhidas três bases de dados públicas da cidade de Chicago, sendo elas: a base Crimes_-_2001_to_Present, a base Sex_Offenders e a base CPD_Parks. Cada uma das bases está disponível no portal da cidade de Chicago[1], e foram escolhidas por sua relevância na análise da incidência e características dos crimes sexuais, e a integração destas proporciona uma visão detalhada do contexto estudado.

**Base SEX OFFENDER:** Esta base contém informações sobre agressores sexuais registrados em Chicago. A pretexto da privacidade, os endereços são agrupados em níveis de blocos, protegendo informações exatas de localização. A base é atualizada diariamente e contém 2.775 registros com 10 atributos descritos na Tabela 2.

**Base CRIMES_-_2001_TO_PRESENT:** Contendo registros desde 2001, esta base detalha incidentes criminais em Chicago. As informações disponíveis incluem tipo de crime, localização, data e hora, entre outros detalhes, como ilustrado na Tabela 3.

[1] https://data.cityofchicago.org/

**Base CPD_PARKS:** Esta base fornece dados geográficos sobre parques em Chicago, para entender as áreas com potencial maior presença de crianças e adolescentes, que podem ser os principais alvos dos criminosos. Os atributos são descritos na Tabela 4.

| Coluna | Descrição |
|---|---|
| FIRST | Primeiro Nome |
| LAST | Último Nome |
| BLOCK | Região do Criminoso Sexual |
| GENDER | Gênero do Criminoso |
| BIRTH DATE | Data de Nascimento |
| HEIGHT | Altura do Criminoso |
| WEIGHT | Peso do Criminoso |
| VICTIM MINOR | Vítima é menor de idade? |

**Tabela 2. Atributos da base SexOffender.**

| Coluna | Descrição |
|---|---|
| Date | Horário do Crime |
| Block | Local do crime |
| Primary Type | Qual foi o crime |
| Description | Descrição do crime |
| Location | Onde ocorreu o crime |
| Latitude | Latitude |
| Longitude | Longitude |

**Tabela 3. Atributos da base Crimes.**

| Coluna | Descrição |
|---|---|
| the_geom | Geolocalização da Praça |
| PERIMETER | Perimetro da Praça |
| LOCATION | Localização |

**Tabela 4. Atributos da Base Parks.**

### 3.2. Procedimentos de Pré-processamento

Antes da aplicação nos modelos, os dados passaram por etapas de pré-processamento para remover informações desnecessárias e prepará-los para a análise. A seguir será descrito cada processo de pré-processamentos realizado em cada uma das bases, culminando na integração de ambas.

**Base SEX OFFENDER:** No pré-processamento da Base Sex Offender (Tabela 2), foram implementadas várias etapas para alinhar o estudo com as diretrizes éticas e enriquecer a análise demográfica dos dados. Após estas etapas, a base contém $1924$ registros e $15$ atributos.

- Remoção de Atributos Sensíveis: Os campos `First` e `Last` foram removidos para proteger a privacidade dos indivíduos, evitando a identificação direta e mantendo a integridade ética do estudo.
- Adição de Idade: A idade dos agressores foi calculada a partir da data de nascimento (`BIRTH DATE`) e adicionada ao dataset. A idade é um elemento crucial para entender o perfil dos agressores e para identificar padrões ou tendências específicas relacionadas à demografia.
- Obtenção de Coordenadas Geográficas: Convertendo endereços em coordenadas de latitude e longitude, foi possível realizar uma análise espacial detalhada, essencial para o mapeamento preciso dos agressores.
- Remoção dos criminosos femininos: Dados foram filtrados para focar nos criminosos masculinos, que representam a maioria dos casos e são o foco principal deste estudo, conforme ilustrado na Figura 2.

- Criação do campo AGE Superior: Introduziu-se um campo para destacar agressores acima de 30 anos, baseado na distribuição etária significativa encontrada, como mostrado na Figura 3. Esse recorte etário possibilita análises mais aprofundadas sobre a prevalência de crimes sexuais em diferentes faixas etárias.

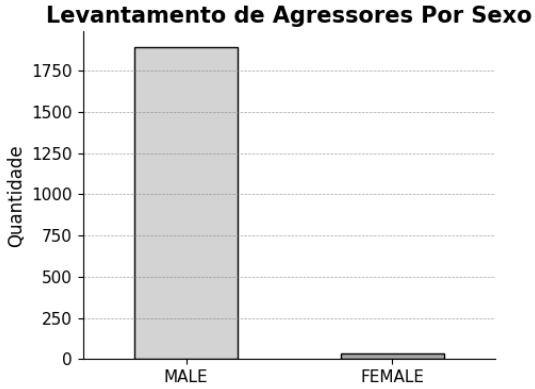


**Figura 2. Distribuição de gênero dos agressores sexuais.**

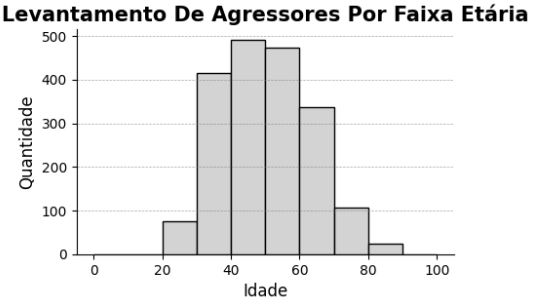


**Figura 3. Distribuição de idade dos agressores sexuais.**

**Base CRIMES:** A base de Crimes (Tabela 3) foi submetida a um processo detalhado de pré-processamento. A base resultante consistiu em 172.600 instâncias e 8 atributos.

- Remoção de Atributos Menos Relevantes: Campos como `Case Number`, `ID`, `IUCR`, `FBI Code`, `X Coordinate` e `Y Coordinate`, `Ward` foram removidos após análise de uma *matriz de correlação*. Estes atributos mostraram-se irrelevantes para os objetivos específicos deste estudo.
- Remoção de Instâncias Inconsistentes: Instâncias inconsistentes foram excluídas para manter a integridade e a confiabilidade dos dados.

**Base PARKS:** Na base de Parques (Tabela 4), o pré-processamento teve como foco a precisão geográfica. Após esta etapa, a base contém 617 instâncias e 84 atributos.

- Inferência de Coordenadas Geográficas: Utilizando a biblioteca `the_geom` [geo ], as colunas `Latitude` e `Longitude` foram criadas a partir do atributo thegeom, permitindo análises espaciais.
- Remoção de Instâncias Inconsistentes: Instâncias com inconsistências nas coordenadas foram removidas para assegurar a exatidão dos dados.

### 3.2.1. Base Final Processada

A junção das bases de dados CRIMES, SEX OFFENDERS e PARKS foi criteriosamente projetada para capturar a natureza da criminalidade sexual em um ambiente urbano. A integração de dados visou criar um mapa interativo da distribuição de agressores sexuais registrados e incidentes relatados, relacionando-os com variáveis ambientais críticas, como a proximidade de espaços públicos – especificamente parques – que podem funcionar tanto como pontos de refúgio quanto de risco no ambiente urbano.

Para alcançar uma representação acurada, as bases de CRIMES e SEX OFFENDERS foram unidas por meio do atributo `Blocks`, oferecendo uma representação detalhada da

localização geográfica. A inclusão de uma variável binária referente à incidência de crimes proporcionou uma perspectiva sobre regiões de risco divergentes, enquanto a manutenção dos detalhes dos tipos de crimes permitiu uma subsequente análise qualitativa.

Em seguida, os dados resultantes foram enriquecidos com informações georreferenciadas da Base de PARKS, ampliando o escopo analítico para incluir a interação entre crimes sexuais e espaços de lazer, potencialmente vulneráveis a tais incidentes. Esta etapa provou ser fundamental na elaboração de estratégias de segurança e prevenção, iluminando a relação entre a geografia dos espaços públicos e a incidência de atividades criminosas.

Para a construção dos modelos de agrupamento, atributos decisivos como `Menor Distância` até os parques e `Existência de Praças` foram calculados usando a distância Euclidiana, favorecendo sua interpretação intuitiva e adequação para análises de geoprocessamento. A meticulosa preparação dos dados culminou em um *dataset* final de 1924 registros com 16 atributos, representando um robusto corpus para dissecar padrões de crimes sexuais na malha urbana de Chicago[2].

### 3.3. Modelos de Agrupamento

Foram selecionados três métodos de agrupamentos estabelecidos em pesquisas acadêmicas para organizar os indivíduos com características semelhantes de maneira supervisionada, esses três métodos estas descritos abaixo.

**K-means:** Este algoritmo é conhecido por sua eficiência em termos de custo computacional, conforme destacado em [Guerreiro 2021]. A escolha desse algorítimo deve-se por sua capacidade de segmentar eficientemente o conjunto de dados em grupos com características semelhantes de crimes o que é instrumental para a identificação de padrões específicos de criminalidade sexual. A escolha do número ótimo de *clusters* foi determinada através do método do cotovelo, garantindo uma divisão que reflete a estrutura natural dos dados.

**Self-Organizing Maps:** É um algorítimo baseado em aprendizado competitivo, baseado em redes neurais, é particularmente eficaz para visualizar e interpretar padrões complexos e de alta dimensão. Sua escolha se deve por sua habilidade de mapear os dados em uma grade bidimensional, preservando a topologia dos dados originais de acordo com [Khribi et al. 2019], no qual permite identificar *clusters* e padrões espaciais nos crimes sexuais. Essa técnica tem um grande peso para esse estudo. pois facilita a visualização das relações não-lineares entre as variáveis.

**DBSCAN:** Este algoritmo diferencia por sua abordagem baseada em densidade, o DBSCAN permite identificação de *clusters* de formas arbitrárias e a detecção de *outliers* conforme descrito em [Azimi et al. 2019]. Ele foi selecionado por sua capacidade de capturar *clusters* baseados na proximidade espacial dos dados, o qual adequá com à natureza geoespacial de nossa análise de crimes, permitindo identificar regiões de alta e baixa incidência de crimes sexuais, para ajudar na prevenção dos crimes.

Cada um desses algoritmos foi escolhido por suas características intrínsecas. Enquanto o K-means oferece uma visão geral rápida dos padrões de agrupamento, o SOM proporciona uma visualização mais detalhada e o DBSCAN permite uma análise focada

[2]Todos os dados utilizados neste estudo estão disponíveis para acesso no seguinte endereço: `https://doi.org/10.5281/zenodo.10884020`.

na densidade espacial, essencial para a compreensão da distribuição geográfica dos crimes em Chicago.

### 3.4. Hiperparâmetros

Os hiperparâmetros são fundamentais para o processo de clusterização, influenciando diretamente a performance e a eficiência dos algoritmos de agrupamento utilizados. Nesta seção, discutiremos os hiperparâmetros escolhidos para os modelos de clusterização e explicaremos as justificativas para tais seleções.

**K-means:** A determinação do número ideal de *clusters* foi realizada usando a técnica de curva de aprendizado. Como mostrado na Figura 4, o número ótimo de *clusters* para o segundo vetor foi identificado como 5. Esta escolha baseia-se na análise comparativa dos resultados, onde se observou que o segundo vetor apresentava resultados mais consistentes.

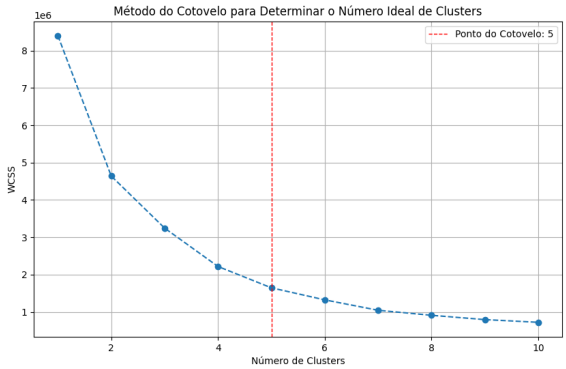


**Figura 4. Curva de aprendizado do K-means.**

**DBSCAN:** Para este algoritmo, a escolha dos hiperparâmetros *épsilon* e *minsamples* foi feita após testes variando de 0.1 a 10 e de 5 a 20, respectivamente. A seleção foi baseada na métrica de Silhouette [Rousseeuw 1987], com os melhores valores encontrados sendo *épsilon*$= 0.8$ e *minsamples*$= 5$. Esses valores permitiram uma distinção eficaz entre as áreas de alta e baixa densidade nos dados.

**SOM:** A determinação dos hiperparâmetros ótimos para o SOM envolveu a criação de um mapa de calor variando o tamanho do *grid* e testando diferentes taxas de aprendizagem e valores de sigma. A análise dos resultados levou à escolha de uma taxa de aprendizagem de $0.5$ e sigma de $1$ conforme a Figura 5, proporcionando uma organização espacial efetiva dos dados.

Cada configuração de hiperparâmetros foi cuidadosamente escolhida para otimizar a performance dos algoritmos de agrupamento no contexto deste estudo, garantindo uma análise mais precisa e eficiente dos padrões de crimes sexuais em Chicago.

## 4. Resultados

Na busca por padrões e características associadas a criminosos sexuais em áreas urbanas de Chicago, três algoritmos de agrupamento foram utilizados: Self-Organizing Map (SOM), K-Means e DBSCAN. Esses algoritmos foram selecionados com base em sua eficácia em lidar com a complexidade e dimensionalidade dos dados, cada um revelando percepções sobre a interação no âmbito deste estudo.

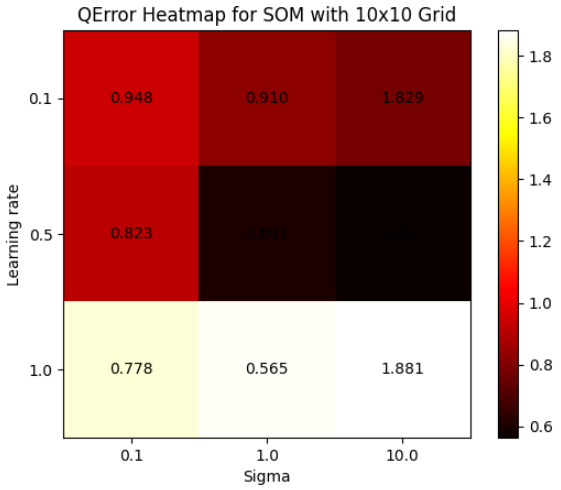


**Figura 5. Mapa de calor para otimização dos hiperparâmetros do SOM.**

### 4.1. Self-Organizing Map (SOM)

A implementação do Self-Organizing Map (SOM) permitiu a identificação de padrões espaciais, sociais e criminais nos dados de crimes sexuais em Chicago. Esta seção discute as descobertas e suas implicações para políticas públicas e intervenções de segurança.

| Grupo | Latitude | Longitude | Praças | Diversidade Racial |
|---|---|---|---|---|
| 133.0 | −0.897 | 0.965 | Baixa | Baixa |
| 221.0 | 0.620 | 0.620 | Alta | Alta |
| 86.0 | 1.071 | 0.147 | Alta | Muito Alta |
| 77.0 | 0.265 | 0.215 | Alta | Alta |
| 105.0 | −0.364 | 0.914 | Baixa | Muito Baixa |

**Tabela 5. Informações Geográficas e Sociais dos *Clusters*.**

| Grupo | Crime | Idade Superior | Vítimas Menores |
|---|---|---|---|
| 133.0 | Alto | Jovem | Média |
| 221.0 | Baixo | Média | Média |
| 86.0 | Baixo | Média | Baixa |
| 77.0 | Baixo | Jovem | Média |
| 105.0 | Baixo | Média | Baixa |

**Tabela 6. Informações sobre Crime, Idade Superior e Vítimas Menores dos *Clusters*.**

A análise SOM revelou variações nas características geográficas e sociais dos nós de crimes sexuais. Os nós variam desde áreas com baixa existência de praças e diversidade racial (Grupos 133.0 e 105.0) até regiões com alta presença de espaços públicos e grande diversidade (Grupos 221.0, 86.0, e 77.0). Esses resultados sugerem que comunidades mais integradas, com maior diversidade e infraestrutura urbana voltada ao lazer e à socialização como praças e locais púbicos, onde esses locais podem ser o ponto de partida na prevenção e detecção de crimes sexuais.

O grupo que mais chamou atenção foi o Grupo 133.0, pelo seu alto nível de crimes sexuais, tendo um destaque a idade predominantes e a presença de vitimas menores de idade. Porém não ocorreu a incidência da existência de parques, sendo assim alguma variável não presente nesse estudo deve ter um impacto significativo para a ocorrência dos crimes. Sendo assim este padrão indica áreas que exigem atenção prioritária e estratégias de prevenção focadas, especialmente considerando a jovem faixa etária predominante dos criminosos e a média incidência de vítimas menores.

A aplicação do SOM forneceu ideias sobre os padrões de crimes sexuais em Chicago, destacando a complexidade e a variedade de fatores que influenciam esses crimes. A compreensão desses padrões é crucial para desenvolver estratégias de segurança pública mais eficazes e adaptadas às necessidades específicas de cada comunidade. O trabalho futuro pode expandir essa análise, incorporando variáveis adicionais e utilizando o SOM para explorar outras dimensões dos crimes sexuais e suas dinâmicas urbanas.

## 4.2. Análise K-Means

A utilização do algoritmo K-Means na base de dados de crimes sexuais em Chicago revelou a existência de cinco *clusters* distintos, cada um com características únicas que espelham a complexa rede de padrões criminais urbanos. A visualização gráfica e as métricas quantitativas apresentadas na Figura 6 e na Tabela 7 oferecem uma visão diferenciada dos agrupamentos encontrados.

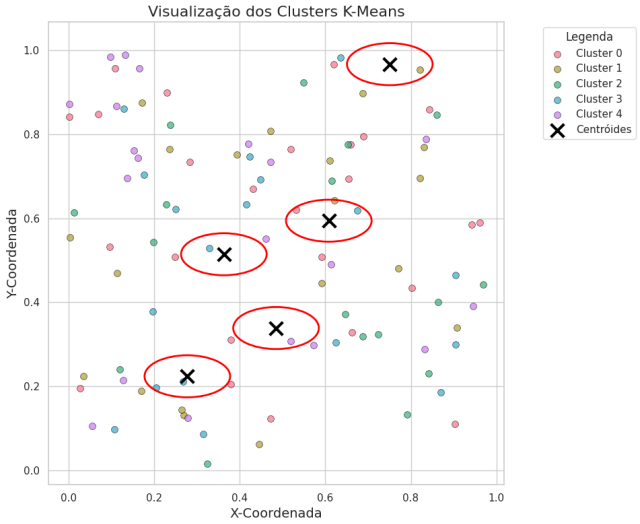


**Figura 6. Representação dos *clusters* identificados pelo K-Means, mostrando a diversidade nos perfis criminais.**

O *Cluster* 2 merece atenção especial por sua elevada média em 'Crime', uma indicação de altas taxas de criminalidade que demandam um entendimento mais detalhado e uma abordagem estratégica específica para prevenção e ação policial.

Por outro lado, o *Cluster* 4 se destaca por sua escassez de 'Existência de Praças' e alta incidência de 'VICTIM MINOR', sinalizando regiões onde a vulnerabilidade de jovens e crianças é um problema emergente. A falta de espaços públicos sugere a necessidade de políticas que promovam a vigilância e o engajamento comunitário, bem como

| Variável | Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|---|---|
| LAT | $-0.50$ | $0.70$ | $-0.13$ | $0.81$ | $-0.93$ |
| LONG | $0.44$ | $-0.60$ | $0.10$ | $-0.61$ | $0.69$ |
| Existência de Praças | $-0.32$ | $0.58$ | $0.04$ | $0.67$ | $-0.86$ |
| RACE Int | $-0.59$ | $0.41$ | $-0.04$ | $0.45$ | $-0.29$ |
| Crime | $-0.24$ | $-0.24$ | $4.20$ | $-0.24$ | $-0.24$ |
| AGE Superior | $0.28$ | $-1.71$ | $0.03$ | $0.59$ | $0.02$ |
| VICTIM MINOR | $-1.49$ | $0.12$ | $0.04$ | $0.35$ | $0.67$ |

**Tabela 7. Perfil predominante dos *clusters* identificados pelo K-Means.**

o desenvolvimento de infraestruturas que possam criar ambientes mais seguros para a população jovem.

Esta análise criteriosa dos *Clusters* $2$ e $4$ não apenas indica áreas que requerem intervenções prioritárias, mas também reforça a importância de considerar a distribuição e a infraestrutura urbana na formulação de estratégias eficazes de combate aos crimes sexuais. O perfil detalhado dos *clusters* fornece um caminho para abordagens mais direcionadas e personalizadas, alinhando os esforços de segurança pública com as características socioespaciais específicas de cada região em Chicago.

### 4.3. DBSCAN

O uso do algoritmo DBSCAN possibilitou a identificação de padrões claros dentro do contexto da criminalidade sexual em Chicago, levando em conta variáveis como localização geográfica, a existência de áreas de lazer, diversidade racial, incidência de crimes, idade da população e a ocorrência de vítimas menores de idade. Esta seção detalha os principais *clusters* encontrados e um grupo representando dados discrepantes, fornecendo um panorama das suas características distintivas.

| Cluster ID | Latitude Média | Longitude Média |
|---|---|---|
| $-1$ (Ruído) | Variável | Variável |
| $0$ | Norte | Oeste |
| $1$ | Centro | Centro |
| $2$ | Sul | Leste |

**Tabela 8. Localização Geográfica Média dos *Clusters* Identificados pelo DBSCAN.**

| Cluster ID | Existência de Praças | (RACE Int)$^3$ | Crime |
|---|---|---|---|
| $-1$ (Ruído) | Variável | Variável | Alto |
| $0$ | Alta | Alta | Baixo |
| $1$ | Média | Média | Muito Alto |
| $2$ | Baixa | Baixa | Baixo |

**Tabela 9. Características Sociais dos *Clusters* Identificados pelo DBSCAN.**

| Cluster ID | Idade Superior | Vítimas Menores |
|---|---|---|
| −1 (Ruído) | Variável | Variável |
| 0 | Velha | Média |
| 1 | Média | Média |
| 2 | Jovem | Alta |

**Tabela 10. Demografia dos *Clusters* identificados pelo DBSCAN.**

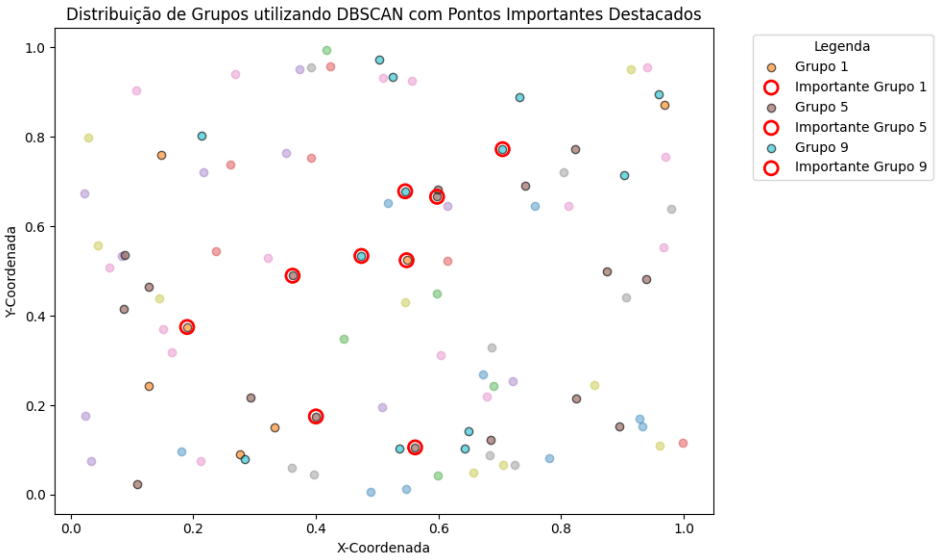


**Figura 7. Visualização dos *Clusters* Importantes Identificados pelo DBSCAN.**

As Tabelas 8, 9, e 10, juntamente com a Figura 7, fornecem um entendimento abrangente dos *clusters* identificados. Esta análise multidimensional contempla aspectos sociais, demográficos e a distribuição geográfica de crimes sexuais, oferecendo uma visão detalhada das condições que prevalecem em cada área estudada.

O *Cluster* -1, marcado como 'Ruído', compreende pontos de dados que não seguem um padrão homogêneo e podem indicar áreas com características peculiares. A variação desses dados aponta para a necessidade de investigações mais detalhadas para compreender os fatores locais e moldar políticas públicas adaptadas às especificidades de cada comunidade.

O *Cluster* 0 é caracterizado por sua localização ao norte e oeste e se distingue por uma alta disponibilidade de áreas de lazer e grande diversidade racial, que coincidem com índices de criminalidade mais baixos. Isso sugere que a presença de espaços públicos bem estruturados e a heterogeneidade cultural podem ser fatores de resiliência contra o crime.

O *Cluster* 1 apresenta elevados níveis de crime, apesar de estar situado na região central e contar com espaços públicos moderados e diversidade racial média. Essa configuração sinaliza a presença de outros elementos que potencializam a criminalidade, exigindo uma atenção redobrada em termos de segurança e medidas sociais para atenuar esses desafios.

Por fim, o *Cluster* 2, localizado nas zonas sul e leste, se notabiliza pela escassez

de espaços de lazer e por uma baixa diversidade racial, mas, ainda assim, destaca-se pela significativa presença de vítimas menores de idade. Tal observação demanda estratégias específicas de proteção e segurança para as populações jovens nessas regiões.

### 4.4. Análise Comparativa

A análise dos padrões de criminalidade sexual em Chicago foi enriquecida pelo uso de três metodologias distintas de clusterização: Self-Organizing Maps (SOM), K-Means e DBSCAN. Cada método revelou aspectos únicos da criminalidade, oferecendo perspectivas complementares que, em conjunto, proporcionam uma compreensão abrangente dos padrões criminais na região.

Self-Organizing Map (SOM): Proporcionou uma representação visual intuitiva, mapeando os dados criminais em uma matriz bidimensional que reflete a complexidade dos crimes sexuais em Chicago. Essa técnica permitiu observar a interação entre diferentes variáveis, como a localização de espaços públicos e a demografia dos envolvidos, oferecendo uma visão detalhada da estrutura dos dados.

K-Means: Este algoritmo, centrado na identificação de centróides de *cluster*, segmentou os dados em cinco grupos distintos com características bem definidas. Dois grupos se destacaram particularmente: um com baixa latitude e escassez de espaços públicos, indicando uma alta incidência de vítimas menores, e outro que, estando geograficamente e demograficamente próximo da média, registrou uma taxa elevada de crimes sexuais. Essa abordagem ilustrou a tendência dos *clusters* formados pelo K-Means a uma distribuição uniforme e separação clara, embora a sensibilidade a *outliers* possa afetar a localização dos centróides.

DBSCAN: Se diferenciou ao agrupar os dados baseando-se na proximidade espacial e densidade, identificando regiões de alta concentração de crimes e distinguindo pontos isolados como ruído. Esse método iluminou áreas críticas de alta criminalidade e revelou a variação dos tamanhos dos *clusters*, demonstrando sua eficácia em capturar distribuições de densidade variável e formas irregulares de *clusters*.

A comparação entre os algoritmos sugere que uma abordagem integrada, aproveitando as características específicas de cada método, pode fornecer a análise mais completa dos padrões de crimes sexuais em Chicago. Enquanto o K-Means oferece uma visão geral estruturada e facilmente interpretável, o DBSCAN destaca-se na identificação de áreas de alta criminalidade e o SOM na visualização de padrões complexos e relações intrínsecas nos dados. Juntos, eles compõem um panorama detalhado e informativo dos crimes sexuais na cidade, essencial para o desenvolvimento de estratégias de segurança pública adaptativas e eficazes baseadas em dados.

## 5. Conclusão e Trabalhos Futuros

Este estudo realizou uma análise de crimes sexuais em Chicago, utilizando algorítimos de agrupamento como SOM, K-Means e DBSCAN. Está metodologia não só apenas elucidou as complexas interações entre variáveis geográficas, demográficas e comportamentais, mas também ofereceu diretrizes para estratégicas de intervenção mais eficazes e focadas.

É crucial destacar que, embora significativos, os resultados deste estudo não devem ser utilizados para reforçar estigmas contra comunidades ou grupos étnicos específicos. O

propósito principal é oferecer *insights* que ajudem na distribuição eficiente de recursos para combater os crimes sexuais, mantendo sempre um compromisso com a responsabilidade ética.

Os resultados alcançados devem ser vistos somente como parte de um contexto analítico mais amplo, devido a complexidade dos fatores que influenciam os crimes sexuais na cidade de Chicago. Este estudo sublinha a necessidade de pesquisas adicionais para uma compreensão mais holística do fenômeno.

Futuras pesquisas poderiam ampliar a abordagem deste estudo para diferentes contextos urbanos, considerando desafios demográficos e de segurança variados. A integração de variáveis socioeconômicas e a adoção de métodos qualitativos, como entrevistas, proporcionariam uma visão mais abrangente e humana do problema. Investigar o impacto das políticas públicas e a eficácia das estratégias de intervenção, bem como implementar um sistema de reclamação para avaliação e melhoria contínua, são direções valiosas para pesquisas futuras.

## Referências


GeoPandas: Python tools for geographic data.

Azimi, J., Jalali, M. S., and Saeidi, R. (2019). Analysis and prediction of crimes by clustering and classification. *International Journal of Information Technology & Decision Making*, 18(4):1091–1114.

Groff, E. R. and McCord, E. S. (2010). The role of neighborhood parks as crime generators. *Security Journal*, 23(1):1–24.

Guerreiro, M. T. (2021). *Análise de Métodos de Agrupamento de Dados para Detecção de Anomalias na Precificação e Categorização de Peças da Indústria Automotiva*. PhD thesis, Universidade Tecnológica Federal do Paraná, Ponta Grossa.

Gómez, C., Barilari, S., and Pinto, D. A. (2021). Crimes prediction using artificial intelligence: A case of study in the city of buenos aires. *Sustainability*, 13(3):1081.

Harous, S., Al Harmoodi, M., and Biri, H. (2019). A comparative study of clustering algorithms for mixed datasets. *Journal of Big Data*, 6(1):1–30.

Khribi, M. K., Jemni, M., and Nasraoui, O. (2019). Analysis of students' behavior through user clustering in online learning settings, based on self organizing maps neural networks. *Journal of Educational Computing Research*, 57(5):1305–1327.

Kumar, S. A., Raj, P. J., et al. (2021). A novel approach for crime detection based on machine learning algorithms. *International Journal of Advanced Science and Technology*, 30(5):6475–6486.

Moreira de Carvalho, I. A. M. (2006). Globalização, metrópoles e crise social no brasil. *EURE (Santiago)*, 32:5 – 20.

News, W. (2020). Arrests made in just 10%-20% of chicago's sex crimes.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Walczak, S. (2021). Predicting crime and other uses of neural networks in police decision making. *Frontiers in Psychology*, 12.