PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ

ESCOLA POLITÉCNICA

PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA (PPGIa)

PREDIÇÃO DE TIPOS DE CRIMES USANDO TÉCNICAS DE MINERAÇÃO DE PROCESSOS

FERNANDA PARIZOTTO PADILHA

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática Aplicada.

CURITIBA
2024

FERNANDA PARIZOTTO PADILHA

PREDIÇÃO DE TIPOS DE CRIMES USANDO TÉCNICAS DE MINERAÇÃO DE PROCESSOS

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Informática Aplicada da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática Aplicada.

Área de Concentração: *Ciência da Computação.*

Orientador: Prof. Dr. Edson Scalabrin.

CURITIBA

2024

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central
Gisele Alves – CRB 9/1578

P123p
2024

Padilha, Fernanda Parizotto
Predição de tipos de crimes usando técnicas de mineração de processos / Fernanda Parizotto Padilha ; orientador : Edson Scalabrin. – 2024.
140 f. ; il. : 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2024
Bibliografia: f. 135-139

1. Redes neurais (Computação). 2. Processos de Markov. 3. Mineração de processos. I. Scalabrin, Edson Emilio. II. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Informática. III. Título.

CDD. 20. ed. – 004

Pontifícia Universidade Católica do Paraná
Escola Politécnica
Programa de Pós-Graduação em Informática

Curitiba, 20 de janeiro de 2025.

07-2025

DECLARAÇÃO

Declaro para os devidos fins, que **FERNANDA PARIZOTTO PADILHA** defendeu a dissertação de Mestrado intitulada "**PREDIÇÃO DE TIPOS DE CRIMES USANDO TÉCNICAS DE MINERAÇÃO DE PROCESSOS"**, na área de concentração Ciência da Computação no dia 06 de dezembro de 2024, no qual foi aprovada.

Declaro ainda, que foram feitas todas as alterações solicitadas pela Banca Examinadora, cumprindo todas as normas de formatação definidas pelo Programa.

Por ser verdade firmo a presente declaração.

gov.br Documento assinado digitalmente
EMERSON CABRERA PARAISO
Data: 21/01/2025 14:10:30-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Emerson Cabrera Paraiso
Coordenador do Programa de Pós-Graduação em Informática

Dedico este trabalho à minha família e aos meus amigos, pelo apoio em momentos de ausência.

Agradecimentos

Primeiramente, gostaria de agradecer aos meus pais, Sirlei e José, pelo pouco tempo em que passamos juntos, por me aturarem nos momentos mais difíceis, desesperadores e chocantes da minha vida e pelos conselhos durante a jornada.

Agradeço, imensamente, ao meu orientador, Professor e Doutor Edson Emílio Scalabrin, por me aceitar na sua linha de pesquisa e trabalhar com assuntos criminais. Por mais que a distância seja grande, momentos no laboratório foram de imenso aprendizado com os outros colegas e principalmente por estar pessoalmente conversando com ele. Grata também aos vários profissionais do programa PPGIa, pelo acolhimento. Principalmente a PUCPR, por me dar a oportunidade de estar ingressando nesta universidade em que era meu sonho estudar.

Agradeço também aos meus pets, primeiramente à minha calopsita de 15 anos, o Nego (em memória), por me acalmar, dormir comigo Segundo, aos meus cachorros Caramelo, por brincar e dormir comigo em alguns momentos e a nova integrante da família a Dory, por ser uma criança doce, meia e caçadora. Por fim, agradeço ao meu galo Frederico (em memória) por me ensinar que os mais fracos também se sobressaem.

Aos meus amigos, sou imensamente grata à minha melhor amiga, Jealice, pelos melhores conselhos, puxões de orelha e paciência quando eu estava em desespero, querendo desistir de tudo e, principalmente, por ceder um espaço da sua residência em Curitiba. Meu excelentíssimo e grandioso amigo de anos de karatê, André, pelos conselhos, puxões de orelha e risadas. Agradeço, imensamente, a minha amiga de faculdade e mestra Gabriela Corbari, que sempre me apoiou nos momentos mais difíceis, assim como eu também dei apoio a ela, nos momentos mais difíceis. Agora desejo o melhor para ela, nessa nova caminhada de doutorado. Aos meus novos amigos adquiridos no ambiente universitário e do laboratório.

Agradeço também aos professores membros da banca, professor Dr. Paulo Varela, pelos anos de ensinamento e inspirações no tema da criminologia na graduação, e participar da banca. Ao professor Braulio Avila, por instigar o conteúdo mais a fundo, em conversas no laboratório. E por último agradeço a UTFPR, campus Francisco Beltrão, por ceder o espaço no laboratório.

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro concedido por meio da bolsa de estudos durante o doutorado. Este trabalho contou com o suporte da CAPES - Código de Financiamento 001.

Velocidade é isso, sentir-se livre. Quando corremos, pensamos na vida, ficamos em silêncio, ouvindo apenas o barulho do motor. (Paul Walker, 2009)

Resumo

O crime é um fenômeno universal, profundamente enraizado nas dinâmicas sociais ao longo da história. O aumento constante na incidência de atividades criminosas exige respostas inovadoras e eficazes, especialmente no contexto do apoio à segurança pública. Este trabalho propõe uma abordagem híbrida e interdisciplinar, integrando Mineração de Processos, Redes Neurais Recorrentes (RNNs, LSTMs, GRUs) e Cadeias de Markov para a predição de tipos de crimes. O objetivo principal é mostrar a viabilidade e a eficácia dessas técnicas, assim como explorar sua aplicação prática no suporte à alocação eficiente de recursos e à formulação de estratégias preventivas pelas forças de segurança. A metodologia desenvolvida é composta por seis fases principais, aplicadas a um conjunto de dados históricos contendo 512.657 registros de crimes da cidade de Denver, Colorado. Inicialmente, utilizam-se técnicas de *mineração de processos* para descobrir o modelo subjacente das ocorrências no formato de DFGs (*Directly-Follows Graphs*). Esses DFGs fornecem uma representação temporal dos eventos, permitindo a criação de uma matriz de transição baseada em Cadeias de Markov. Posteriormente, as Redes Neurais Recorrentes são utilizadas para realizar a predição de tipos de crimes para o ano/período seguinte (n+1), identificando aqueles com maior probabilidade de ocorrência. Os resultados obtidos mostram alta eficácia do método proposto. As redes neurais atingiram uma acurácia de 94% na predição dos tipos de crimes, enquanto as métricas de precisão e *recall* para os crimes mais prováveis (*top-1*) alcançaram médias de 95%. Além das predições, o método oferece visualizações interpretáveis dos padrões subjacentes e dos elementos que fundamentam as predições, facilitando discussões estratégicas e tomadas de decisão pelas autoridades competentes. A análise destaca a capacidade do método em identificar comportamentos frequentes e padrões relevantes ao longo do tempo. Essa pesquisa contribui para o avanço no uso de técnicas híbridas em problemas de predição criminal, destacando a integração de aprendizado de máquina, modelagem estatística e mineração de processos. Os resultados podem impactar a eficácia das políticas de segurança pública, promovendo uma abordagem mais informada, preditiva e estratégica no enfrentamento da criminalidade.

Palavras-chave: Palavras-chave: Predição, Rede Neural, Cadeia de Markov, Mineração de Processo, Tipo de Crime.

Abstract

Crime is a universal phenomenon deeply rooted in the social dynamics throughout history. The continuous increase in criminal activities demands innovative and effective responses, especially in the context of public safety support. This work proposes a hybrid and interdisciplinary approach, integrating Process Mining, Recurrent Neural Networks (RNNs, LSTMs, GRUs), and Markov Chains for crime type prediction. The main objective is to demonstrate the feasibility and effectiveness of these techniques, as well as to explore their practical application in supporting efficient resource allocation and the formulation of preventive strategies by security forces. The developed methodology consists of six main phases, applied to a historical dataset containing 512,657 crime records from the city of Denver, Colorado. Initially, process mining techniques are used to discover the underlying model of occurrences in the form of Directly-Follows Graphs (DFGs). These DFGs provide a temporal representation of events, enabling the creation of a transition matrix based on Markov Chains. Subsequently, Recurrent Neural Networks are employed to predict crime types for the next year/period (n+1), identifying those with the highest likelihood of occurrence. The results demonstrate the high effectiveness of the proposed method. The neural networks achieved an accuracy of 94% in predicting crime types, while precision and recall metrics for the most probable crimes (top-1) reached averages of 95%. In addition to predictions, the method provides interpretable visualizations of underlying patterns and the elements supporting these predictions, facilitating strategic discussions and decision-making by competent authorities. The analysis highlights the method's ability to identify frequent behaviors and relevant patterns over time. This research contributes to the advancement of hybrid techniques in criminal prediction problems, emphasizing the integration of machine learning, statistical modeling, and process mining. The findings have the potential to significantly impact the effectiveness of public safety policies, fostering a more informed, predictive, and strategic approach to combating criminality.

Keywords: Prediction, Neural Network, Markov Chain, Process Mining, Crime Type.

Lista de Figuras

Lista de Tabelas

Lista de Gráfico

Equações

Lista de Abreviaturas e Siglas

DFG – Directly-Follows Graph.

PM4PY – Plataforma de mineração de processos de código aberto escrita em Python.

QGIS – Sistema de Informação Geográfica.

MSL - Mapeamento Sistemático da Literatura.

GQM - Goal-Question Metric.

SPIDER - Sample; Phenomen of Interest; Design; Evaluation; Research type.

LSTM - Long Short Term Memory .

PNL - Processamento de Linguagem Natural.

RTM - Risk Terrain Modeling.

RNN - Recurrent Neural Networks

GRU - Gated Recurrent Unit .

LBSN - Location-Based Social Networks.

SVM - Support Vector Machine.

CDR - Registros De Detalhes De Chamadas.

RF - Random Forest.

RFR - Random Forest Regression.

DTR - Regressão De Árvore De Decisão.

MLR - Multi-Layer Perceptron.

SLR - Regressão Linear Simples.

SVR - Support Vector Regression.

IPC - Código Penal Indiano.

KNN - K-Nearest Neighbor.

ARIMA - Autoregressive Integrated Moving Average.

KDE - Kool Desktop Environment.

MPL - Multi-Layer Perceptron.

DT - Decision Tree.

LB - Lasso Bayesian.

LR - Linear Regression.

GIS - Informação Geográfica Técnicas.

NN - Neural Networks.

MAXEN T - Maximum Entropy Classifier.

SLDA - Scaled Linear Discriminant Analysis.

MPL - Perceptron Multicamadas.

XGBoost - eXtreme Gradient Boosting.

VAR - Vector Autoregression.

CAS - Crime Anticipation System.

RN - Redes Neurais.

LR - Regressão Logistica.

CART - Classification and Regression Tree.

LDA - Linear Discriminant Analysis.

MCLR - Multi-class Logistic Regression.

GBM - Gradient Boosting Machines.

CF - Collaborative Filtering.

BPMN - Business Process Model and Notation.

MT - Matriz Transição.

MA - Matriz Adjacente.

CTM - Current Time Matrix.

SVP - State Vector Present.

SVE - State Vector Evaluation.

Sumário

1. INTRODUÇÃO

A criminalidade é um fenômeno universal que atravessa séculos, impactando sociedades de diferentes formas e intensidades, principalmente com a evolução tecnológica e com o crescimento urbano que contribuíram para mudanças nos padrões de crimes, tornando a identificação e a predição desses eventos mais desafiadoras. Diante desse cenário, compreender as dinâmicas criminais é essencial para desenvolver estratégias de segurança pública mais eficazes. Este trabalho explora métodos inovadores para prever tipos de crimes, integrando técnicas como mineração de processos, redes neurais recorrentes (RNNs), com o objetivo de oferecer uma abordagem robusta e aplicável.

A segurança pública pode implementar melhorias significativas ao obter uma ferramenta que, baseada em dados históricos e predições precisas, permite identificar padrões criminais e prever a ocorrência de diferentes tipos de crimes. Nesse sentido, a análise detalhada, oferecida pela mineração de processos e pelas redes neurais, possibilita aos gestores de segurança pública uma melhor alocação de recursos, maior assertividade em estratégias preventivas e o aprimoramento da tomada de decisão. Além disso, ao integrar abordagens preditivas inovadoras, como redes RNNs e cadeias de Markov, esta dissertação promove um avanço no desenvolvimento de sistemas de apoio à segurança pública, fortalecendo a capacidade de prevenção e intervenção em contextos urbanos.

Por longos períodos, sofremos com crimes: independentemente de ser uma bala roubada, eles são bíblicos. Diante desse cenário cronológico, os crimes foram agregando mais violência nos seus atos, tornando a sociedade, nos dias atuais, mais aniquiladora. (Sullivan & Piquero, 2016) salientam que, em 1970, foi introduzido o conceito de carreira criminal, que teve impacto significativo no âmbito da criminologia e tornou-se uma estrutura conceitual para orientar as proposições teóricas e as de investigação.

Nesse aspecto, a criminalidade é um conjunto de atividades ilegais que uma pessoa pode escolher como meio para ganhar dinheiro ou poder, sendo altamente lucrativa. Além disso, as atividades criminosas podem causar danos significativos às comunidades e à sociedade como um todo, como o tráfico de drogas, roubo, extorsão, branqueamento de dinheiro, fraude, entre outros.

Durante um período na criminalidade, o indivíduo se interessa em aprender sobre características de tipos de delito, possuindo vários fatores-chave realizáveis — internos ou externos — por parte do indivíduo. Ou seja, estes fatores podem ser psicológicos, psíquicos, traumáticos, de infância, entre outros fatores possíveis. (Blumstein, Cohen, & Farrington, 1988, p. 2) ressaltam que, na idealização da vida do sujeito, este poderá aprender características dos tipos de crimes, levando a padrões criminais ofensivos. Estes fatores podem sugerir uma especialização em atos ofensivos que o encaminhem a uma vida criminal. Além disso, as atividades ilícitas podem causar danos significativos às comunidades e à sociedade como um todo.

Sob esse viés, há hipóteses que tentam explicar o porquê de as pessoas escolherem envolver-se com a criminalidade. Elas incluem as experiências e escolhas que, ao longo da vida, podem levar ao envolvimento no crime. (Blumstein, Cohen, & Farrington, 1988) analisam, de forma quantitativa, os tipos de crime e a ofensa à justiça criminal, proporcionando assim a oportunidade de testar, de confirmar e de refutar teorias sobre carreiras criminosas. Diante disso, vários fatores teóricos podem indicar o envolvimento em atividades criminosas, entre eles, a escolha racional, a trajetória de vida criminosa e a subcultura.

Ademais, o crime é um fenômeno universal que perdura ao longo dos tempos, enraizado na sociedade mundial, porém, o crescente aumento na incidência de atividades criminosas sempre constituiu um alerta constante. Por gerações, temos confrontado com esses desafios em diferentes cidades, estados e países ao redor do globo. Além disso, ao traçarmos a cronologia da história criminal, é evidente que a sociedade tem se ambientado com o avanço das tecnologias para frear este problema (Blumstein, Cohen, & Farrington, 1988).

Diante desses avanços tecnológicos, surge a necessidade premente de explorar abordagens inovadoras para compreender, prever e combater o crime de maneira mais eficaz. Pois, a fusão entre tecnologia e criminalidade tornou-se uma realidade tangível, com ferramentas como a análise de dados, inteligência artificial e mineração de processos, oferecendo novas perspectivas e capacidades sem precedentes.

Nesse sentido, a análise e a identificação de padrões criminais desempenham um papel crucial no desenvolvimento de estratégias eficazes de aplicação da lei e na promoção da

segurança pública. Nos últimos anos, o campo da mineração de processos, aliado ao aprendizado de máquina e à modelagem probabilística, tem se destacado como uma abordagem para compreender e antecipar tendências criminais. Esta dissertação se concentra na aplicação conjunta dessas três abordagens — Mineração de Processos, redes neurais recorrentes (RNNs) e cadeias de Markov — para a predição de tipos de crimes.

A Mineração de Processos, derivada da área de gestão de processos de negócios, permite extrair conhecimento a partir de sequências de eventos registrados em sistemas de informação. No contexto criminal, ela possibilita a visualização e análise detalhada de como os crimes ocorrem, identificando etapas, sequências e correlações entre diferentes tipos de atividades criminosas. Complementarmente, modelos de aprendizado profundo, como as RNNs e suas variantes GRU e LSTM, são ferramentas para capturar padrões temporais complexos e para realizar predições com maior precisão, especialmente em contextos nos quais fatores históricos e temporais têm relevância.

Adicionalmente, as Cadeias de Markov, com sua abordagem probabilística baseada na transição de estados em sistemas dinâmicos, são utilizadas para prever a evolução de tipos de crimes de forma mais simples e eficiente em situações em que as dependências de curto prazo entre eventos são predominantes. Embora não possuam a mesma capacidade de modelagem de sequências complexas que as redes neurais, sua simplicidade e interpretabilidade oferecem um complemento em análises exploratórias e no suporte à validação de modelos.

O objetivo desta dissertação é explorar a viabilidade e eficácia da integração dessas abordagens. Combinando a análise detalhada, proporcionada pela Mineração de Processos com o poder preditivo de redes neurais profundas como GRU e LSTM, avaliou-se a *accuracy*, a *precision* e o *recall* dessas ferramentas no suporte à predição de tipos de crimes. Além disso, um complemento probabilístico reforça a análise de padrões de transição mais simples entre tipos de crimes. A partir da integração desses métodos, ampliaram-se a compreensão e a capacidade de antecipação de padrões criminais, contribuindo significativamente para a predição e a prevenção de crimes no contexto da segurança pública.

1.1. Problema

O crime é um fenômeno universal que atravessa séculos, impactando sociedades de diferentes formas e intensidades. A evolução tecnológica e o crescimento urbano têm contribuído para mudanças nos padrões de crimes, tornando a identificação e predição desses eventos cada vez mais desafiadora. Nota-se que o crime está presente na sociedade e é considerado por muitos, senão pela maioria, como um dos problemas sociais mais inquietantes. Desse modo, constitui-se como um grande problema de segurança humana, confrontando a todos. Além disso, o público em geral tende a possuir uma percepção de aumento da incidência de atividades criminosas, como homicídio, roubo, assalto à mão armada, sequestro, banditismo, tráfico de drogas, infração de trânsito, estupro, assassinato, abuso de drogas, corrupção, assalto e perseguição, entre outros (Jonathan, Olusola, Bernadin, & Inoussa, 2021).

A criminalidade não apenas reflete os desafios sociais contemporâneos, mas também exige novas abordagens para análise e combate. Conforme destacado por (Heidensohn, 1989), o principal problema no estudo do crime é demonstrar uma problemática clara e lidar com as consequentes dificuldades em estudá-la. Em meio a essa complexidade, a predição de tipos de crime surge como uma ferramenta poderosa para auxiliar órgãos competentes na formulação de estratégias mais eficazes.

A segurança pública enfrenta o desafio de alinhar recursos escassos com demandas crescentes. Neste cenário, a mineração de processos e as redes neurais recorrentes (RNNs) oferecem uma solução promissora, pois permitem identificar padrões criminais a partir de dados históricos e prever a ocorrência de crimes futuros. Essa predição pode ajudar na alocação mais eficiente de recursos, na formulação de políticas públicas de prevenção e na criação de estratégias específicas para mitigar os efeitos de diferentes tipos de crimes.

Embora técnicas de mineração de processos e redes neurais sejam amplamente utilizadas em diferentes domínios, sua aplicação conjunta para a predição de crimes ainda é limitada. Essa lacuna motivou a presente pesquisa, que busca unir essas ferramentas para aprimorar a predição de tipos de crimes e potencializar o impacto na segurança pública. Assim, esta dissertação não apenas avança o estado da arte, mas também oferece contribuições práticas, como a possibilidade de prever crimes com maior precisão e apoiar estratégias de prevenção baseadas em dados.

Complementando essas abordagens, redes neurais recorrentes (RNNs), incluindo suas variantes como GRU e LSTM, têm se mostrado altamente eficazes na captura de padrões temporais complexos em grandes volumes de dados sequenciais. Essas ferramentas são particularmente úteis na análise de séries temporais, permitindo prever eventos futuros com base em históricos detalhados, como a ocorrência de crimes em diferentes regiões e contextos. Ao contrário de métodos tradicionais, as redes GRU e LSTM lidam eficientemente com longas dependências temporais, aprimorando significativamente a precisão das predições.

Adicionalmente, as cadeias de Markov, com abordagem probabilística baseada na transição de estados, oferecem uma alternativa simples e eficiente para prever a probabilidade de transição entre diferentes tipos de crimes. Esses modelos podem ser combinados com redes neurais para criar sistemas preditivos mais completos, que consideram tanto padrões simples quanto relações complexas nos dados criminais.

Diante disso, a presente pesquisa propõe explorar a integração dessas técnicas — mineração de processos, cadeias de Markov e redes neurais recorrentes — para a predição de crimes. Essa abordagem visa não apenas analisar dados criminais de maneira abrangente, mas também identificar padrões e prever a probabilidade de ocorrência de diferentes categorias de crimes com maior precisão e eficiência. Assim, questiona-se: **Como predizer a ocorrência de diferentes tipos de crimes em áreas urbanas, permitindo identificar padrões e tendências em dados criminais, utilizando técnicas de mineração de processos, cadeias de Markov e redes neurais recorrentes (RNNs), incluindo suas variantes LSTM e GRU?**

1.2. Motivação

O principal fator motivacional para a elaboração desta dissertação se encontra no fato de que a luta contra o crime é uma tarefa complexa e desafiadora a qual requer a colaboração de múltiplos atores — as forças de segurança, o poder judiciário e a sociedade em geral. Além disso, é importante destacar que o combate ao crime deve ser feito de forma justa e equilibrada, respeitando os direitos e as garantias dos cidadãos.

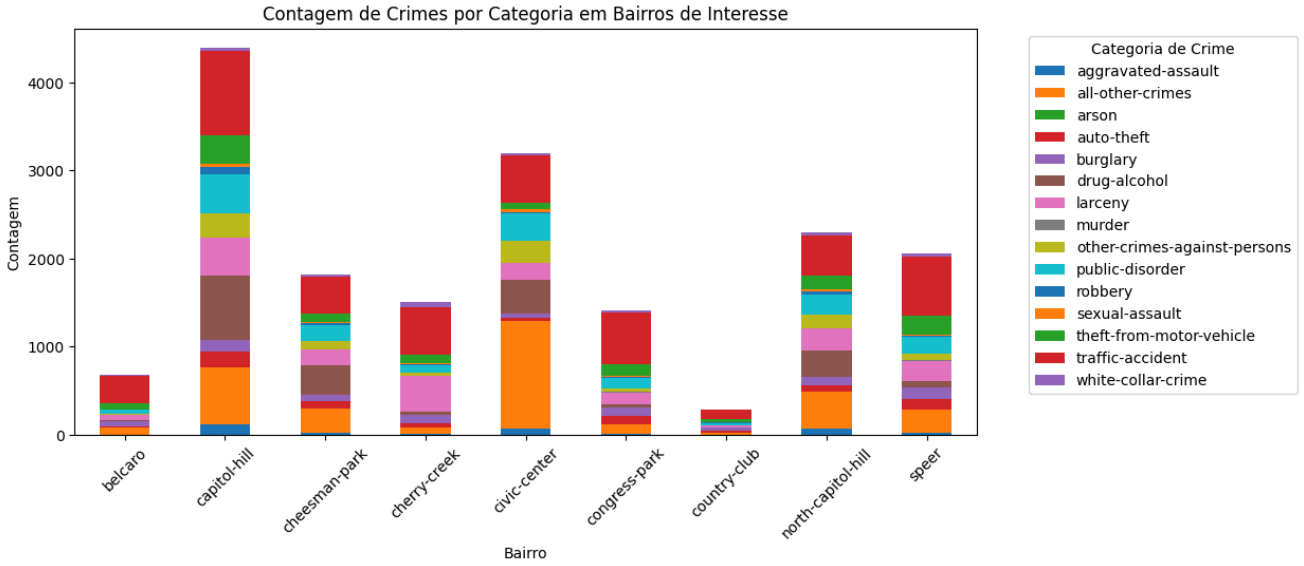
Observa-se que a maioria dos sistemas jurídicos também classifica crimes com o objetivo de atribuir casos a diferentes tipos de tribunal. Pois, as mudanças sociais, muitas

vezes, resultam na adoção de novas leis criminais e na obsolescência das mais antigas. A principal motivação diz respeito em facilitar a detecção de onde ocorrerão os próximos crimes, cujo índice (de crimes ocorridos) até a atualidade está crescendo significativamente, e não somente saber como diminuir este índice, mas ocasionar descongestionamento nas prisões.

Para uma melhor análise, o gráfico de barras empilhadas representa a contagem de crimes por categoria em diferentes bairros. Cada barra corresponde a um bairro, e as cores diferentes dentro de cada barra representam categorias específicas de crimes. O eixo Y do gráfico, denominado "Contagem", indica o número de crimes, enquanto o eixo X lista os nomes dos bairros.

Diante disso, o modelo integrado proposto nesta dissertação apresenta uma motivação prática clara: fornecer melhorias para as forças de segurança pública ao identificar padrões criminais que auxiliem na alocação de recursos, no planejamento estratégico e na formulação de políticas públicas. Além disso, a predição de crimes pode ajudar a antecipar cenários críticos, permitindo uma intervenção mais eficiente e preventiva, o que pode melhorar significativamente a segurança em comunidades urbanas e reduzir os custos associados a ações reativas e de investigação.

Gráfico 1 - Contagem de crimes por categoria nos bairros da área central de Denver.



Fonte: A autora.

Há uma legenda à direita que corresponde às cores das barras com as categorias de crimes. As categorias incluem: "aggravated-assault" (agressão agravada), "all-other-crimes" (todos os outros crimes), "arson" (incêndio criminoso), "auto-theft" (furto de veículo automotor), "burglary" (roubo), "drug-alcohol" (drogas-álcool), "larceny" (furto), "murder" (homicídio), "other-crimes-against-persons" (outros crimes contra pessoas), "public-disorder" (desordem pública), "robbery" (assalto), "sexual-assault" (assalto sexual), "theft-from-motor-vehicle" (furto de veículo motorizado), "traffic-accident" (acidente de trânsito) e "white-collar-crime" (crime de colarinho branco).

Os bairros listados no eixo X são, da esquerda para a direita: "belcaro", "capito-lhill", "cheesman-park", "cherry-creek", "civic-center", "congress-park", "country-club", "north-capitol-hill" e "speer".

Nesse sentido, é possível observar que o bairro "Civic Center" possui a contagem total mais alta de crimes, com uma grande quantidade de crimes classificados como "all-other-crimes", enquanto "Belcaro" possui a menor contagem total de crimes visível no gráfico. Pois, o gráfico fornece uma visão geral comparativa da segurança dos bairros em questão, baseada na contagem e tipos de crimes reportados.

Em resumo, o *Gráfico 1* fornece uma visão geral das ocorrências de diferentes tipos de crimes em cada um dos bairros listados. Ela é útil para entender como a incidência de diferentes tipos de crimes varia de um bairro para outro.

Diante disso, combater o crime pode ser muito desafiador e complexo, porém, fornece informações para os órgãos competentes e para a sociedade em geral, a fim de melhorar a segurança.

1.3. Objetivos

Objetivo Geral

Predizer a ocorrência de diferentes tipos de crimes em áreas urbanas uma vez que permite identificar padrões e tendências em dados criminais — utilizando técnicas de mineração de processos, cadeias de Markov e redes neurais recorrentes (RNNs) —, além de incluir suas variantes LSTM e GRU.

Para a consecução desse objetivo, foram estabelecidos os seguintes objetivos específicos:

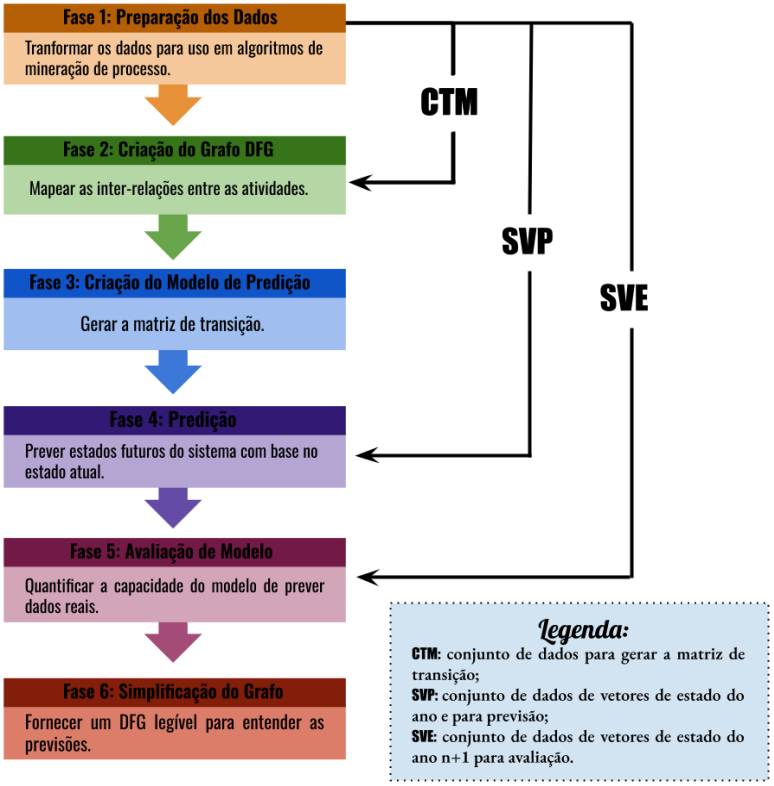
- Criar um modelo de predição que mostre a relação entre atividades;
- Representar as inter-relações diretas das atividades criminais;
- Gerar informações que ajudem na alocação de recursos.

1.4. Método da Pesquisa

O método proposto visa prever, considerando o estado *n*, que crime será o próximo $(n + 1)$, a partir do histórico de crimes reais de cada indivíduo. Tais predições para o crime seguinte $(n + 1)$ e as visualizações dos elementos podem subsidiar discussões estratégicas em outros contextos.

O método, exibido na *Figura 1,* é uma adaptação de técnicas de mineração de processos e de princípios de Markov (Dupuis, Dadouchi, & Agard, 2022). Consiste em 6 fases que levam a uma predição do tipo de crime que pode ocorrer $(n + 1)$ e a um grafo (de atividades e transições) que representa as relações entre os tipos de crimes. Além disso, essa preparação também permitiu a aplicação de redes neurais recorrentes (RNNs) e suas variantes avançadas, como LSTM e GRU, para capturar padrões temporais complexos e aprimorar a precisão das predições, especialmente em contextos em que os dados apresentam dependências de longo prazo.

Figura 1 - Fases da predição de Crime.



Fonte: Adaptado de (Dupuis, Dadouchi, & Agard, 2022, p. 03)

A primeira fase de preparação trata de dados que possam ser duplicados, dados ausentes e uma generalização dos dados, a fim de obter uma tabela utilizada em um algoritmo de mineração de processos. Já na segunda fase, os dados, no formato correto, obtêm as informações necessárias e descobrem o grafo *Directly Follows Graph* (DFG). Assim, a representação visual é dada por este grafo, no qual cada aresta representa a relação entre duas atividades e seu peso correspondente ao número de ocorrências dessa relação no conjunto de dados.

A terceira fase de criação de um modelo de predição se dá pelo método aplicado, de modo a seguir os princípios da cadeia de Markov e gerar uma matriz de transição para representar a probabilidade de transição entre os tipos de crimes. Na quarta fase de predição, a matriz, previamente criada a partir da matriz de transição, é usada como um conjunto de dados e um dicionário de codificação, associando um número a cada atividade presente no conjunto de dados. Para obter um vetor probabilístico, utiliza-se o cruzamento de um produto

de um vetor SVP e a matriz de transição; assim, o conjunto de dados SVP é transformado em um conjunto de dados de vetores binários.

A quinta fase — a avaliação do modelo — consiste em um vetor de probabilidade associado a cada atividade, indicando a probabilidade de que ocorra no tempo n+1. Dessa forma, as atividades podem ser classificadas de acordo com a probabilidade de ocorrência, utilizando as métricas de *recall* e *precisão*. Na sexta e última fase, a simplificação do grafo fornece informação para explicar a origem dos resultados previstos: a utilização desse grafo DFG permite destacar as relações mais frequentes entre as atividades.

Finalmente, os dados preparados e estruturados em etapas anteriores foram utilizados para treinar redes neurais recorrentes (RNNs) e suas variantes avançadas, LSTM e GRU, com o objetivo de predizer tipos de crimes de forma mais precisa. Essas redes foram ajustadas para identificar padrões temporais complexos e longas dependências nos dados, oferecendo predições complementares e mais robustas em relação ao método baseado na matriz de transição e ao grafo DFG.

1.5. Escopo

Optou-se por utilizar um conjunto de probabilidades de transição que determina a probabilidade de um sistema passar de um estado para outro. Portanto, a probabilidade de transição para um determinado estado depende apenas do estado atual, e não dos estados anteriores. Como se evidencia, é uma ferramenta para a modelagem de sistemas complexos e a predição de eventos futuros, com base em dados históricos. No entanto, para o conjunto de dados analisado neste estudo, os resultados apresentaram uma performance satisfatória (TOP-3, entre 69% e 79%), indicando em certa medida a captura de padrões mais complexos e de longa dependência temporal nos dados criminais.

Diante disso, foram utilizadas redes neurais recorrentes (RNNs) e suas variantes LSTM e GRU como complementos à abordagem probabilística. Essas redes, projetadas para lidar com sequências temporais e identificar dependências de longo prazo, demonstraram uma capacidade superior de capturar padrões mais complexos nos dados, oferecendo uma performance preditiva mais robusta e alinhada às expectativas. Essa integração permitiu combinar a simplicidade das probabilidades com a sofisticação das redes neurais para obter resultados mais consistentes e precisos na predição de tipos de crimes.

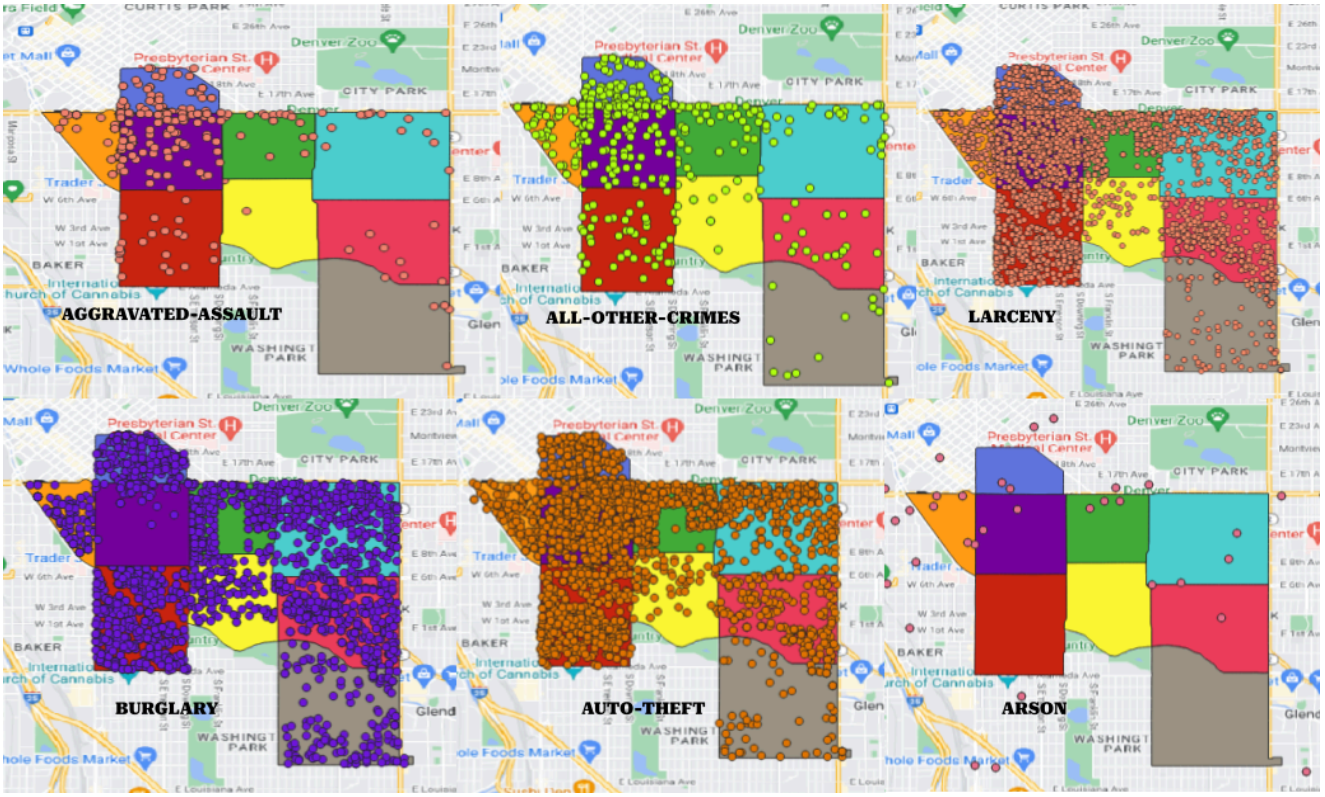
Além disso, os resultados obtidos neste estudo possuem grande potencial de aplicação prática no contexto da segurança pública. Ao permitir a identificação de padrões criminais recorrentes e prever cenários futuros, o modelo proposto pode auxiliar na alocação mais eficiente de recursos das forças policiais, na formulação de estratégias preventivas e na criação de políticas públicas mais eficazes. A aplicação desses resultados contribui para a melhoria da segurança urbana, reduzindo custos operacionais e promovendo uma abordagem mais preventiva e proativa na gestão da criminalidade.

Nesse contexto, optou-se por utilizar a biblioteca PM4PY, que suporta algoritmos de mineração de processos modernos, amplamente testados e reconhecidos como soluções de última geração, destacando-se por sua implementação em Python. Para descrever o código de predição de roubos em ambiente urbano, utilizou-se o *Google Colab* devido a sua acessibilidade, exigindo apenas uma conta no Gmail, e por ser uma plataforma on-line com salvamento automático. Essas características tornam o *Colab* especialmente útil em casos de imprevistos, como falhas na máquina, problemas técnicos de conexão à internet ou interrupções de energia.

Além disso, foi transcrito para o *software* PYCHARM, que facilita a organização das etapas do processo em código, permitindo que cada fase do estudo seja apresentada de forma clara e sequencial, culminando na predição do crime representado por um grafo para melhor análise. A escolha da biblioteca PM4PY foi motivada por sua facilidade de uso e pela ampla adoção dentro do laboratório de pesquisa, a fim de fortalecer sua aplicação neste trabalho em detrimento de outras ferramentas.

O software QGIS foi escolhido para possibilitar a visualização dos tipos de crime em forma de mapa, podendo observar o mapeamento na área central da cidade de estudo e analisar os tipos de crime de forma individual/bairro, conforme a *Figura 2*. Para tanto, é de fácil manuseio o livre acesso aos acadêmicos e com suporte rápido via e-mail, porém, não se optou por outro software por ser de relevância nacional brasileira.

Figura 2 - Tipos de crimes, da categoria OFFENSE_CATEGORY_ID.



Fonte: A autora.

1.6. Background

A predição de tipos de crimes é uma tarefa crucial para órgãos de segurança e para outras áreas (Dupuis, Dadouchi, & Agard, 2022). Em vista disso, a mineração de processos (W. van der Aalst, 2011), (W. M. P. van der Aalst, 2016) e as cadeias de Markov (Douc, Moulines, Priouret, & Soulier, 2018) são técnicas que podem ser utilizadas para prever tipos de crimes com base em dados históricos. Nesse sentido, a mineração de processos é uma técnica que analisa dados de eventos para exibir padrões e relacionamentos entre eles. Já as cadeias de Markov são modelos estatísticos que descrevem a probabilidade de um evento ocorrer com base no evento anterior.

A integração das técnicas de mineração de processos e redes neurais recorrentes proposta neste trabalho possibilita uma análise mais detalhada dos padrões temporais e espaciais associados à criminalidade, permitindo não apenas identificar tendências, mas também fornecer subsídios para a tomada de decisões estratégicas por parte das forças de segurança. Essa abordagem oferece ferramentas para a alocação eficiente de recursos,

planejamento de ações preventivas e desenvolvimento de políticas públicas que respondam de forma proativa aos desafios impostos pela criminalidade em áreas urbanas.

Embora redes neurais recorrentes (RNN), incluindo suas variantes LSTM e GRU, não tenham sido incluídas na fundamentação teórica deste trabalho, elas foram empregadas como ferramentas de comparação para avaliar o desempenho dos modelos de predição dos tipos de crimes. Essas redes oferecem uma abordagem alternativa baseada em aprendizado profundo para lidar com dependências temporais complexas nos dados, permitindo comparar a eficácia e as limitações das abordagens tradicionais, como mineração de processos e cadeias de Markov, com técnicas mais avançadas de modelagem preditiva.

Para utilizar essas técnicas na predição de tipos de crimes, é necessário coletar e analisar dados históricos sobre ocorrências criminais — localização, horário, tipo de crime e outras variáveis relevantes. Esses dados são, então, utilizados para treinar o modelo de predição, o qual é testado em um conjunto de dados de validação para avaliar sua precisão na predição de tipos de crimes futuros.

Em termos de conteúdo de predição de tipos de crimes, usando técnicas de mineração de processos e cadeias de Markov, há várias áreas de estudo, como as seguintes formas: análise e predição de relatórios criminais em Bangladesh (Pavel Rahman et al., 2021), mitigando vulnerabilidades por meio de predições e análises de tendências criminais (Orong, Sison, & Hernandez, 2018), crime em áreas urbanas: uma perspectiva de mineração de dados (X. Zhao & Tang, 2018), análise exploratória de dados e predição de crimes para cidades inteligentes (Pradhan, Potika, Eirinaki, & Potikas, 2019b), VisCrimePredict: um sistema para predição e visualização da trajetória do crime a partir de fontes de dados heterogêneas (Morshed et al., 2019), uma pesquisa sobre análise e predição do crime (Thomas & Sobhana, 2022). Predição da hora e localização de crimes futuros com métodos de recomendação (Y. Zhang, Siriaraya, Kawai, & Jatowt, 2020b), o uso da análise preditiva na predição do crime espaço-temporal: construindo e testando um modelo em um contexto urbano (Rummens, Hardyns, & Pauwels, 2017). Uso da modelagem de terreno de risco para prever crimes relacionados a moradores de rua em Los Angeles, Califórnia (Yoo & Wheeler, 2019). Predição de eventos criminais com recursos dinâmicos (Rumi, Deng, & Salim, 2018b). Prevenção do crime de furtos de ônibus em Pequim, China: a qualidade do ar afeta o crime? (Ding & Zhai, 2019). Melhoria da predição da criminalidade a curto prazo com fluxos de mobilidade humana e arquiteturas de aprendizagem profunda (J. Wu, Abrar, Awasthi,

Frias-Martinez, & Frias-Martinez, 2022). Algoritmos de aprendizado de máquina para predição de crimes sob o Código Penal Indiano (Aziz, Sharma, & Hussain, 2022a). Sistema de identificação de escritor para manuscrito offline pré-segmentado Caracteres Devanagari usando k-NN e SVM (Dargan, Kumar, Garg, & Thakur, 2019). Predição de eventos recorrentes usando técnicas de mineração de processos e princípios de Markov (Dupuis, Dadouchi, & Agard, 2022). Essa última técnica foi a inspiração para o nosso projeto.

1.6.1. Predição de risco de crimes de roubo em comunidades urbanas

A predição de crimes é um método que analisa e investiga conjuntos de dados relacionados a crimes, com referências notáveis em trabalhos como os de (Pavel Rahman et al., 2021), (Orong, Sison, & Hernandez, 2018), (Thomas & Sobhana, 2022), (Y. Zhang, Siriaraya, Kawai, & Jatowt, 2020b), (Yoo & Wheeler, 2019), (Ding & Zhai, 2019), (Dupuis, Dadouchi, & Agard, 2022). Esse método emprega técnicas de análise de dados para identificar áreas ou momentos com maior probabilidade de ocorrência de roubos em áreas urbanas, além de centrar-se na predição e inferência da probabilidade de crime ou reincidência.

Denota-se que as técnicas, empregadas nesse processo, incluem análise de séries temporais, regressão logística e redes neurais, que possibilitam a identificação de padrões e de tendências nos dados históricos de roubos. A partir desses dados, é viável criar modelos de predição de riscos de roubos em comunidades urbanas.

Por outra premissa, a predição espaço-temporal, abordada em trabalhos como os de (Rummens, Hardyns, & Pauwels, 2017), fundamenta-se na teoria de repetição de crimes, que sugere que o risco de um determinado tipo de crime pode se propagar em um intervalo de tempo específico. Desse modo, a aplicação de tais modelos permite uma compreensão mais aprofundada dos fenômenos estudados e, consequentemente, o desenvolvimento de medidas preventivas mais eficazes.

Embora estudos anteriores tenham fornecido referências para a pesquisa de predição de crimes, poucos exploraram a aplicação de técnicas de mineração de processos e cadeias de Markov na predição de crimes em ambientes urbanos e roubos em cidades inteligentes, com o intuito de antecipar o risco de crimes comunitários.

Nesse contexto, a presente dissertação busca preencher essa lacuna, propondo um modelo que combina mineração de processos e cadeias de Markov para predizer diversos tipos de crimes em ambientes urbanos. Além disso, redes neurais recorrentes (RNN),

incluindo LSTM e GRU, foram utilizadas como ferramentas de comparação para avaliar a eficácia do modelo proposto em relação a técnicas avançadas de aprendizado profundo. Essa abordagem tem o potencial de contribuir significativamente para o aprimoramento da prevenção de crimes em comunidades urbanas, fortalecendo a eficácia das medidas preventivas e apoiando a tomada de decisão em segurança pública.

1.6.2. Mineração de processos de mobilidade humana para predição de crimes

A análise da mobilidade humana, por meio da mineração de processos, envolve a aplicação de técnicas de processamento de dados para identificar padrões de deslocamento e de movimentação de indivíduos em áreas específicas. Essas informações podem ser combinadas com dados relativos à incidência de crimes em uma determinada região, cujo objetivo é construir modelos de predição de atividades criminosas.

Além disso, a utilização dessa técnica é viabilizada por algoritmos de aprendizado de máquina, como redes neurais e regressão logística, que permitem a análise eficiente de grandes volumes de dados. Pois, os modelos de predição, resultantes deste processo, proporcionam uma base sólida para a implementação de medidas preventivas, contribuindo, assim, para a redução da criminalidade em áreas específicas.

De acordo com pesquisas conduzidas por (X. Zhao & Tang, 2018), (Pradhan, Potika, Eirinaki, & Potikas, 2019b), (Morshed et al., 2019), (J. Wu, Abrar, Awasthi, Frias-Martinez, & Frias-Martinez, 2022), (Aziz, Sharma, & Hussain, 2022a), (Dargan, Kumar, Garg, & Thakur, 2019), os crimes em áreas urbanas são percebidas por sua alta densidade populacional, infraestrutura urbana, atividades comerciais e uma variedade de desafios e de oportunidades associados à vida na cidade. Esses delitos podem abranger uma ampla gama de atividades criminosas, como crimes contra a propriedade, crimes contra pessoas, delitos de colarinho branco, crimes relacionados a substâncias ilícitas, infrações de trânsito, delitos cibernéticos e outros.

Embora estudos anteriores tenham contribuído para a pesquisa sobre mobilidade urbana, uma lacuna importante reside na aplicação de técnicas de aprendizagem de máquina e de predição de crimes em ambientes urbanos, especialmente em cidades inteligentes, com o intuito de antecipar o risco de atividades criminosas.

Nesse contexto, esta dissertação apresenta uma proposta inovadora: a criação de um modelo integrado de predição de crimes que combina a mineração de processos com o uso de cadeias de Markov. Essa abordagem tem o potencial de antecipar diversos tipos de atividades criminosas em ambientes urbanos, contribuindo para o desenvolvimento de estratégias de prevenção mais eficazes e para o fortalecimento da segurança pública em comunidades urbanas.

Adicionalmente, para avaliar a eficácia do modelo proposto, foram realizadas comparações com redes neurais recorrentes (RNN) e suas variantes mais avançadas, como LSTM e GRU. Essas redes, reconhecidas por sua capacidade de capturar padrões temporais complexos e dependências de longo prazo, foram utilizadas como *benchmarks* para identificar vantagens e limitações do modelo baseado em mineração de processos e cadeias de Markov, ampliando a análise da abordagem proposta.

1.7. Considerações Finais

Esta seção apresentou uma contextualização detalhada sobre o fenômeno da criminalidade, a problemática que permeia sua análise e combate, a motivação para explorar abordagens preditivas e inovadoras no contexto da segurança pública, objetivos geral e específicos, técnicas de mineração de processos, que consistem em seis fases de predição de tipo de crime, escopo do projeto e *background*. Tal discussão fornece o embasamento necessário para compreender a relevância e a aplicação do modelo integrado proposto nesta dissertação.

.

2. MAPEAMENTO SISTEMÁTICO DA LITERATURA

O MSL (**M**apeamento **S**istemático da **L**iteratura) foi baseado nas diretrizes recomendadas por (Lazar, Feng, & Hochheiser, 2010). O objetivo deste MSL é definido de acordo com o paradigma métrico de questão de objetivos do *Goal-Question Metric* (GQM), como mostrado na *Tabela 1*. As etapas deste MSL são expostas abaixo.

Tabela 1 - Base para a construção do protocolo de mapeamento do MSL.

Analisar	Publicações científicas.
Com o propósito de	Caracterizar.
Em relação a	Predizer a recorrência de tipos de crimes em ambientes urbanos.
Do ponto de vista dos	Pesquisadores.
No contexto de	Fontes primárias disponíveis no mecanismo de busca das bibliotecas digitais.

Fonte: A autora.

2.1. Questão da Pesquisa

Este MSL tem como questão principal de pesquisa: “Quais são os sistemas de predição de tipos de crimes utilizando mineração de processos e cadeia de Markov em ambientes urbanos?”. Dessa forma, temos duas subquestões que foram definidas, conforme descrito na *Tabela 2*.

Tabela 2 – Subquestões da pesquisa.

ID	Subquestão da pesquisa	Objetivos
SQ1	Como predizer crimes/roubos em ambiente urbano?	Analisar qual é o método de predição de tipos de crimes, com base na exploração de dados em ambientes urbanos, como método, técnica, entre outros.
SQ2	Quais métodos de predição são usados para predizer roubos em ambiente urbano?	Analisar quais aplicações estão sendo utilizadas.

Fonte: A autora.

2.2. Estratégia da Pesquisa

Para otimizar a melhoria da estratégia de buscas nas bibliotecas digitais, SPIDER (*Sample; Phenomen of Interest; Design; Evaluation; Research type*) foi definido como método de busca automática. Realizado por mecanismo de busca utilizando palavras-chave definidas com base em (Donato & Donato, 2019), utilizou-se SPIDER para definir as palavras-chave, as quais foram agrupadas em duas partes: (1) Criminologia: indica onde o tema de pesquisa será contextualizado, citado por (Blumstein, Cohen, & Farrington, 1988). (2) Intervenção: refere-se aos recursos utilizados - nesse caso, a predição, utilizando mineração de processos e cadeias de Markov, citados por (W. van der Aalst, 2011) e (Dupuis, Dadouchi, & Agard, 2022) . Na *Tabela 3* são apresentadas as *strings* de busca do MSL.

Tabela 3 – *Strings* de busca do MSL.

Criminologia	(“crime”) AND (“theft”)	AND
Intervenção	(“prediction”)	AND

Fonte: a autora.

Os estudos foram investigados em quatro bibliotecas digitais: IEEE Xplore, ACM, SCIENCE DIRECT e SPRINGER. Estas bibliotecas foram priorizadas pela sua qualidade e desempenho de busca, pela variedade de publicações de diferentes áreas que possuem e por

fazerem parte da área de pesquisa de Inteligência Artificial e Predição, conforme se vê na *Tabela 4*:

Tabela 4 – Fonte utilizada no MSL.

Nome da fonte	Link	Tipo de Pesquisa
IEEEXplore	https://ieeexplore.ieee.org/	Máquina de busca
ACM	https://dl.acm.org/	Máquina de busca
SCIENCE DIRECT	https://www.sciencedirect.com/	Máquina de busca
SPRINGER LINK	https://link.springer.com/	Máquina de busca

Fonte: A autora.

2.3. Critérios de Seleção dos Artigos

Os critérios de seleção foram baseados de acordo com (Budgen & Brereton, 2006), em que há os critérios de inclusão dos artigos (I) e critérios de exclusão dos artigos (E).

Critérios de inclusão (I) do artigo:

- (I1) Devem responder pelo menos às duas questões do MSL;
- (I2) Artigos publicados em *Journal* e *Conferences;*
- (I3) Artigos que apresentam informações e discussões sobre predição de crimes/comportamento criminosos em ambientes urbanos;
- (I4) Artigos publicados de 2013 a 2023.

Critérios de exclusão (E) do artigo:

- (E1) Não serão aceitos artigos que não atendam aos critérios de inclusão;
- (E2) Artigos publicados em outras línguas, que não sejam a inglesa e a portuguesa;
- (E3) Artigos publicados em *Early Access Articles, Magazines, Books*;
- (E4) Artigos que não possuam acesso livre, ou seja, artigos pagos;
- (E5) Artigos duplicados, como, por exemplo, publicados em diferentes bibliotecas virtuais;
- (E6) Artigos publicados em anos anteriores a 2013.

O processo de seleção dos artigos foi realizado em três etapas: a primeira (1º filtro) consistiu na leitura do título e do resumo dos artigos retornados pela biblioteca online, sendo feita a análise dos critérios de inclusão e exclusão; a segunda (2º filtro) consistiu na leitura total dos artigos que passaram pela primeira etapa, sendo também avaliados os critérios de inclusão e exclusão; e a terceira (3° filtro) consistiu novamente na leitura dos artigos que passaram pela segunda etapa, visto que foram escolhidos somente os que obtiveram assuntos semelhantes à proposta do projeto. A pesquisa foi realizada somente pela autora, em especial, as três etapas e de forma individual. Da mesma forma, a *string* de busca, rodada de 01/02/2022 até 30/10/2023, pois, dentro do programa de ensino laboratorial, não há alunos nesta linha de pesquisa para ser dividido o processo de seleção e análise dos artigos. As *strings* de busca e o critério de inclusão foram aplicados na biblioteca digital, sendo retornados 2.582 artigos. Desses, 165 passaram pelo primeiro filtro, 81, pelo segundo e 15, pelo terceiro filtro, conforme a *Tabela 5*.

Tabela 5 – Totalização de artigos selecionados no 1º, 2º e 3° filtros.

Fonte	**Retornado (s)**	**1° Filtro**	**2° Filtro**	**3° Filtro**
IEEEXplore	25	9	2	2
Springer	996	71	16	3
ACM	617	40	17	5
Science Direct	944	45	46	5
TOTAL	2.582	165	81	15

Fonte: A autora.

Os 15 artigos, com suas referências, são apresentados na *Tabela 6*.

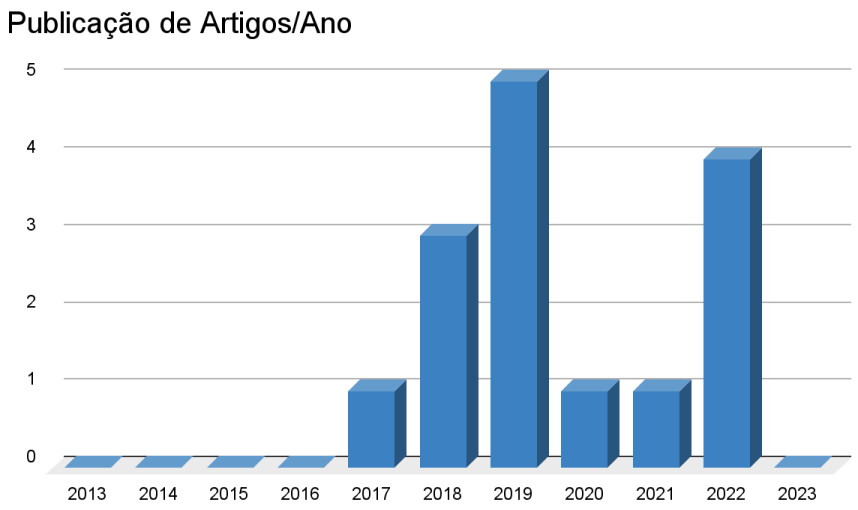
Tabela 6 – Referências das iniciativas.

Fontes	**Referências**
IEEExplore	(Pavel Rahman et al., 2021) e (Orong, Sison, & Hernandez, 2018).
ACM	(X. Zhao & Tang, 2018), (Pradhan, Potika, Eirinaki, & Potikas, 2019b), (Morshed et al., 2019).
SPRINGER	(Rumi, Deng, & Salim, 2018b), (Ding & Zhai, 2019), (J. Wu, Abrar, Awasthi, Frias-Martinez, & Frias-Martinez, 2022), (Aziz, Sharma, & Hussain, 2022a), (Dargan, Kumar, Garg, & Thakur, 2019).
SCIENCE DIRECT	(Thomas & Sobhana, 2022), (Y. Zhang, Siriaraya, Kawai, & Jatowt, 2020b), (Rummens, Hardyns, & Pauwels, 2017), (Yoo & Wheeler, 2019), (Dupuis, Dadouchi, & Agard, 2022).

Fonte: A autora.

Os artigos selecionados foram publicados entre 2017 e 2022, conforme indicado abaixo, no *Gráfico 2*. Com início em 2017, foram apresentados trabalhos sobre a predição de crimes em ambientes urbanos, utilizando mineração de processos e cadeia de Markov, demonstrando um interesse por parte dos pesquisadores neste tópico. O maior índice de estudo selecionado no MSL ocorreu nos anos de 2018, 2019 e 2022. Por outro aspecto, nos anos de 2013, 2014, 2015, 2016 e 2023, não houve tantos resultados de estudo relacionados. Já em 2017, 2020 e 2021, obtiveram o resultado de um artigo.

Gráfico 2 – Ano de publicação dos artigos selecionados do MSL.



Fonte: A autora.

No MSL, foram analisados os locais de publicações dos artigos selecionados. Os *Gráficos 3 e 4* contribuíram para uma visão geral dos locais de publicação das Conferências e dos *Journals.* Como se observa na *Tabela 7*, os quatro artigos — publicados em conferências — foram apresentados, de modo que um para cada evento.

Tabela 7 - Representação das conferências por fontes.

Fontes	**Conferências**
IEEE	• International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM); • 5th International Conference on Business and Industrial Research (ICBIR).
ACM	• International Conference on Knowledge Discovery & Data Mining (SIGKDD); • International Database Applications & Engineering Symposium (IDEAS); • Symposium on Applied Computing (SIGAPP);
SCIENCE DIRECT	• International Conference on Artificial Intelligence & Energy Systems (ICAIES)

Fonte: A autora.

Os dez artigos publicados em *Journal*s foram apresentados unicamente para cada evento, conforme a *Tabela 8*.

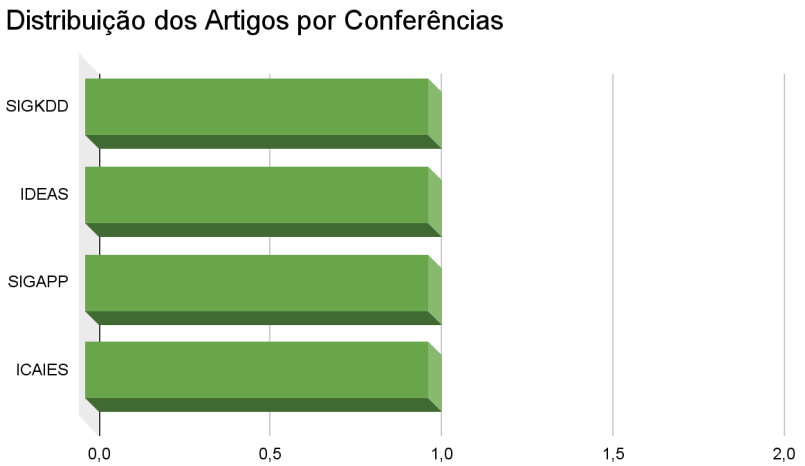
Tabela 8 - Representação dos *Journals* por fontes.

Fontes	***Journals***
SPRINGER	• *Data Science* (EPJ); • Security Journal (SJ); • *Annals of Data Science* (AODS)*;* • Soft Computing (SC).
SCIENCE DIRECT	• Knowledge-based Systems (KBSes); • Applied Geography (AG); • Computers and Electronics in Agriculture (CEA).

Fonte: A autora.

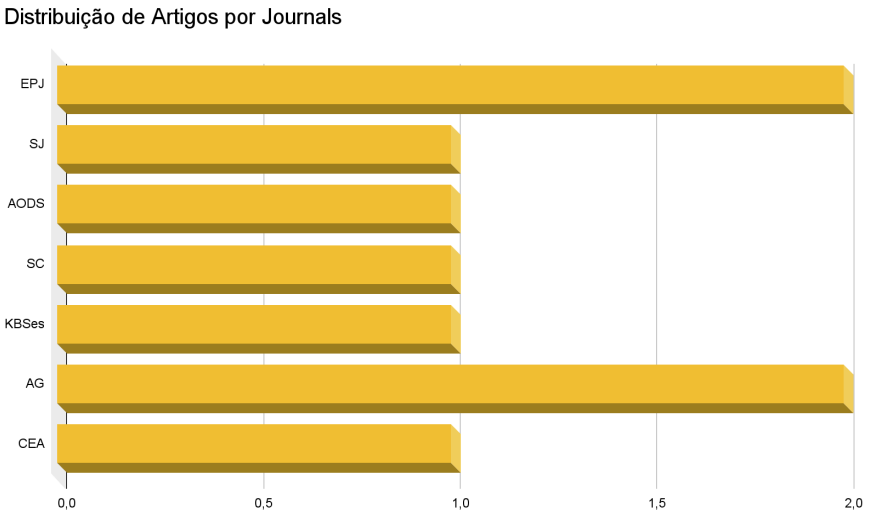
A representação de conferências da *Tabela 7* está gerada no *Gráfico 3*, assim como a representação de *Journals* da *Tabela 8* está representada no *Gráfico 4*.

Gráfico 3 – Distribuição de Artigos por Conferências.



Fonte: A autora.

Gráfico 4 – Distribuição de Artigos por *Journals*.



Fonte: A autora.

2.4. Resposta às Subquestões da Pesquisa

Na busca por compreender e mitigar o fenômeno da criminalidade em centros urbanos, é imperativo focar em métodos que possam antecipar e, consequentemente, prevenir ocorrências ilícitas. Esta seção se dedica a responder perguntas essenciais que pavimentam o caminho para tal entendimento: "Como predizer crimes/roubos em ambiente urbano?" e "Quais métodos de predição são usados para predizer roubos em ambiente urbano?". Aqui, discutiremos os esforços interdisciplinares que combinam diversas áreas de estudo e análise de dados, proporcionando uma visão abrangente das técnicas preditivas mais eficazes atualmente. Por intermédio da síntese de estudos de caso e literatura especializada, delinearemos como os métodos preditivos podem tornar-se uma ferramenta indispensável para táticas preventivas e no alocamento otimizado de recursos.

2.4.1. Predição de crimes/roubos em ambiente urbano (SQ1)

A questão SQ1 aborda a área de pesquisa relacionada à predição de crimes e roubos em ambientes urbanos. Trata-se de um campo interdisciplinar que faz uso da análise de dados e de algoritmos para investigar dados históricos, com o objetivo de identificar padrões que possam indicar a probabilidade de ocorrência futura de roubos e de crimes. Essa é uma

questão de grande relevância, uma vez que a predição de crimes desempenha um papel fundamental na manutenção da segurança pública e na otimização dos recursos das agências de aplicação da lei.

Ademais, a síntese dos artigos relacionados à questão SQ1 no Mapeamento Sistemático da Literatura (MSL) será realizada com foco em três aspectos principais:

Base de Dados: um aspecto central na predição de crimes/roubos é a utilização de uma base de dados sólida e confiável. Pois, os artigos analisados forneceram *insights* sobre as fontes de dados utilizadas, a qualidade dos dados e a abrangência temporal e, em espacial, bem como quaisquer desafios enfrentados na coleta e na preparação dos dados.

Área do Crime/Tipo de Espaço do Crime: a localização geográfica desempenha um papel crítico na predição de crimes. Nesse caso, os estudos abordados na SQ1 destacaram a importância de considerar a área do crime ou o tipo de espaço do crime, como áreas urbanas, subúrbios, bairros específicos, entre outros. Além disso, foi explorada a relevância de fatores geoespaciais na análise, como densidade populacional, características ambientais e infraestrutura.

Tipos de Crimes: a natureza dos crimes a serem previstos é diversificada. Pois, os artigos, analisados na SQ1, abordaram os diferentes tipos de crimes, incluindo roubos, furtos, agressões, crimes contra a propriedade e outros. Destacou-se ainda a importância de adaptar os modelos de predição para cada tipo de crime, levando em consideração suas características específicas.

Com essa abordagem detalhada, a questão SQ1 forneceu uma visão abrangente sobre a predição de crimes/roubos em ambiente urbano, permitindo a compreensão das metodologias, dos desafios e das descobertas nessa área crucial de pesquisa. Essa expansão proporciona uma visão mais completa da questão SQ1, detalhando os principais tópicos que foram abordados na síntese dos artigos relacionados.

- Análise e predição de relatórios criminais em Bangladesh (Pavel Rahman et al., 2021): analisou-se a predição da criminalidade em Bangladesh por ser um país de terceiro mundo, no qual a polícia conta com a ajuda dos distritos e dos seus respectivos órgãos para registrar os tipos de crimes ocorridos visto que existe, no país, uma alta taxa de criminalidade. Neste artigo, utilizou-se a abordagem de aprendizado de máquina para analisar dados de crimes anteriores, predizendo ocorrências futuras de crimes. Como o autor relata, a abordagem de aprendizado de máquina serve para

coletar e analisar estatísticas criminais no sítio eletrônico da Polícia de Bangladesh, em que os algoritmos de aprendizagem automática e técnicas de mineração de dados podem ser usados para analisar dados criminais e prever eventos futuros, beneficiando, em última análise, as agências e as autoridades responsáveis pela aplicação da lei na prevenção e resolução de crimes. Com os crimes não relatados na maioria dos acontecimentos em Bangladesh, nos anos de 2001 a 2020, o conjunto de dados deve ser analisado com cautela para encontrar dados consistentes ao ano. O autor iniciou a análise da latitude e longitude, realizando várias predições, como o número de eventos que ocorreram por distritos. Quanto ao método deste artigo, argumentar-se-á na questão SQ2.

- Mitigando vulnerabilidades por meio de predições e análises de tendências criminais (Orong, Sison, & Hernandez, 2018); analisaram-se dados indexados sobre crimes da província de Misamis Ocidental, Filipinas, e fornecem uma predição de sua ocorrência nos próximos cinco anos. Em vista disso, salienta-se que o crime é uma ofensa contra a sociedade e, muitas vezes, é punível pela lei. O autor destaca que a mineração de dados é uma técnica que pode auxiliar a autoridade em problemas de detecção de crimes, tendo como ideia capturar anos de experiência humana em modelos computacionais por meio de mineração de dados. Assim, para mitigar as vulnerabilidades trazidas pelos crimes cometidos na comunidade, os estudos preveem as ocorrências dos crimes e realizam análises de tendências dos delitos utilizando técnicas de mineração de dados. Uma das primeiras tarefas, na mineração de dados e na detecção criminal, envolve a visualização dessas associações, sendo que os analistas de inteligência, juntamente com os investigadores criminais, devem frequentemente relacionar a grande quantidade de informações sobre pessoas em associações criminosas enganosas e políticas, baseadas no medo, opiáceos e outras associações criminosas. Quanto ao método deste artigo, será argumentado na questão SQ2.
- Crime em áreas urbanas: uma perspectiva de mineração de dados (X. Zhao & Tang, 2018): apresenta-se uma visão geral das principais teorias da criminologia, resume a análise do crime em dados urbanos, revisa algoritmos de última geração para vários tipos de tarefas de crime computacional e discute algumas direções de pesquisa atraentes que podem trazer a análise do crime urbano para uma nova fronteira. O

crime é um evento complexo e multidimensional que ocorre quando a lei, o infrator e o alvo (pessoa ou objeto) convergem dentro do tempo e do lugar, entendendo que os infratores, padrões de comportamento e criminalidade desempenham um papel essencial na compreensão do crime. Consequentemente, é conveniente familiarizar-se às teorias da criminologia. Pois, a criminologia ambiental se concentra nos padrões criminais dentro do ambiente particularmente construído e analisa os impactos das variáveis externas sobre o comportamento emocional das pessoas. Os crimes consistem em espaço (geografia), tempo, lei e infrator, além de alvo ou vítima, como: Teoria da Atividade Rotineira, Teoria do Padrão Criminal, Teoria da Escolha Racional, Teoria da Conscientização, Teoria das Janelas Quebradas, Teoria das Oportunidades do Crime e Teorias Criminais Sociais. As teorias desta família são usadas em diversas abordagens sobre como a teoria do conflito ou perspectiva do conflito estrutural em sociologia estão associadas ao crime. Teorias criminais sociais enfatizam a pobreza, a ausência de educação, a falta de recursos comercializáveis e habilidades e valores subculturais como causas fundamentais de crime, como: Teoria da Desorganização Social, Teoria da Tensão Social, Teoria do Conflito Cultural, Teoria da Eficácia Social, Teoria Subcultural, Teoria do Controle e Teoria da Rotulagem. A predição da taxa de criminalidade opera na prevenção de uma futura porcentagem criminal em uma determinada região urbana. Neste artigo, o autor categoriza modelos de predição da taxa de criminalidade de acordo com os dados que usam como predição, baseada em dados criminais, contexto ambiental e dados de mídia social, como predição baseada em dados criminais, predição baseada em dados de contexto ambiental e predição baseada em dados de mídia social. Em relação ao método deste artigo, será explicado na questão SQ2.

- Análise exploratória de dados e predição de crimes para cidades inteligentes (Pradhan, Potika, Eirinaki, & Potikas, 2019b): analisam-se os principais tipos de crimes e determinam quais atributos contribuem para crimes específicos. Para realizar uma análise exploratória de dados e identificar padrões de crimes, utiliza-se o conjunto de dados criminais de São Francisco. O problema abordado pelos autores foi realizado em uma análise exploratória de dados para identificar padrões de crime e observar os padrões existentes no crime em toda a cidade de São Francisco. Assim, os escritores sugerem melhorias, que determinam as classes de crimes em diferentes áreas da

cidade, e analisam a propagação e o impacto do crime, buscando construir um modelo de predição que trate a classificação multiclasse, isto é, configurando novos crimes em uma das categorias de crime para predizer qual pode ocorrer. Para solucionar o problema de um conjunto de dados desequilibrado, introduzem-se tarefas adicionais de pré-processamento de dados para melhorar a precisão e a necessidade de todas as classes de dados. No aprendizado de máquina/predição, os autores utilizam diferentes tipos de algoritmos, como *K-Nearest Neighbor, Multi-class Logistic Regression, Decision Tree, Random Forest* e *Naïve Bayes*. Os métodos aplicados pelos autores foram *Support Vector Machine (SVM), Random Forest, Gradient Boosting Machines* e Redes Neurais. Antes de aplicar as técnicas de aprendizado de máquina, utiliza-se um pré-processamento de dados. O método deste artigo será explicado na questão SQ2.

- VisCrimePredict: trata-se de um sistema para a predição e a visualização da trajetória do crime a partir de fontes de dados heterogêneas. Nesse sentido, (Morshed et al., 2019) afirmam que os dados multidimensionais abertos — provenientes das redes sociais e de fontes semelhantes — transportam frequentemente informações perspicazes sobre questões sociais. Com o aumento do volume de dados e com a proliferação de plataformas de análise visual, torna-se mais fácil para os usuários interagirem, além de selecionarem informações significativas de grandes conjuntos de dados. Como se observa, a capacidade de visualizar padrões criminais e de predizer incidentes iminentes, com precisão, amplia novas possibilidades na prevenção do crime e, portanto, os autores evidenciam o *VisCrimePredict*, um sistema que utiliza análises visuais e preditivas para mapear crimes ocorridos em uma região/bairro. Para prever crimes futuros, a partir de uma trajetória criminosa, a criminologia depende de uma série de teorias interligadas. Em particular, as teorias do crime ambiental discutem a influência do ambiente no crime e assumem que atores, relativamente racionais, efetuem ações deliberadas com o objetivo de maximizar o seu retorno quando cometem crimes. Nesse sentido, os métodos de predição do crime têm usado uma variedade de métodos estatísticos e de aprendizado de máquina, como aprendizado profundo, análise de regressão, estimativa de densidade de Kernel (KDE), métodos de vetores de suporte e métodos semelhantes, em que todos os modelos dependem da localização geográfica, do tempo e da natureza dos crimes.

Outros estudos mostraram a adequação da Memória Longa e de Curto Prazo LSTM para o desenvolvimento de modelos de trajetória do crime. Para comprovar, os autores utilizam o *Twitter*, abarcando, assim, uma visão geral do *VisCrimePredict*. Primeiramente, é construída uma camada de gerenciamento de dados, em que são coletados dados de código aberto em tempo real e armazenados em um banco de dados NoSQL em formato JSON para facilitar sua manipulação e análise. Também são coletados dados de mídia social em tempo real da API do *Twitter* durante seis meses. Na camada analítica, são processados os dados de mídia social usando ferramentas existentes de Processamento de Linguagem Natural (PNL) para remover frases-chave desnecessárias e extrair recursos, como a natureza dos crimes, localização e horário do incidente no formato JSON. Para conseguir isso, os autores treinaram o modelo de PNL existente uma vez que anotaram, manualmente, dois mil *tweets* relacionados ao crime e usaram esse modelo para extrair apenas os *tweets* relacionados ao crime, em que todas as informações capturadas são integradas aos dados existentes no banco de dados NoSQL. Estes dois conjuntos de dados funcionam como entradas para o componente 'Modelo de Trajetória', que calcula trajetórias de crime usando o algoritmo TbSTS, as quais, ao serem calculadas, são usadas como entradas para o modelo de "Predição de Trajetória". Por fim, o componente de análise visual, desenvolvido com a utilização de *Kepler.js*, apresenta visualmente as trajetórias de crimes gerados, compreendendo a predição de trajetórias em uma região, a natureza de cada crime e informações adicionais obtidas do *Twitter* sobre a região. O método deste artigo será explicado na questão SQ2.

- Uma pesquisa sobre análise e predição do crime (Thomas & Sobhana, 2022): muitas forças policiais, em todo o mundo, adotaram mecanismos que utilizam dados estatísticos para orientar a sua tomada de decisões, denominando-se policiamento preditivo. Nessa abordagem, os departamentos de polícia examinam dados históricos e estatísticos para prever em que áreas geográficas existe maior probabilidade de ocorrência de atividade criminosa. Este tipo de dado é, frequentemente, utilizado pelas autoridades para aplicar eficientemente os seus recursos e impedir o comportamento criminoso. O policiamento preditivo não pode substituir os métodos convencionais de policiamento, mas melhora estas práticas habituais através da aplicação de modelos e algoritmos estatísticos avançados. Alguns estudos empíricos

concluem que estratégias de policiamento preditivo causam uma diminuição na criminalidade. Porém, de modo que a urbanização aumenta a cada dia, é importante observar as atividades criminosas de cada região, procurando reduzir a ocorrência de comportamentos indesejados. A predição de crimes só pode ser feita através da análise dos padrões de atividades criminosas, utilizando os dados anteriores disponíveis dos indivíduos em questão, principalmente os dados históricos, analisando-os a partir do *Deep Learning*, Modelos Estatísticos e Algoritmos. O método deste artigo será explicado na questão SQ2.

- Predizer a hora e a localização de crimes futuros com métodos de recomendação (Y. Zhang, Siriaraya, Kawai, & Jatowt, 2020b): ainda que a detecção de *hotspots* seja eficaz para compreender a distribuição geográfica dos crimes, não considera os aspectos temporais, embora assuma que não há efeito temporal e prediz a ocorrência futura de crimes apenas por intermédio de um mapa de pontos críticos. Como tal, há uma série de estudos que propõem um método integrado de predição do crime espaço-temporal, avaliando a probabilidade de um crime acontecer em um determinado local no dia seguinte. No entanto, isso pode não ser tão útil na prática, uma vez que a polícia teria de patrulhar o local durante todo o dia, caso fosse verificada a elevada probabilidade de ocorrência de um crime. Nesse aspecto, a predição da hora e do local do crime tem sido um tema de investigação potencialmente benéfico tanto para os governos, como para os cidadãos. Destarte, os autores estudam a predição do crime como um problema de recomendação, usando dados criminais abertos e refinados. Definidas as unidades espaço-temporais refinadas, os dados sobre a criminalidade tornar-se-iam muito escassos. Por isso, modelar a predição do crime como um problema de recomendação, no entanto, permite utilizar métodos em sistemas de recomendação que lidam com a escassez de dados. O método deste artigo será explicado na questão SQ2.
- O uso da análise preditiva na predição do crime espaço-temporal: Construindo e testando um modelo em um contexto urbano (Rummens, Hardyns, & Pauwels, 2017): as bases de dados policiais contêm uma grande quantidade de dados criminais que poderiam ser usados para informar sobre tendências e padrões de criminalidade atuais e futuros, visto que a análise preditiva busca otimizar o uso desses dados para antecipar eventos criminais e utilizar métodos estatísticos específicos para predizer a

probabilidade de novos eventos criminais em pequenas unidades espaço-temporais de análise. O objetivo deste estudo é investigar o potencial da aplicação de análise preditiva em um contexto urbano partindo dos conceitos de *Big data* e análise preditiva, os quais, embora sejam relativamente novos em criminologia, se tornaram uma prática padrão em disciplinas como finanças, marketing, inteligência de negócios e ciências biomédicas. No contexto da análise criminal, a grande quantidade de dados criminais disponíveis nas bases de dados policiais pode ser considerada uma fonte valiosa de *Big data* que pode ser utilizada para obter novas perspectivas e conhecimentos úteis sobre tendências e padrões de criminalidade atuais e emergentes. A aplicação de métodos estatísticos avançados para obter essa inteligência a partir de *big data* é comumente chamada de análise preditiva, cujo uso em aplicações criminológicas é frequentemente referido como policiamento preditivo. Na análise prospectiva de *hotspots*, estes são formados não pelas áreas com maior concentração de criminalidade, mas pela agregação de zonas de risco ao redor de cada incidente. Tais zonas de risco são temporárias, tornando os *hotspots* mais dinâmicos. Outro desenvolvimento é a modelagem de terreno de risco, que cria um mapa de risco de locais sensíveis a altas taxas de criminalidade, com base em suas propriedades espaciais e em suas interações. Dessa forma, tanto a análise prospectiva de pontos críticos quanto o RTM resultam em um mapa de calor que mapeia o risco de crime para cada área, garantindo a obtenção de uma imagem mais dinâmica das tendências futuras da criminalidade. O policiamento preditivo pode, portanto, ser considerado um passo à frente na evolução do mapeamento da criminalidade devido ao seu foco específico nas predições espaço-temporais da criminalidade, permitindo assim uma estimativa mais precisa dos padrões futuros. Este processo está totalmente alinhado à evolução recente do uso de níveis micro geográficos de análise na pesquisa criminológica para explicar a distribuição desigual do crime, sendo sugerido como o novo padrão no mapeamento e análise do crime. O nível micro geográfico é considerado mais adequado e preciso, pois reflete melhor a variabilidade existente nesse nível tanto da criminalidade, como das variáveis socioeconômicas, e fornece padrões de criminalidade mais previsíveis em comparação com unidades geográficas de análise mais elevadas, como setores censitários, bairros ou distritos. O objetivo deste estudo foi explorar o potencial de aplicação da análise preditiva em um contexto

urbano, especificamente em uma grande cidade (população > 250.000) na Bélgica, explorando as possibilidades e limites deste método na criminologia espacial. O método deste artigo será explicado na questão SQ2.

- Usando modelagem de terreno de risco para prever crimes relacionados a moradores de rua em Los Angeles, Califórnia (Yoo & Wheeler, 2019): apresenta a *Risk Terrain Modeling* (RTM) aplicada à predição de áreas espaciais de alto risco de crimes relacionados com moradores de rua em Los Angeles, Califórnia. As populações sem-abrigo, cujo número de indivíduos tem aumentado nos Estados Unidos nos últimos anos, com um total estimado de mais de meio milhão, em 2017, apresentam desafios significativos para a aplicação da lei. O. A resolução dos problemas relacionados com os sem-abrigo está, intrinsecamente, relacionada com o policiamento e com o ambiente construído já que as diferentes estruturas exteriores fornecem abrigo temporário para indivíduos transitórios (como pontes), e estes fatores são fixos no espaço. Uma vez que esses fatores espaciais se aplicam igualmente a todos os sem-abrigo, pode-se prever adicionalmente que funcionam igualmente para predizer a infração ou a vitimização dos sem-abrigo. Acredita-se, também, que a aplicação da RTM aos crimes relacionados com os sem-abrigo pode ajudar a informar futuras estratégias de justiça criminal de várias maneiras. Devido a isso, uma delas é que, ao identificar os potenciais fatores espaciais que contribuem para um maior risco de criminalidade entre os sem-abrigo, a RTM pode orientar o policiamento para os problemas ou para a prevenção da criminalidade por meio de estratégias de concepção ambiental para mitigar esse risco. Em segundo lugar, as repressões policiais anteriores contra distúrbios relacionados aos sem-abrigo, centradas no *Skid Row*, apenas resultaram em pequenas reduções da criminalidade e, provavelmente, causaram o subsequente deslocamento daqueles indivíduos. A RTM, desde a concepção, pode identificar áreas em que há provável deslocamento do crime e, assim, sugerir estratégias proativas para abordar simultaneamente locais de alto risco, que vão além das estratégias típicas de policiamento em pontos críticos. Isso sugere que estratégias espaciais específicas podem reduzir tanto a infracção dos sem-abrigo como o risco de vitimização, haja vista que a maioria dos indivíduos sem-abrigo é apenas de forma intermitente, e as estratégias baseadas no local podem ser uma forma mais eficaz de

limitar o risco do que as que se concentram nos indivíduos. O método deste artigo será explicado na questão SQ2.

- Predição de rotações de culturas usando técnicas de mineração de processos e princípios de Markov (Dupuis, Dadouchi, & Agard, 2022): Satisfazer uma procura crescente de alimentos e, ao mesmo tempo, preservar o ambiente é um dos desafios mais importantes do século XXI. Para enfrentar tal impasse, a agricultura de conservação pode contar com a antiga prática da rotação de culturas. Acerca disso, o objetivo do artigo é desenvolver uma metodologia de predição e visualização de rotações de culturas, apoiando discussões entre agrônomos e produtores. Com base nos dados históricos das culturas, a metodologia de seis fases utiliza cadeias de Markov para a predição das N culturas mais prováveis cultivadas no ano n + 1, sendo que a mineração de processos e os grafos DFG permitem a modelagem e a visualização dos resultados, e as operações de generalização e filtragem destacam os comportamentos frequentes dos produtores. Aplicada para analisar o histórico de cultivo de 10.376 campos de 409 fazendas de cultivo em Quebec, Canadá, a metodologia é competitiva com o desempenho de várias redes neurais recorrentes (LSTM, RNN, GRU), com uma taxa de predição de sucesso que excede 90%, ao mesmo tempo que permite uma inteligibilidade dos resultados e uma relativa simplicidade computacional. Este artigo não apenas lida com a previsão em um setor crítico (a agricultura), mas também emprega métodos preditivos que podem ser, analogicamente, aplicados ao contexto urbano de predição de crimes. O método, descrito no artigo, utiliza cadeias de Markov e mineração de processos para prever a rotação de culturas, alcançando uma taxa de sucesso superior a 90% em comparação com modelos complexos de redes neurais. Sobre a respectiva análise, o presente estudo visa extrair *insights* sobre a aplicabilidade de tais métodos estatísticos e computacionais para o problema da previsão de crimes, aproveitando a alta precisão, inteligibilidade dos resultados e eficiência computacional que essas técnicas oferecem. Tal transposição de metodologias de um campo para outro ilustra a versatilidade e o potencial de técnicas preditivas baseadas em dados quando aplicadas a desafios complexos e multifacetados, como a predição de crimes urbanos.
- Predição de eventos criminais com recursos dinâmicos (Rumi, Deng, & Salim, 2018b): As *Location-Based Social Networks* (LBSN) recolhem uma vasta gama de

informações, que podem ajudar a compreender a dinâmica regional — a mobilidade humana —, em toda uma cidade, e oferecem oportunidades sem precedentes para enfrentar vários problemas sociais. Neste artigo, exploraram-se características dinâmicas derivadas de dados de *check-in* do *Foursquare* na predição de eventos criminais de curto prazo com granularidade espaço-temporal fina. Embora a predição de eventos criminais tenha sido amplamente investigada devido à sua importância social, a taxa de sucesso está longe de ser satisfatória. Estudos existentes baseiam-se em características relativamente estáticas, como características regionais, informações demográficas e tópicos obtidos a partir de *tweets*, mas poucos estudos centram-se na exploração da mobilidade humana por meio das redes sociais. Neste artigo, os autores identificam uma série de características dinâmicas com base nos resultados da pesquisa em Criminologia e relatam suas correlações com diferentes tipos de eventos criminais. Em particular, observa-se que alguns tipos de eventos criminais estão mais correlacionados com as características dinâmicas — roubo, crime relacionado com drogas, fraude, entrada ilegal e agressão — do que outros, como crime relacionado com o trânsito. Um desafio-chave do artigo é o fato de que a informação dinâmica é muito escassa em comparação com a informação relativamente estática. Para resolver esse problema, desenvolveu-se uma abordagem baseada em fatoração matricial para estimar as características dinâmicas ausentes em toda a cidade. Curiosamente, as características dinâmicas estimadas ainda mantêm a correlação com a ocorrência de crimes em diferentes tipos. Após avaliados os métodos propostos em diferentes intervalos de tempo, os resultados verificaram que o desempenho da predição do crime pode ser, significativamente, melhorado com a inclusão de características dinâmicas em diferentes tipos de eventos criminais. O método deste artigo será explicado na questão SQ2.

- Predição do crime de furtos de ônibus em Pequim, China: a qualidade do ar afeta o crime? (Ding & Zhai, 2019): com o desenvolvimento do sistema de transporte público urbano, os ônibus tornaram-se uma parte indispensável das atividades diárias das pessoas e, nesse contexto, os batedores de carteira — os ladrões — aproveitam roubar itens das roupas e das bolsas das pessoas enquanto elas estão no interior do ônibus. De tal forma que as estações e rotas de transportes coletivos são áreas de tráfego intenso e os passageiros possuem pouca consciência de segurança, os batedores de carteira

podem escapar facilmente após um crime, e a maioria das vítimas teme retaliação se for confrontada. Essas questões dificultam o controle dos furtos de ônibus, embora o crime afete seriamente a segurança pessoal e a propriedade dos passageiros. O artigo examinou o impacto de fatores ambientais nas taxas de criminalidade nas estações de ônibus e a diferença entre crimes estáticos e não estáticos em locais públicos. Ainda, neste artigo, é apresentada uma estrutura completa e clara, incluindo a descoberta de padrões com base em dados, da aplicação da teoria criminológica para explicar padrões e o uso de método de investigação empírica para verificar a explicação teórica. Descobriu-se que a temperatura e a estação do ano não estão claramente correlacionadas com os furtos de ônibus. Os índices AQI consideram várias concentrações de poluentes atmosféricos, como partículas (PM10 e PM2.5), dióxido de enxofre (SO2), monóxido de carbono (CO), dióxido de nitrogênio (NO2) e ozônio (O3). O PM2,5 se refere a partículas em suspensão no ar que têm um diâmetro menor do que 2.5 micrômetros, o que é cerca de 3% do diâmetro de um fio de cabelo humano. Estas partículas são tão pequenas que podem ser inaladas, profundamente, nos pulmões e até mesmo entrar na corrente sanguínea, causando uma variedade de problemas de saúde, desde irritações respiratórias até efeitos mais graves, como doenças cardíacas e pulmonares. Em seguida, foram realizadas duas investigações empíricas para verificar se a teoria dos padrões de crime e a teoria da escolha racional podem ser usadas para explicar o impacto da qualidade do ar no crime de furto de ônibus. O método SVM foi usado para predizer o risco diário de furtos em ônibus. O método deste artigo será explicado na questão SQ2.

- Melhorar a predição da criminalidade a curto prazo com fluxos de mobilidade humana e arquiteturas de aprendizagem profunda (J. Wu, Abrar, Awasthi, Frias-Martinez, & Frias-Martinez, 2022): os modelos de predição de crimes de curto prazo, baseados em locais, aproveitam os padrões espaço-temporais de crimes históricos para prever volumes agregados de incidentes criminais em locais específicos ao longo do tempo. Acerca disso, a teoria da oportunidade do crime sugere que a mobilidade humana pode desempenhar um papel na geração do crime e, por isso, dá-se mais atenção ao poder preditivo da mobilidade humana em modelos de criminalidade de curto prazo baseados no local. Para isso, os pesquisadores usaram registros de detalhes de chamadas (CDR) e dados de serviços baseados em localização, como o *Foursquare*,

ou de mídias sociais, para caracterizar a mobilidade humana, e mostraram que as métricas de mobilidade, juntamente com os dados históricos da criminalidade, podem melhorar a precisão da predição da criminalidade a curto prazo. Neste artigo, propõe-se usar um conjunto de dados de mobilidade humana refinada, disponível publicamente, de uma empresa de inteligência de localização, para explorar os efeitos dos recursos de mobilidade humana na predição de crimes em curto prazo. Para esse efeito, realiza-se uma avaliação abrangente em múltiplas cidades com diversas características demográficas, diferentes tipos de crimes e vários modelos de aprendizagem profunda, mostrando que adicionar recursos de fluxo de mobilidade humana a crimes históricos pode melhorar as pontuações F1 para uma variedade de modelos neurais de predição de crimes em cidades e tipos de crimes, com melhorias variando de 2% a 7%. A análise também mostra que algumas arquiteturas neurais podem melhorar ligeiramente o desempenho da predição de crimes quando comparadas com modelos de regressão não neural em no máximo 2%. O método deste artigo será explicado na questão SQ2.

- Algoritmos de aprendizado de máquina para predição de crimes sob o Código Penal Indiano (Aziz, Sharma, & Hussain, 2022a): propõe uma abordagem baseada em dados para extrair conhecimento perspicaz dos dados criminais indianos. A abordagem proposta pode ser útil para a polícia e para outros órgãos de aplicação da lei, na Índia, controlarem e prevenirem o crime em toda a região. Na abordagem proposta, diferentes modelos de regressão são construídos com base em diferentes algoritmos de regressão, a saber, regressão “florestal aleatória” (RFR), regressão de árvore de decisão (DTR), regressão linear múltipla (MLR), regressão linear simples (SLR) e suporte regressão vetorial (SVR) após pré-processamento dos dados usando MySQL Workbench e programação R. Esses modelos de regressão podem predizer 28 tipos diferentes de crimes conhecidos do IPC e também um número total de crimes conhecidos do Código Penal Indiano (IPC) em termos regionais, estaduais e anuais (para todo o país). Técnicas de visualização de dados, nomeadamente diagramas de cordas e mapas, são utilizadas para visualizar dados pré-processados — correspondentes aos anos de 2014 a 2020 — e dados previstos pelo modelo de regressão relativamente melhor para o ano de 2022. Para os dados escolhidos, conclui-se que a RFR, que prediz o crime total conhecido pelo IPC, ajusta-se

relativamente melhor, com um valor de *R2* ajustado de 0,96 e um valor MAPE de 0,2. Ainda, entre os modelos de regressão que preveem a contagem de crimes de roubo em toda a região, a RFR é o modelo baseado em regressão que se ajusta melhor, com um valor *R2* ajustado de 0,96 e um valor MAPE de 0,166. Estes modelos de regressão prevêem que o estado de Andhra Pradesh terá a maior contagem de crimes, posicionando o distrito de Adilabad no topo, com 31.933 contagens de crimes previstos. O método deste artigo será explicado na questão SQ2.

- Sistema de identificação de escritor para manuscrito offline pré-segmentado Caracteres Devanagari usando k-NN e SVM (Dargan, Kumar, Garg, & Thakur, 2019): um sistema de identificação biométrica baseado em modalidades únicas e múltiplas tem sido um conceito em evolução para resolver questões criminais, de segurança e manutenção da privacidade para verificar a autenticação de um indivíduo. O sistema de identificação do escritor é um tipo de identificação biométrica em que a caligrafia de um indivíduo é considerada como um identificador biométrico no qual o escritor pode ser identificado com base no seu texto manuscrito. Esses sistemas empregam algoritmos de aprendizado de máquina e reconhecimento de padrões para a geração de uma estrutura. Neste artigo, os autores apresentaram um novo sistema para a identificação do escritor baseado nos personagens pré-segmentados da escrita *Devanagari* e também apresentando um trabalho abrangente e de última geração. O experimento foi realizado em corpus composto por cinco cópias de cada personagem da escrita *Devanagari* grafadas por cem escritores diferentes, selecionados aleatoriamente em locais públicos, cujo experimento foi composto por um total de 24.500 amostras de caracteres *Devanagari*. Quatro metodologias de extração de características (baseadas em zoneamento, diagonal, transição e extensão de pico) e métodos de classificação, como k-NN e SVM linear, são usados com precisão de identificação de 91,53% ao usar recursos baseados em zoneamento, transição e extensão de pico com um linear Classificador SVM. O método deste artigo será explicado na questão SQ2.

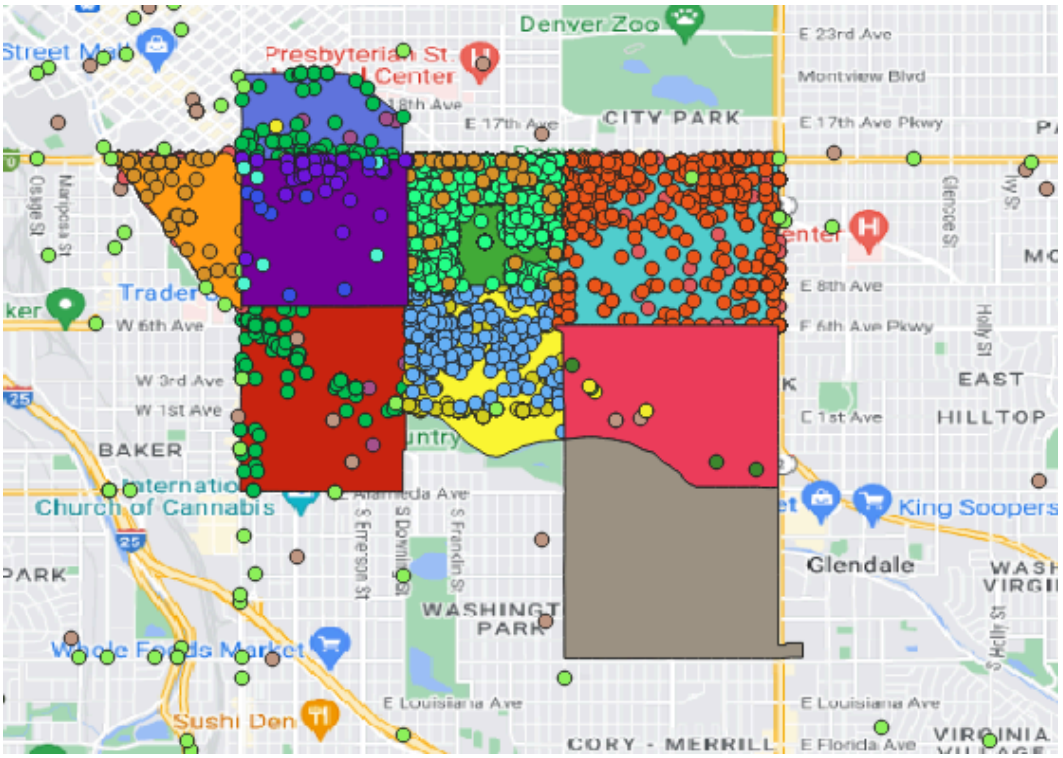
A predição de crimes/roubos em ambiente urbano é uma tarefa complexa e essencial para a segurança pública. Para abordar essa questão, é fundamental coletar e analisar dados históricos relacionados a crimes e roubos. Conforme observado por (Dupuis, Dadouchi e

Agard, 2022), em seu estudo, dados consolidados em um banco de dados são essenciais para uma análise precisa, pois incluem informações detalhadas sobre cada incidente, como localização, horário, data, tipos de crimes e outros fatores relevantes.

Além disso, para uma visualização mais eficaz da predição de crimes/roubos em ambientes urbanos, recorre-se ao uso de um software de geoprocessamento de código aberto: o QGIS. Isso permite mapear os diferentes tipos de crimes, tornando a percepção desses eventos mais intuitiva para análise e para compreensão. No processo de implantação do banco de dados no programa QGIS, selecionou-se a região central da cidade de Denver, no estado do Colorado nos Estados Unidos da América, como a área de foco.

Posteriormente, foram divididos os tipos de crimes por bairro na região central da cidade. O resultado desse processo pode ser visualizado na imagem a seguir:

Figura 3 - Exemplo de crimes do *software* QGIS.



Fonte: A autora.

Ao observar a *Figura 3*, obtém-se uma visão clara dos nove bairros na região central da cidade de Denver e cada um representado por uma cor distinta: Belcaro (cinza), Capitol Hill (roxo), Cheesman Park (verde), Cherry Creek (rosa), Civic Center (laranja), Congress

Park (azul claro), Country Club (amarelo), North Capitol Hill (azul escuro) e Speer (vermelho).

Em decorrência da diversidade dos tipos de crime, uma ampla gama de cores foi utilizada para garantir a clareza na distinção entre eles. Cada círculo colorido no mapa possui uma localização geográfica específica, permitindo identificar as áreas exatas em que ocorreram os crimes. Tais círculos podem estar dispersos em várias partes de um bairro, de modo a destacar a heterogeneidade da distribuição dos crimes, além de possibilitar a identificação de áreas de risco e a alocação mais eficiente de recursos policiais. Nesse aspecto, é relevante notar que os tipos de crimes citados também estão presentes em outros bairros da cidade, embora o estudo esteja centrado na área central. Assim, a *Figura 3* foca em mapear os tipos de crimes na região central, oferecendo *insights* valiosos para a compreensão da distribuição espacial desses eventos na cidade de Denver.

Esta abordagem integrada, que combina análise de dados históricos detalhados para a visualização, fornece uma base importante para a predição de crimes e de roubos em ambientes urbanos. A compreensão aprofundada dos padrões de ocorrência de crimes é crucial para o desenvolvimento de modelos de predição mais precisos, que podem auxiliar as autoridades na prevenção e na resposta a incidentes.

Além disso, a temporalidade dos dados pode ser usada para discernir tendências ao longo do tempo, permitindo previsões dinâmicas que se ajustam às mudanças de padrão e comportamento dos criminosos. Combinando técnicas de visualização com algoritmos avançados de mineração de processos e modelos estatísticos, pesquisadores e autoridades também buscam implementar medidas preventivas baseadas em evidências para aumentar a segurança pública.

2.4.2. Métodos de predição são usados para predizer roubos em ambiente urbano (SQ2)

A questão SQ2 aborda a aplicação de métodos estatísticos na predição de roubos em ambientes urbanos. Esta é uma área de pesquisa de extrema importância, uma vez que métodos de predição desempenham um papel fundamental na prevenção e na resposta a crimes, contribuindo para a segurança pública.

Desse modo, os resultados da SQ2 destacam a aplicação de uma variedade de métodos para predizer roubos em ambiente urbano. Como se evidencia, a análise de dados é

uma técnica amplamente adotada em todo o mundo para identificar padrões e tipos de crimes/roubos, possibilitando a aplicação de medidas preventivas eficazes.

A síntese dos artigos relacionados no Mapeamento Sistemático da Literatura (MSL) com a questão SQ2 se concentra, principalmente, em um tópico central: os artigos que foram analisados na SQ2 ofereceram *insights* detalhados sobre os métodos empregados na predição de roubos em ambiente urbano. Isso inclui desde uma exploração de métodos tradicionais, como análise estatística e regressão logística, até abordagens de vanguarda, como aprendizado de máquina, análise geoespacial e inteligência artificial. Cada método foi discutido em termos de suas vantagens, de seus desafios e de sua eficácia na predição de roubos.

Essa abordagem permite que os leitores do MSL compreendam a diversidade de métodos de predição disponíveis e como eles podem ser aplicados na predição de roubos em ambientes urbanos. Além disso, essa análise fornece uma base sólida para a seleção de métodos apropriados com base nos objetivos da pesquisa e nas características dos dados disponíveis.

Essa elaboração amplia a resposta à questão SQ2, visto que destaca a importância dos métodos na predição de roubos em ambientes urbanos e fornece uma visão abrangente sobre os diferentes métodos disponíveis.

- Análise e predição de relatórios criminais em Bangladesh (Pavel Rahman et al., 2021): propõe-se um método baseado em dados, em que todos aqueles considerados desnecessários fossem removidos. Pois, a técnica aplicada foi a regressão linear para treinar e classificar os dados para ACLED, no intuito de prever diferentes tipos de eventos — brigas, explosões, protestos, motins, desenvolvimentos estratégicos, violência contra civis com geolocalização dos eventos. (Pavel Rahman et al., 2021) usou vários modelos de aprendizado de máquina, como *Random Forest*, MLP e *Decision Tree* para predizer o ano de 2019. As estatísticas criminais são coletadas no site da Polícia de Bangladesh para analisar e predizer *dacoity*, roubo, assassinato, repressão de mulheres e crianças, sequestro, roubo, furto e outros crimes. A regressão com múltiplas saídas, como regressão linear, KNN e de decisão, foi usada para predizer longitude e latitude. Vários algoritmos de aprendizado de máquina, como árvore de decisão, “floresta aleatória”, SVM e *perceptron* multicamadas, foram usados para o modelo. Para isso, os autores extraíram os dados por meio da

ferramenta PNL e “palavra-gatilho”, e um classificador BUH foi usado para pré-processar os dados, os quais foram coletados de 2010 a 2018, além de alguns dados de 2019. O pré-processamento foi feito nesses dados, com o uso de um codificador de rótulo, normalização e, posteriormente, divididos em dados de treinamento e teste e usados para predição. (Pavel Rahman et al., 2021) previram que haveria 262 bandidos, 562 roubos e 3.830 assassinatos em 2019.

- Mitigando vulnerabilidades por meio de predições e análises de tendências criminais (Orong, Sison, & Hernandez, 2018): o artigo utilizou o algoritmo de agrupamento *k-means* e o modelo de média móvel integrada autoregressiva (ARIMA) para agrupar e prever os dados criminais indexados, respectivamente. O método de mineração consiste em seis etapas, que incluem: compreensão do negócio, compreensão dos dados, preparação de dados, modelagem, avaliação e implantação. Desse modo, a predição da criminalidade, a curto prazo, contrasta-se com a precisão da predição de modelos de séries temporais univariadas com métodos simples, habitualmente, efetuados pela polícia.
- Crime em áreas urbanas: uma perspectiva de mineração de dados (X. Zhao & Tang, 2018): como os autores não possuem um método, revisam algoritmos de última geração para vários tipos de tarefas de crime computacional, entre eles: (I) Detecção de Pontos Críticos de Crime: é um mapeamento de técnica espacial com foco na identificação da concentração de eventos criminais em toda a cidade. Os métodos de detecção de pontos críticos de crime, com base nos tipos de técnicas do KDE, são: (a) um método não paramétrico para calcular a probabilidade da função de densidade de crimes; (b) técnica baseada em reação-difusão: uma estrutura matemática baseada na reação-difusão de equações diferenciais parciais para aprender a dinâmica dos focos de crime; e (c) outras técnicas: incluem mapeamento temático de limites geográficos, grade temática, elipses espaciais e otimização de pontos de acesso. (II) Predições do próximo local e a incidência de crimes urbanos: o crime é investigado com o objetivo de auxiliar estudos futuros sobre predição inteligente de crimes. (X. Zhao & Tang, 2018) utilizam a fórmula de Rossmo, que é introduzida para prever a próxima localização do crime com dados geográficos e com técnicas de perfil. Em outras palavras, uma rede de tráfego é incorporada ao perfil geográfico e o próximo local de predição é tratado na rede de tráfego ponderado, no qual o caminho mais curto entre

os nós é aproveitado para substituir a distância euclidiana pela localização da predição do próximo crime, ou seja, oferece uma técnica para modificar padrões espaço-temporais como consequência de algumas atividades especiais *a priori*. (III) Análise de rede criminosa: categorizar os métodos de análise de redes criminosas de acordo com as técnicas que empregam, como as técnicas baseadas em agentes que se movem em rede e interagem entre si. Assim, podem-se produzir dinâmicas e padrões complexos a partir de simples comportamentos, baseados em regras e em técnicas, de acordo com a teoria de grafos, cujas métricas e técnicas são empregadas para análise de redes criminosas, além de técnicas baseadas em GIS — Informação Geográfica Técnicas. (IV) Vitimização quase repetida: o crime não acontece de forma aleatória ou uniforme ao longo do tempo ou espaço, mas a referida tarefa computacional indica o aumento do risco de vitimização repetida na mesma região ou em regiões vizinhas dentro de um determinado período de tempo. (V) Patrulhamento Policial: planejamento adequado da rota de patrulha é uma aplicação importante dos sistemas de análise criminal, uma vez que ajuda a aumentar a eficácia do patrulhamento policial e a melhorar, simultaneamente, a segurança pública.

- Análise exploratória de dados e predição de crimes para cidades inteligentes (Pradhan, Potika, Eirinaki, & Potikas, 2019b): aplicam-se os métodos de classificação: *Support Vector Machines* (SVM), *Random Forests* (RF), *Neural Networks* (NN), *Maximum Entropy Classifier* (MAXENT) e *Scaled Linear Discriminant Analysis* (SLDA), buscando prever o tipo de crime que pode ocorrer com base no local e no horário determinados. No entanto, os autores observaram que as SVM realizaram predições consistentemente melhores e destacaram que os dados foram balanceados para evitar resultados distorcidos. Antes de aplicar as técnicas de aprendizado de máquina, para predizer a categoria do crime, realizaram-se pré-processamentos de técnicas, como transformação de dados, discretização, limpeza e redução. Usaram, ainda, dois dados conjuntos: no primeiro caso, um grupo demográfico com informações utilizadas sem alterações, com técnicas *Random Forest* ou *Gradient Boosting*, que acarretou resultados melhores. No segundo caso, utilizaram-se dados para predizer os valores ausentes nos dados originais definidos, com técnicas SVM e Redes Neurais, que mostraram resultados promissores.

- VisCrimePredict: trata-se de um sistema para predição e visualização da trajetória do crime a partir de fontes de dados heterogêneas. Nesse sentido, (Morshed et al., 2019) propõem um novo algoritmo que cria trajetórias a partir de fontes de dados heterogêneas, como dados abertos e mídias sociais, com o objetivo de relatar incidentes criminais. *VisCrimePredict* usa um algoritmo *Long Short Term Memory* (LSTM) para predição de trajetória, criando uma implementação de prova de conceito do *VisCrimePredict* e uma avaliação experimental da precisão da predição da trajetória do crime, usando a rede neural LSTM. Diante das análises dos autores, tem-se uma implementação do *VisCrimePredict*, o qual exibe um sistema para análise visual de dados multidimensionais nos níveis macro e microscópico para mostrar trajetórias de crimes com base em suas características espaciais e temporais. Tal sistema *VisCrimePredict* incorpora um novo algoritmo de Segmentação Temporal Espacial, baseado em Limites (TbSTS, como a janela deslizante) para a segmentação da trajetória do crime, e usa o algoritmo LSTM, uma variante da rede neural recorrente para predição de trajetória.
- Uma pesquisa sobre análise e predição do crime (Thomas & Sobhana, 2022): os métodos são categorizados e estuda-se sua eficácia em diversas áreas com base na precisão e na exatidão em sua predição, de modo a destacar as metodologias existentes e a necessidade de futuros desenvolvimentos. Para isso, os autores utilizam métodos de análise e predição, comumente usados, e agrupam as abordagens nas seguintes categorias, como abordagens de (I) redes neurais, (II) abordagens estatísticas e (III) abordagens espaço-temporais, estudando o nível de precisão mostrado por alguns dos diferentes métodos. (I) Redes Neurais: o aprendizado profundo é uma categoria de aprendizado de máquina que lida com algoritmos inspirados pela composição e função do cérebro — redes neurais artificiais. Além disso, possui redes capazes de aprender sem supervisão e regressão “florestal aleatória” com a finalidade de prever as tendências e padrões de crimes. Nos dados de treinamento para a predição do crime, são realizados testes com diferentes modelos de regressão, como: Regressão linear, *Random Forest*, autorregressão e árvore de decisão. Outros modelos utilizam diferentes algoritmos de aprendizado de máquina, como a regressão logística, o vetor de suporte máquina (SVM), *Naïve Bayes*, k-vizinhos mais próximos (KNN), árvore de decisão, *perceptron* multicamadas

(MLP), floresta aleatória, *eXtreme Gradient Boosting* (XGBoost), análise por LSTM e média móvel integrada autoregressiva (ARIMA) para predizer crimes. (II) Abordagens Estatísticas: para analisar os registros do crime, no intuito de descobrir as informações sobre diversas perspectivas de um evento criminal, dados e técnicas de mineração foram sugeridas, por exemplo: regras de associação, classificação, regressão e agrupamento. Esses métodos têm sido amplamente utilizados junto com outros modelos, como LSTMs e RNNs; método de classificação, como algoritmo Gradient Boosting; método de seleção: baseado em árvores e florestas; e algoritmos de aprendizado de máquina como o KNN. (III) Abordagens Espaço-temporais: esta é uma metodologia de predição, que tem a capacidade de fazer uso de toda a informação acessível coletada de diversos locais. Como resultado, têm-se grandes conjuntos de dados diversos, e a predição espaço-temporal do crime que se torna muito importante. Este método é, geralmente, implementado em cidades inteligentes, e estudos dão a ideia de que ocorrências dos eventos criminais não são distribuídas, uniformemente, dentro de uma cidade.

- Predizer a hora e a localização de crimes futuros com métodos de recomendação (Y. Zhang, Siriaraya, Kawai, & Jatowt, 2020b): discute como a predição do crime pode ser modelada como um problema de recomendação, mais especificamente como os fatores espaço-temporais podem ser modelados como usuários e itens. Ao modelar o crime dessa forma, as técnicas, utilizadas nos sistemas de recomendação, podem ser aplicadas de forma eficaz na predição do crime, a qual é considerada, neste artigo, como um problema de recomendação. Com base na teoria da atividade rotineira, os autores de crimes associam suas atitudes ilícitas a uma oportunidade que está presente. Acerca disso, o crime que ocorre em uma dada área e hora é determinado por dois fatores: a presença de criminosos nesta área, que é propriedade da área, e a rotina diária das pessoas nessa hora (deslocamento). Outrossim, os autores descobriram que esses dois fatores podem ser modelados adequadamente como um problema de recomendação. Em um sistema de recomendação típico, existem três componentes: usuários, itens e interações usuário-item. Esta última é, geralmente, considerada como uma avaliação dada por um usuário para um item ou um registro de compra. Os métodos usados em sistemas de recomendação que mitigam a esparsidade são, portanto, adequados para o contexto refinado de predição de crimes. Como se

verifica, a fatoração de matrizes usa duas matrizes de baixa dimensão para usuários e itens para aproximar a tabela de classificação original, mas a fatoração matricial tradicional, porém, não incorpora informações contextuais. A extensão deste artigo permite a fatoração de matrizes para incorporar o contexto social para um determinado local e horário, obtido a partir de postagens georreferenciadas em mídias sociais, como o *Twitter*. Dessa maneira, os resultados experimentais mostram que o método supera os métodos recomendados e não recomendados. No experimento, comparou-se a abordagem com três linhas de base não recomendadas, a saber, ARIMA, VAR e KDE, bem como cinco métodos de recomendação, tais quais, CF baseado em item (CF-item), CF baseado em usuário (CF-usuário), código binário (BC), decomposição tensorial (TD) e fator oculto como tópicos (HFT). Para testar o efeito da adição de viés de usuário — item e contexto —, também se testou o método MF original e o BMF.

- O uso da análise preditiva na predição do crime espaço-temporal: Construindo e testando um modelo em um contexto urbano (Rummens, Hardyns, & Pauwels, 2017): para este fim, os dados criminais, disponíveis para três tipos de crime — assalto a residências, roubo na rua e agressão —, são agregados espacialmente em *grids* de 200 por 200m e analisados retrospectivamente. Nesse sentido, é aplicado um método sintetizando os resultados de um modelo de regressão logística e de rede neural, resultando em predições quinzenais para 2014, com base em dados de criminalidade dos três anos anteriores, sendo que predições mensais desagregadas temporalmente (predições diurnas versus noturnas) também foram feitas. A qualidade das predições é avaliada com base nos seguintes critérios: taxa de acerto direto, precisão e índice de predição. Já os resultados indicam que é possível obter predições funcionais, aplicando análise preditiva aos dados criminais em nível de rede. As predições mensais com distinção entre dia e noite produzem melhores resultados globais do que as predições quinzenais, indicando que a resolução temporal pode ter um impacto importante no desempenho da predição. A análise preditiva é aplicada aos dados criminais disponíveis, e o seu desempenho preditivo é avaliado. Estes modelos também apresentam a vantagem de serem seguramente documentados em outros domínios em que a análise preditiva é aplicada. Foram escolhidos os assaltos a residências e os roubos nas ruas por serem crimes de alto impacto priorizados pelos

serviços policiais locais. Uma bateria de testes foi incluída para permitir uma comparação entre crimes violentos e crimes contra a propriedade, sendo que a comparação entre o dia e a noite é incluída, primeiro, porque muitas vezes há uma necessidade prática de mapas diferenciados dependendo dos turnos policiais e, segundo, para testar a suposição intuitiva de que existem diferenças importantes nos padrões de criminalidade que poderiam influenciar o desempenho da predição. A abordagem utilizada, neste artigo, baseia-se em *Crime Anticipation System* (CAS), desenvolvido pelo Departamento de Polícia de Amsterdã, na Holanda. O CAS recolhe dados históricos sobre mais de 200 variáveis, que consistem em indicadores de criminalidade, demográficos, socioeconômicos e de uso da terra/espaço. Essas variáveis são modeladas usando redes neurais (RN), e as células risco 3% identificadas pelo modelo são, então, mapeadas usando uma grade com resolução de 125 por 125 m.

- Usando modelagem de terreno de risco para prever crimes relacionados a moradores de rua em Los Angeles, Califórnia (Yoo & Wheeler, 2019): os dados deste estudo constroem os modelos RTM com base em crimes relatados entre moradores de rua, de 2013 a 2017, e usam dados de 2018 para testar a precisão dos modelos. Para a amostra de 2018, há um total de 6.307 incidentes criminais, os quais foram analisados coletivamente, mostrando que os fatores de risco identificados tendem a ser muito semelhantes em cada um dos quatro modelos. Os resultados da seleção obtidos do modelo da análise RTM, aplicada aos quatro resultados diferentes, são: violência contra os sem-abrigo, violência cometida pelos sem-abrigo, crimes contra a propriedade e contra os sem-abrigo e crimes contra a propriedade cometidos pelos sem-abrigo. Cada tipo de risco indica a operacionalização selecionada, seja como densidade (D) ou proximidade (P), juntamente com o limite de distância que resultou no modelo de melhor ajuste. Nesse sentido, os modelos foram combinados, visto que os *logs* do risco relativo previsto e a pontuação *z* do valor registrado se somaram aos *scores z* e, depois, divididos por quatro. Comparado com o mapa de densidade do *kernel* original, isso fornece uma visão mais detalhada de predição de risco com base nos geradores de crime de entrada. Utilizando o RTM, os autores identificaram vários fatores de risco diferentes que estavam associados à criminalidade relacionada com os sem-abrigo. Diante disso, os fatores de risco mais fortes para predizer a criminalidade,

detenções anteriores por drogas e proximidade de abrigos para sem-abrigo são consistentes tanto nos tipos de crimes contra a propriedade, como nos crimes violentos e na infração ou vitimização dos sem-abrigo. Ademais, os mapas do risco de criminalidade dos sem-abrigo também ilustraram que, embora a criminalidade relacionada a eles esteja concentrada em bairros pobres, há várias áreas ao redor de Los Angeles que apresentam um grande número de incidentes de crimes contra pessoas em situação de rua, sendo previstos como alto risco de crimes, de acordo com os modelos RTM. Assim, discutiu-se a importância dessas descobertas em termos de avanço da teoria do crime em micro locais, bem como as aplicações potenciais das descobertas.

- Predição de rotações de culturas usando técnicas de mineração de processos e princípios de Markov (Dupuis, Dadouchi, & Agard, 2022): a predição da rotação de culturas, para um ano real *(n),* é um problema amplamente abordado na literatura. O objetivo desta dissertação é predizer, para cada tipo de crime, qual será o mais frequente do próximo ano (n + 1), utilizando um histórico de crimes disponibilizado pela polícia. Estas predições para o ano seguinte *(n + 1)* e a visualização dos elementos que levaram a essas predições podem subsidiar discussões estratégicas entre polícia e governo, além de facilitar a implementação de medidas mais preditivas para os próximos crimes, otimizando o uso da tecnologia em prol da redução de crimes. O método consiste em seis fases que levam a um predição do tipo de cultura a ser cultivada em um campo no próximo ano *(n + 1)* e a um grafo que representa as relações entre as culturas, baseado em técnicas de mineração de processos e Markov princípios. Fase I: Preparação dos Dados: Em primeiro lugar, os dados devem ser formatados de modo que, para cada registro, exista um identificador único do caso, da atividade e do marcador de tempo associado. Se houver muita variabilidade nas atividades, alguma fusão pode ser realizada. No exemplo, notamos a existência das atividades 'A1' e 'A2', cujos dados são generalizados, substituindo cada ocorrência de 'A1' (milho transgênico) e 'A2' (milho orgânico) pelo valor 'A' (milho). Todas as duplicatas, no conjunto de dados, são removidas e os dados são reordenados de acordo com os marcadores de tempo. Aqui, a unidade de tempo utilizada é o ano, e o período estudado é de 2012 a 2017. Assim, para cada caso, as atividades devem ser associadas a cada ano do período [2012–2017]. Qualquer informação ausente será

substituída por 'XX'. Três conjuntos de dados são criados na fase de preparação de dados: SVP representa os vetores de estado do ano *n* usados para predição, enquanto o conjunto de dados SVE representa os vetores de estado do ano *n + 1* usados para avaliação; assim, o conjunto de dados CTM é utilizado para criar o primeiro grafo DFG, que permite obter uma matriz adjacente e, em seguida, uma matriz de transição. Fase II: representação visual DFG de um grafo direcionado ponderado, em que cada aresta representa a relação direta entre duas atividades e o peso corresponde ao número de ocorrências dessa relação no conjunto de dados. Portanto, o grafo DFG é obtido com dados do exemplo CTM. Fase III: Criação de um modelo de predição embasado nos princípios de Markov, o qual é baseado em uma matriz de transição, representando a probabilidade de uma transição entre cada estado. Esta matriz pode ser estimada usando as probabilidades condicionais entre cada atividade como uma probabilidade de transição. Fase IV (predição): em uma matriz de transição previamente criada para gerar a matriz adjacente associada, são usados o conjunto de dados SVP e um dicionário de codificação. Este pode ser criado associando um número a cada atividade presente no conjunto de dados, com o objetivo de obter um vetor probabilístico usando o cruzamento. Fase V (Avaliação do modelo): é um vetor de probabilidade que associa cada atividade possível à probabilidade de que ela ocorra no tempo *n + 1*. As atividades podem, portanto, ser classificadas de acordo com sua probabilidade de ocorrência. A obtenção de uma lista das *N* atividades mais prováveis (*Top-N Activities*) é útil, em particular, para propor uma recomendação. Fase VI (simplificação do grafo): o uso de grafos DFG permite destacar as relações mais frequentes entre as atividades. No entanto, a dificuldade se encontra na legibilidade do grafo DFG, que, quando usado em dados reais com alguma diversidade, pode rapidamente se tornar muito grande. Para contornar este problema, duas operações são propostas: a operação de agrupamento e a operação de filtragem. Após as seis fases, comparamos a abordagem com as redes neurais LSTM, GRU e RNN.

- Predição de eventos criminais com recursos dinâmicos (Rumi, Deng, & Salim, 2018b): selecionaram-se quatro modelos de predição, que são: Floresta Aleatória (RF), Rede Neural (NN), máquina de vetores de suporte (SVM) e modelo de regressão logística (LR). Além destes, também usou-se uma estrutura de aprendizagem baseada em conjunto para predição de eventos criminais: (I) Dividiu-se

o espaço de treinamento com base nos tipos de características. Dividiu-se, também, o espaço de treinamento em três subconjuntos: o histórico, o geográfico e o problema. (II) Construíram-se classificadores, usando diferentes algoritmos de aprendizagem para cada subconjunto de treinamento não sobreposto. Quando divididos os espaços com base em tipos de características, podem-se criar estruturas lineares e combinação não lineares de características. Em particular, escolheu-se a *Support Vector Machine* (SVM) como um modelo baseado em *kernel*; *Classification and Regression Tree* (CART) como modelo baseado em árvore; *Random Forest* (RF) como *ensemble;* e *Linear Discriminant Analysis* (LDA) como modelo linear. (III) Reuniram-se os classificadores para cada subconjunto de treinamento separadamente e combinaram-se as saídas dos classificadores de cada subconjunto. Entre diferentes tipos de técnica de combinação, usou-se a regra da soma para modelo *ensemble*. (IV) Agregaram-se os resultados de cada subconjunto de características para o resultado final e treinou-se um modelo SVM para agregar as predições feitas por cada subconjunto de treinamento.

- Prevenção do crime de furtos de ônibus em Pequim, China: a qualidade do ar afeta o crime? (Ding & Zhai, 2019): o coeficiente de correlação de *Pearson* foi calculado para determinar a correlação entre quatro fatores ambientais e dados diários de furtos de ônibus, usando o SPSS 19. Os três parâmetros (AQI, aPM2,5 e bPM2,5) foram significativamente correlacionados com o número de incidentes diários de furtos em ônibus. Portanto, o coeficiente de correlação de *Pearson* foi positivo, implicando que, quanto pior a qualidade do ar, maior o risco de furtos em ônibus. Este fenômeno pode ser explicado a partir de dois aspectos: (1) o impacto do declínio da visibilidade e (2) o impacto psicológico. Quando a qualidade do ar é fraca, o número de partículas finas no ar aumenta, reduzindo assim a visibilidade e, em ambientes de baixa visibilidade, a probabilidade de condução perigosa aumenta. Todos os condutores, incluindo os que conduzem ônibus, podem reduzir a velocidade para melhorar o tempo de resposta em caso de emergência. À medida que a velocidade média dos ônibus urbanos diminui, os passageiros permanecem neles por mais tempo. Essas condições proporcionam aos batedores de carteira mais tempo para cometer vários crimes em um só lugar, aumentando a eficiência da atividade criminosa e o risco de crimes. No entanto, os resultados do SPSS, neste estudo, mostram que os coeficientes de temperatura e

sazonalidade de *Pearson* excedem 0,05; portanto, como segunda condição na teoria da AR, esses fatores não parecem exercer um impacto considerável sobre os batedores de carteira de ônibus, ao contrário da qualidade do ar, uma vez que os autores utilizaram os métodos SVM e *Naive Bayes*. Embora SVM tenha maior precisão e *recall*, sua precisão e seu *Score* F1 não foram os melhores. Se em alguns locais, onde o número de furtos de ônibus for maior, o modelo pode ser melhorado em dois aspectos: primeiro, o nível de criminalidade pode ser dividido em três níveis ou cinco níveis de acordo com a situação real, e as medidas de prevenção e controle serão mais flexíveis; segundo, o impacto do pequeno valor, na precisão da predição, será correspondentemente reduzido.

- Melhorar a predição da criminalidade a curto prazo com fluxos de mobilidade humana e arquiteturas de aprendizagem profunda (J. Wu, Abrar, Awasthi, Frias-Martinez, & Frias-Martinez, 2022): as análises mostraram que arquiteturas não neurais visto que usam dados de mobilidade têm pior desempenho do que algumas arquiteturas neurais, incluindo o uso de métodos de aprendizagem profunda para modelos de predição do crime baseados em locais que incorporam dados de mobilidade, refletindo em dados, agregados em fluxos, por meio dos setores censitários. Os resultados mostraram que características de fluxo de mobilidade extraídas de dados de GPS — coletados por empresas de inteligência de localização — podem melhorar o desempenho da predição de crimes no curto prazo com modelos de arquiteturas neurais e não neurais (HALR), curiosamente GRU, Attn e *NbConv*. No entanto, a diferença é pequena, com arquiteturas neurais produzindo *Score* F1 de predição de crime de curto prazo de até 2% do que abordagens não neurais. Revelou-se também que os fluxos de mobilidade, utilizados em conjunto com dados históricos de criminalidade, fornecem sistematicamente os maiores aumentos no *Score* F1 quando comparados aos modelos que utilizam apenas dados históricos de crimes, e que essas melhorias são generalizadas como modelos neurais e não neurais em diversas cidades e tipos de crime. Usando apenas fluxos de mobilidade como preditores de criminalidade, em vez de dados históricos de criminalidade, também produziram melhorias sistemáticas no *Score* F1 entre cidades e tipos de crime, embora apenas para o modelo neural *NbConv*.

- Algoritmos de aprendizado de máquina para predição de crimes na Índia (Aziz, Sharma, & Hussain, 2022a): dados de crimes espaço-temporais podem ser usados por analistas de dados para realizar análises de séries temporais e para construir regressores que podem prever diferentes tipos de contagem de crimes a serem conhecidos pelo IPC em termos regionais, estaduais e anuais para o país. Os dados previstos desses regressores podem ser posteriormente utilizados pelos analistas para obter informações úteis no auxílio à polícia e a outras agências de aplicação da lei para controlar e prevenir o crime. Pode-se afirmar que, sem uma abordagem baseada em dados, é impossível atingir tais objetivos, uma vez que a taxa de criminalidade é afetada direta e indiretamente por numerosos fatores que também mudam de tempos em tempos. Para construir regressores (preditores de contagem de crimes), foram utilizados diferentes algoritmos: Regressão Linear Simples, Regressão Linear Múltipla, Regressão da Árvore de Decisão, Regressão Florestal Aleatória e Regressão do Vetor de Suporte. Para os dados escolhidos, a regressão florestal aleatória, que prevê o crime total conhecível pelo IPC, ajusta-se relativamente melhor com um valor de R^2 ajustado de 0,96 e um valor MAPE de 0,2 entre os modelos de regressão que preveem a contagem de crimes de roubo em toda a região. O modelo baseado em regressão "florestal aleatória" ajusta-se relativamente melhor com um valor de R^2 ajustado de 0,96 e um valor MAPE de 0,166. De acordo com estes modelos de regressão, a maior contagem de crimes em 2022 permaneceu com o distrito de Adilabad liderando com uma contagem de crimes prevista de 31.933 e o distrito de Anantapur liderando com uma contagem de crimes de roubo prevista de 12.000. Madhya Pradesh, por outro viés, é o estado com maior taxa média de criminalidade. Este artigo também mostra o padrão de criminalidade no estado de MP, cujos dados previstos pelos regressores podem ser usados com técnicas de visualização de dados, como mapa, ou outras técnicas geoespaciais, como mapa de calor, mapa coroplético, mapa de pontos, mapa de *cluster*, mapa de bolhas, mapa de cartograma, *binning hexagonal* etc. Assim, a abordagem proposta fornece uma estrutura para outros analistas de dados usarem-na para a predição e para a visualização de dados criminais indianos a partir dos algoritmos de regressão.
- Sistema de identificação de escritor para manuscrito offline pré-segmentado Caracteres Devanagari usando k-NN e SVM (Dargan, Kumar, Garg, & Thakur, 2019):

a motivação para o sistema e o desenvolvimento de uma estrutura de identificação do escritor originou-se da utilidade e necessidade contínua de registros forenses, análise, verificação de registros no sistema bancário, reconhecimento biométrico e assim por diante. Como a caligrafia se trata de um equilíbrio entre a diversidade de soluções e a convergência da arte, que é tão flexível e dinâmica, há uma variedade de desafios que enfatizam mais interesse e entusiasmo dos pesquisadores pelo desenvolvimento do sistema proposto baseado no texto manuscrito. Os vários desafios são complexos para distinguir os diferentes estilos de caligrafia de indivíduos, diversas formas e tamanho dos alfabetos, a qualidade do documento, estilos de caligrafía restritos e caligrafías irrestritas, entre outras. Nesse sentido, *Script Devanagari* e a coleta de dados são escritas da esquerda para a direita e têm forte preferência pelas formas arredondadas simétricas dos caracteres, em maiusculas e minúsculas distintas, sendo reconhecidas pela linha horizontal ao longo da parte superior da letra completa. Os autores desenvolvem um sistema robusto com fase eficiente de extração de características que realiza a extração de características discriminativas seguida da fase de classificação. A estrutura proposta consiste em várias etapas: (I) Digitalização e pré-processamento: a fase de digitalização trata da conversão de caracteres manuscritos em imagem digital. Na fase de pré-processamento, diversas operações serão aplicadas à imagem de um personagem. Primeiramente, a imagem do caractere é normalizada usando o método de interpolação do vizinho mais próximo e, em seguida, as imagens são convertidas em uma imagem *bitmap*; assim, o *bitmap* é transformado em uma imagem "desbastada", usando um algoritmo de desbaste paralelo. (II) Extração de recursos: extrai traços, características e propriedades relevantes sobre o personagem e finalmente classifica o escritor. (III) Características de zoneamento (F1): neste método, foi feita a decomposição da imagem reduzida de um caractere em n (= 100) número de zonas estimadas equivalentes. Assim, o número de pixels da área frontal em cada zona é determinado. (IV) Características diagonais (F2): trata da divisão da imagem original reduzida de um caractere em n (= 100) número de zonas estimadas equivalentes. Essas características são extraídas dos pixels de cada zona, movendo-se ao longo de suas diagonais. (V) Características de transição (F3): depende das estimativas e das áreas de transições dos pixels de fundo para primeiro plano nas direções vertical e horizontal. Para isso, a imagem é examinada da esquerda para a

direita e de cima para baixo. (VI) Características baseadas em extensão de pico (F4): implementam características baseadas na extensão de pico apresentados no conjunto de dados. Nessa técnica, as características são extraídas considerando a soma da extensão do pico que ajusta sucessivos pixels pretos ao longo de cada zona. As características — baseadas na extensão de pico — podem ser extraídas horizontal e verticalmente. (VII) Classificação: identificar o redator diante das características extraídas na fase anterior. Para a classificação, os autores consideraram duas técnicas de classificação, nomeadamente k-NN e SVM. Uma taxa de precisão de 91,53% foi obtida usando a estrutura proposta com a combinação de características baseadas em zoneamento, diagonal, transições e extensão de pico e classificador SVM linear.

Antes de mergulhar nas metodologias e técnicas aplicadas à previsão de crimes e roubos em ambientes urbanos, é essencial reconhecer a complexidade e as nuances subjacentes à atividade criminal. A criminalidade é um fenômeno multifacetado, influenciado por uma confluência de fatores sociais, econômicos, psicológicos e ambientais. A capacidade de prever com precisão a ocorrência de crimes e roubos não apenas facilita uma melhor alocação de recursos de segurança pública, mas também ajuda a moldar políticas de prevenção eficazes. Este texto explora as abordagens mais avançadas e interdisciplinares na ciência da predição de crimes, destacando como a integração de diversas fontes de dados e metodologias analíticas pode aprimorar significativamente as estratégias de combate ao crime em centros urbanos. Com este entendimento, avança-se para examinar especificamente os métodos empregados na predição de crimes e de roubos, seus pontos fortes, limitações e o impacto potencial na segurança e no bem-estar das comunidades urbanas.

Para predizer crimes/roubos em ambiente urbano, é necessário coletar e analisar dados históricos sobre as criminalidades de cada indivíduo, como local, horário, data, localização e outros aspectos necessários. Com base nesses dados, é possível construir um modelo de predição que leve em consideração as variáveis relevantes para a ocorrência de crimes/roubos. Diversos métodos de predição são utilizados para essa finalidade, incluindo:

Análise Estatística: trata-se de uma abordagem tradicional que envolve a identificação de relações estatísticas entre variáveis e a ocorrência de roubos. A regressão logística, por exemplo, é usada para modelar a probabilidade de ocorrência de roubos com base em

variáveis explicativas, como horário do dia, localização e condições ambientais. No entanto, essa abordagem pode ser limitada quando se relaciona com dados complexos e não lineares.

Aprendizado de Máquina: algoritmos de aprendizado de máquina são mais populares na predição de roubos devido à sua capacidade de lidar com conjuntos de dados complexos. Árvores de decisão, redes neurais, máquinas de vetores de suporte e outros algoritmos podem identificar padrões não lineares imperceptíveis em abordagens puramente estatísticas. No entanto, esses modelos exigem grande quantidade de dados de treinamento e ajustes apropriados.

Análise Geoespacial: é fundamental na predição de roubos, uma vez que a localização desempenha um papel crítico na ocorrência de crimes. Sistemas de Informação Geográfica (SIG) são empregados para analisar a distribuição espacial de roubos, considerando fatores geográficos, como proximidade às áreas de alto risco, presença de iluminação pública e densidade populacional. Essa abordagem leva em consideração o contexto espacial dos crimes.

Modelos de Séries Temporais: em algumas situações, esses modelos são utilizados para prever roubos com base em padrões temporais. Isso é particularmente útil quando se deseja prever a evolução dos roubos ao longo do tempo e identificar sazonalidades ou tendências, como picos durante feriados ou estações do ano.

Aprendizagem de Máquina: técnicas de aprendizagem de máquina, como aprendizado profundo, têm sido aplicadas com sucesso na predição de roubos. Esses modelos são capazes de aprender automaticamente com os dados e aperfeiçoar suas predições com o tempo, permitindo a identificação de padrões complexos e não lineares.

Ao ser contemplado o horizonte de estratégias empregadas para antecipar e mitigar crimes e roubos em ambientes urbanos, torna-se evidente que a integração dessas metodologias oferece a mais promissora direção para avanços futuros. A combinação de análise estatística, aprendizado de máquina, análise geoespacial e modelos de séries temporais não só abrange uma gama de variáveis e condições, mas também capitaliza sobre as forças complementares de cada técnica. O emprego inteligente de grandes volumes de dados e análises sofisticadas representa a vanguarda da criminologia preditiva, prometendo uma nova era de segurança urbana otimizada e responsiva. À medida que se refinam essas ferramentas e tecnologias, a expectativa é que as agências de segurança pública se tornem

mais capacitadas para prevenir crimes antes que eles ocorram, garantindo a segurança e a tranquilidade das comunidades a que servem.

Para construir um modelo de predição eficaz, é necessário existir um conjunto de dados históricos de alta qualidade que serão utilizados para treinar o modelo. A preparação dos dados desempenha um papel fundamental, pois envolve a limpeza e a formatação dos dados, tratando questões como dados duplicados, genéricos, vazios e outros. A partir disso, o modelo é, então, testado em um conjunto de dados de validação para avaliar sua precisão na predição de valores futuros.

Em resumo, a predição de roubos, em ambiente urbano, é uma tarefa complexa que envolve a aplicação de uma variedade de métodos, desde análises estatísticas tradicionais até técnicas avançadas de aprendizado de máquina. A escolha do método depende da natureza dos dados e dos objetivos da predição, com a análise geoespacial, desempenhando um papel central na compreensão da distribuição espacial dos roubos na cidade.

2.5. Tabela de Síntese de Artigos

A luta contra o crime é uma tarefa complexa e desafiadora que requer a colaboração de múltiplos fatores, incluindo as forças de segurança, o poder judiciário e a sociedade em geral. Nesse aspecto, estudar a predição de tipos de crime com mineração de processo e Cadeias de Markov é essencial para melhorar a segurança pública, otimizar recursos e tomar decisões informadas. Para tanto, a proposta dessa dissertação é utilizar grafos de frequência e transição de atividades gerados por algoritmos de mineração de processos, cuja estrutura será o ponto de entrada para a matriz de transição de Cadeias de Markov.

A tabela a seguir apresenta uma comparação entre vários artigos relevantes ao tema desta pesquisa, os quais foram selecionados com o objetivo de destacar as principais características, objetivos e resultados de cada estudo, fornecendo uma visão abrangente das contribuições à literatura existente. A análise comparativa desses artigos ajuda a identificar tendências, lacunas e áreas de convergência ou divergência na pesquisa existente e, portanto, a *Tabela 9* serve como uma ferramenta útil para resumir as informações-chave e auxiliar na discussão dos resultados e conclusões deste estudo, resumindo as principais características da predição de tipo de crimes dos artigos selecionados.

Tabela 9 - Tabela dos Artigos Pesquisados.

Fonte	Autor	Área	Local	Tipo	Método	Base de Dados
IEEE	Pavel Rahman et al., 2021	Relatórios Criminais.	Bangladesh.	Predição de Criminalidade.	Random Forest (RF), Decision Tree (DT), Support Vector Regression (SVR), Multi-Layer Perceptron (MLP), Lasso Bayesian (LB), Ridge, e Linear Regression (LR).	https://github.com/pavstar619/NahidSir_prototype_bdpolice_V2;
IEEE	Orong, Sison, & Hernandez, 2018	Província de Misamis Ocidental.	Filipinas.	Análise de Dados.	Autoregressive Integrated Moving Average (ARIMA).	Não disponibilizada pelos autores.
ACM	X. Zhao & Tang, 2018	Crimes.	Urbano.	Modelo de Previsão de Taxa de Criminalidade.	Kool Desktop Environment (KDE).	Não disponibilizada pelos autores.
ACM	Pradhan, Potika, Eirinaki, & Potikas, 2019b	Cidades Inteligentes.	São Francisco - CA.	Análise exploratória de Dados e Padrões de Crimes.	K-Nearest Neighbor (KNN) , Multi-class Logistic Regression (MCLR) , Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), Gradient Boosting Machines (GBM), Neural Networks (NN), Maximum Entropy Classifier (MAXEN T) e Scaled Linear Discriminant Analysis (SLDA).	https://data.gov/open-gov/; https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-Historical-2003/tmnf-yvry

ACM	Morshed et al., 2019	Previsão e Visualização do Crime Com Fonte de Dados Heterogênea.	Região/Bairro.	Mapear Padrões Criminais.	Long Short Term Memory (LSTM).	VisCrimePredict: um Sistema de Análises Visuais e Preditivas Para Mapear Crimes.
SCIENCE DIRECT	Thomas & Sobhana, 2022	Policiamento Preditivo.	Áreas Geográficas.	Ocorrência de Atividade Criminosa.	Linear Regression (LR), Random Forest (RF) , Autoregression, Decision Tree (DT), Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), Random Forest (RF), eXtreme Gradient Boosting (XGBoost), Long-Short Term Memory (LSTM) e Autoregressive Integrated Moving Average (ARIMA).	Não disponibilizada pelos autores.
SCIENCE DIRECT	Y. Zhang, Siriaraya, Kawai, & Jatowt, 2020b	Método de Recomendação.	Distribuição Geográfica.	Prever a Hora e Localização de Crimes Futuros.	Autoregressive Integrated Moving Average (ARIMA), Vector Autoregression (VAR), Collaborative Filtering (CF) e Kool Desktop Environment (KDE).	Não disponibilizada pelos autores.
SCIENCE DIRECT	Rummens, Hardyns, & Pauwels, 2017	Investigar o Potencial de Aplicação.	Holanda.	Análise Preditiva.	Crime Anticipation System (CAS), Neural Networks (RN)	Não disponibilizada pelos autores.
SCIENCE DIRECT	Yoo & Wheeler, 2019	Modelagem de Terreno	Los Angeles	Crimes Relacionados a Moradores de	Risk Terrain Modeling (RTM).	Não disponibilizada pelos autores.

				Rua		
SCIENCE DIRECT	Dupuis, Dadouchi, & Agard, 2022	Agricultura.	Canadá	Previsão de Rotação de Culturas.	Process Mining (PM), Markov Chains (MC), Recurrent Neural Networks (LSTM, RNN, GRU), DFG Graph, Prediction, Grouping Operation and Filtering Operation.	Não disponibilizada pelos autores.
SPRINGER	Rumi, Deng, & Salim, 2018b	Previsão de Eventos Criminais Com Recursos Dinâmicos.	Informações de *Tweets*	Dados de Check-in do Foursquare.	Location-Based Social Networks (LBSN), Random Forest (RF), Neural Network (NN), Support Vector Machine (SVM), Logistic Regression (LR).	https://www.police.qld.gov.au/; http://www.predpol.com/how-predictive-policing-works/; https://data.qld.gov.au; http://www.abs.gov.au/; https://data.cityofnewyork.us/Public-Safety/; https://www.census.gov/about/regions/new-york.html
SPRINGER	Ding & Zhai, 2019	Furtos de Ônibus.	Pequim.	Batedores de Carteira e Qualidade do Ar.	Support Vector Machine (SVM) e Naive Bayes (NB).	Não disponibilizada pelos autores.
SPRINGER	J. Wu, Abrar, Awasthi, Frias-Martinez,	Previsão de Crimes de Curto Prazo	Mobilidade Urbana	Espaço-temporais de Crimes	Gated Recurrent Unit (GRU).	https://data.census.gov/table?q=demographic

	& Frias-Martinez, 2022					
SPRINGER	Aziz, Sharma, & Hussain, 2022a	Previsão de Crimes Sobre o Código Penal Indiano.	Índia	Aplicação de Leis.	Random Forest Regression (RFR), Decision Tree Regression (DTR), Multiple Linear Regression (MLR), Simple Linear Regression (SLR) e Support. Vector Regression (SVR)	Todos os dados utilizados são *benchmark* e estão disponíveis gratuitamente em repositórios, assim como todos os códigos utilizados estão disponíveis gratuitamente na rede, mas os autores não disponibilizam um determinado link.
SPRINGER	Dargan, Kumar, Garg, & Thakur, 2019	Identificação Biométrica	Caligrafia	Modalidades Únicas e Múltipla	K-Nearest Neighbor (KNN), Support Vector Machine (SVM).	Não disponibilizada pelos autores.

Fonte: a autora.

O artigo de (Pavel Rahman et al., 2021) aborda a predição da criminalidade e utiliza as técnicas *Support Vector Regression* (SVR), *Random Forest* (RF), *Decision Tree* (DT), *Support Vector Regression* (SVR), *Multi-Layer Perceptron* (MLP), *Lasso Bayesian* (LB), *Ridge* e *Linear Regression* (LR). A aplicação do SVR oferece *insights* valiosos sobre a eficácia dessa técnica na predição de eventos criminais. (Orong, Sison, & Hernandez, 2018) abordam a análise de dados na província de Misamis Ocidental, nas Filipinas, utilizando a *Autoregressive Integrated Moving Average* (ARIMA). (X. Zhao & Tang, 2018) tratam de modelo de predição de taxa de criminalidade urbana, aplicando *Kool Desktop Environment* (KDE). (Pradhan, Potika, Eirinaki, & Potikas, 2019b) utilizam a análise exploratória de dados e padrões de crimes nas cidades inteligentes, aplicando *K-Nearest Neighbor* (KNN), *Multi-class Logistic Regression* (MCLR), *Decision Tree* (DT), *Random Forest* (RF), *Naïve Bayes* (NB), *Support Vector Machine* (SVM), *Gradient Boosting Machines* (GBM), *Neural Networks* (NN), *Maximum Entropy Classifier* (MAXEN T) e *Scaled Linear Discriminant Analysis* (SLDA).

(Morshed et al., 2019) utilizam a predição e visualização do crime com fonte de dados heterogênea, mapeando padrões criminais e utilizando *Long Short Term Memory* (LSTM). (Thomas & Sobhana, 2022) e policiamento preditivo em áreas geográficas de ocorrência de atividade criminosa. Para tanto, utilizam-se os métodos *Linear Regression* (LR), *Random Forest* (RF), *Autoregression, Decision Tree* (DT), *Support Vector Machine* (SVM), *Naïve Bayes* (NB), *K-Nearest Neighbors* (KNN), *Multilayer Perceptron* (MLP), *Random Forest* (RF), *eXtreme Gradient Boosting* (XGBoost), *Long-Short Term Memory* (LSTM) e *Autoregressive Integrated Moving Average* (ARIMA). Em relação a (Y. Zhang, Siriaraya, Kawai, & Jatowt, 2020b), o artigo se concentra na predição da hora e na localização de crimes futuros, um aspecto de grande relevância para este estudo, aplicando as variações de rede neural: *Autoregressive Integrated Moving Average* (ARIMA), *Vector Autoregression* (VAR), *Collaborative Filtering* (CF) e *Kool Desktop Environment* (KDE).

(Rummens, Hardyns, & Pauwels, 2017) investigam o potencial de aplicação de análise preditiva na Holanda com os métodos *Crime Anticipation System* (CAS) e *Neural Networks* (RN). (Yoo & Wheeler, 2019) se pautam na modelagem de terreno de crimes relacionados a moradores de rua em Los Angeles, utilizando *Risk Terrain Modeling* (RTM). (Rumi, Deng, & Salim, 2018b) se embasam na predição de eventos criminais com recursos dinâmicos com dados de check-in do *Foursquare*, aplicando os métodos *Location-Based Social Networks* (LBSN), *Random Forest* (RF), *Neural Network* (NN), *Support Vector Machine* (SVM) e

Logistic Regression (LR). (Ding & Zhai, 2019) analisam furtos em ônibus de batedores de carteira e qualidade do ar a partir dos métodos *Support Vector Machine* (SVM) e *Naive Bayes* (NB). (J. Wu, Abrar, Awasthi, Frias-Martinez, & Frias-Martinez, 2022) partem da predição de crimes de curto prazo na mobilidade urbana espaço-temporais de crimes, aplicando *Gated Recurrent Unit* (GRU).

Com base em pesquisas anteriores sobre predição de crimes, o artigo de (Dupuis, Dadouchi, & Agard, 2022) inspirou nossa pesquisa na prevenção de crimes, pois descreve um método que utiliza mineração de processos e cadeias de Markov para prever rotações de culturas com sucesso, demonstrando o potencial da aplicação dessas técnicas na prevenção de crimes para possibilitar a predição de futuros crimes para medidas proativas. Assim, utilizar-se-ão, na presente dissertação, os passos desse método, quais sejam: criação dos três conjuntos de dados na fase de preparação de dados, SVP, SVE e CTM; criação de um grafo DFG, gerando uma matriz adjacente com os valores das arestas desse grafo; e criação de um modelo de predição baseado nos princípios de Markov, em que a matriz de transição é criada e usada com o conjunto de dados SVP, produzindo-se um dicionário de codificação e gerando o SVE-Predição. Na avaliação do modelo, utilizar-se-á o desempenho das métricas *Recall* e *Precision*. Por fim, a simplificação do grafo DFG permite destacar as relações mais frequentes entre as atividades; no entanto, a legibilidade do grafo DFG pode se tornar difícil, visto que, para contornar este problema, duas operações são utilizadas: a operação de agrupamento e a operação de filtragem, comparando-as com modelos de rede neural LSTM, GRU, e RNN.

Comparando o desempenho com modelos de rede neurais, o uso de cadeias de Markov aborda o problema de predição de sequência, que é considerado fundamental no campo do aprendizado de máquina. Muitas pesquisas, como as de (Morshed et al., 2019), (Dupuis, Dadouchi, & Agard, 2022), (J. Wu, Abrar, Awasthi, Frias-Martinez, & Frias-Martinez, 2022), especialmente em modelos de redes neurais recorrentes (RNN, LSTM, GRU), fornecem soluções que usam explicitamente uma sequência para prever o valor futuro. Para comparar o desempenho do modelo de cadeias de Markov com o que pode ser obtido por meio de redes neurais, os dados do estudo de caso são reutilizados em três modelos de redes neurais recorrentes (SimpleRNN, LSTM e GRU), que devem passar novamente por etapas de preparação. Quatro conjuntos de dados (*treino*: conjunto de treinamento, *treino*: rótulos do conjunto de treinamento, *teste X*: conjunto de teste e *teste Y*: rótulos do conjunto de teste) são gerados.

No contexto da predição de crimes, métricas como *Recall* e *Precision* desempenham um papel fundamental na avaliação da eficácia de nossos modelos. *Recall* mede a capacidade do modelo em identificar corretamente casos positivos, ou seja, crimes que realmente ocorreram; *Precision*, por outra acepção, avalia a precisão das predições positivas, indicando as corretas em relação ao total de predições.

Planejamos utilizar *Recall* e *Precision* como métricas de desempenho em nosso projeto, uma vez que mostram o quão bem o modelo é capaz de identificar e predizer eventos criminais. Nesse aspecto, um alto *Recall* indica que estamos capturando a maioria dos crimes reais, enquanto um alto *Precision* significa que nossas predições positivas são altamente confiáveis. Desse modo, tais métricas são essenciais para medir a eficácia de nossa abordagem na prevenção de crimes.

2.6. Consideração Final

Os resultados das iniciativas do MSL de predição de tipos de crimes, usando técnicas de mineração de processos e Cadeias de Markov (SQ1), indicam que a principal contribuição dessa abordagem é a possibilidade de predizer os tipos de crimes que podem ocorrer em uma determinada região com base em padrões de comportamento criminoso identificados em dados históricos. Ademais, a mineração de processos é uma técnica de análise de dados que permite identificar padrões em eventos sequenciais (SQ2), que, na predição de crimes, pode ser usada para analisar dados históricos destes crimes e identificar padrões de comportamento criminoso, como a escolha de determinadas áreas geográficas, horários ou tipos de vítimas.

As cadeias de Markov, por sua vez, são modelos probabilísticos que permitem predizer a probabilidade de um evento futuro, isto é, de um determinado tipo de crime ocorrer com base em eventos anteriores, como o número de crimes semelhantes ocorridos na região. Complementarmente, redes neurais recorrentes (RNN), incluindo suas variantes LSTM e GRU, foram utilizadas para capturar padrões temporais mais complexos e dependências de longo prazo nos dados. Essas redes oferecem uma abordagem robusta para prever tipos de crimes, especialmente em contextos onde as transições entre eventos possuem relações mais sofisticadas do que as consideradas pelas cadeias de Markov.

Ao analisar esses principais resultados identificados no MSL, evidenciou-se que diferentes abordagens e elementos foram utilizados na predição de tipos de crimes/roubos em

ambientes urbanos. Essa diversidade de métodos e técnicas oferece uma base significativa para o desenvolvimento de estratégias eficazes na prevenção de crimes e na promoção da segurança urbana.

3. CONCEITOS PRELIMINARES DE MINERAÇÃO DE PROCESSOS E CADEIAS DE MARKOV

Esta seção apresenta o referencial teórico e exemplificações com as seguintes subseções: Mineração de processos, *Logs* de Eventos, Modelo de Processos, DFG, Cadeias de Markov e Considerações Finais.

3.1. Mineração de processos

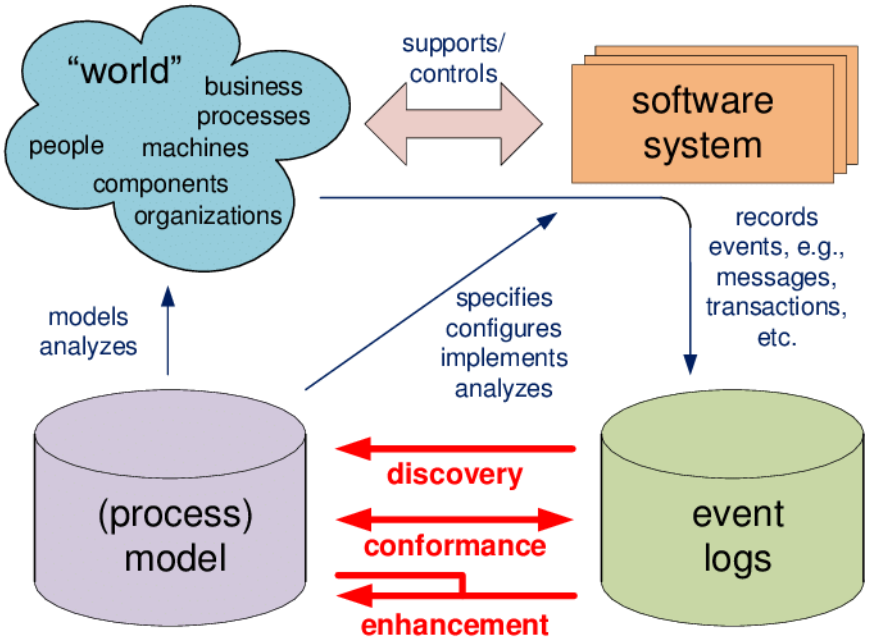
A mineração de processos é uma abordagem interdisciplinar que surgiu como uma ferramenta essencial para mapear informações e analisar dados de um banco de dados, com fortes conexões às áreas de aprendizado de máquina e mineração de dados, incluindo negócios, saúde e ciência. Esta técnica, introduzida pelo pesquisador (W. van der Aalst, 2011) e (W. M. P. van der Aalst, 2016), tornou-se uma ferramenta valiosa para descobrir, validar e melhorar fluxos de trabalho em diversas áreas, transformando os registros de eventos em *insights* valiosos e permitindo a compreensão e melhoria de processos organizacionais e a tomada de decisões baseadas em dados.

Sob outra premissa, a mineração de processos é definida como o processo de descobrir informações, padrões e conhecimento a partir de *logs* de eventos gerados por sistemas de informação. Seus objetivos primordiais incluem a identificação e a sequência das atividades, a modelagem de fluxos de trabalho, a análise de desvios em relação aos processos planejados e a melhoria contínua das operações. Para tanto, tal processo faz uso de diversas técnicas e métodos, que incluem a descoberta de processos, a partir da identificação automática dos processos subjacentes a partir de *logs* de eventos, a análise de conformidade, para garantir que os processos estejam em conformidade com as regras e regulamentos, e a melhoria de processos, concentrando-se em otimizar fluxos de trabalho para aumentar a eficiência.

Em suma, a mineração de processos é uma disciplina que desempenha um papel fundamental na extração de informações valiosas a partir de *logs* de eventos, permitindo a melhoria contínua de processos, a identificação de problemas e a tomada de decisões informadas. Sua aplicação abrange várias áreas e desempenha um papel crítico na análise de processos organizacionais.

Um modelo de processos normalmente é gerado automaticamente a partir de dados encontrados por meio de *logs* de eventos, acompanhando três tipos de mineração de processos: descoberta, conformidade e aprimoramento, conforme a *Figura 4*.

Figura 4 - Posicionamento dos três principais tipos de mineração de processos: descoberta, conformidade e aprimoramento.



Fonte: (W. van der Aalst, 2011, p. 9) e (W. M. P. van der Aalst, 2016, p. 32).

A descoberta de processos concentra-se na extração automática e na representação do fluxo de trabalho de um processo a partir de *logs* de eventos; em outras palavras, visa identificar os processos subjacentes a partir dos dados reais gerados durante a execução desses processos. O objetivo é criar modelos de processos que representem, com precisão, o comportamento real dos processos organizacionais, podendo assumir várias formas, como redes de petri, grafos de fluxo de trabalho ou diagramas de atividades. Pois, eles são úteis para visualizar e compreender de que maneira as atividades estão interconectadas em um processo, bem como identificar padrões de comportamento, cujos resultados podem ser usados para análise, otimização e automação de processos.

A análise de conformidade compara o modelo de processo com *logs* de eventos ou um processo real com modelo referência e/ou modelo distinto do processo. A técnica determina se o processo real corresponde ao processo de referência, e objetiva verificar semelhanças e

discrepâncias entre o comportamento modelado e o comportamento observado. Para a verificação, utilizam-se dois tipos de discrepância: a primeira é o *log* desajustado, em que o comportamento observado do *log* não é permitido pelo modelo; já a segunda é o comportamento adicional do modelo: o comportamento é permitido no modelo, porém nunca observado pelo *log*. Além disso, a mineração de conformidade também permite identificar gargalos e ineficiências em processos, uma vez que qualquer atividade, que não esteja em conformidade, pode afetar o fluxo geral do processo.

O aprimoramento de processos tem como objetivo otimizar e melhorar os processos existentes com base em *insights* extraídos dos *logs* de eventos em vez de apenas verificar a conformidade, cujo foco é tornar os processos mais eficientes, reduzindo custos, tempo e recursos. Isso envolve a análise do desempenho do processo e a identificação de áreas que podem ser aprimoradas, sendo que os modelos de processos simulam diferentes cenários e preveem o impacto de mudanças propostas. O aprimoramento de processos é, particularmente, valioso em ambientes de negócios, pois pode levar a melhorias significativas na produtividade e na qualidade dos serviços, ajudando as organizações a entender como seus processos podem ser aprimorados, seja por meio da automação, da otimização de recursos ou da eliminação de atividades redundantes.

Em analogia a (W. van der Aalst, 2011) e (W. M. P. van der Aalst, 2016), os tipos de processos citados acima apresentam apenas o fluxo de controle; portanto, esses tipos são estendidos a perspectivas adicionais. Diante disso, o primeiro e o segundo tipos (descoberta e conformidade) não se enquadram ao controle de fluxo.

Adicionadas ao modelo de fluxo de controle original, usando atributos dos *logs* de evento, (W. van der Aalst, 2011) e (W. M. P. van der Aalst, 2016) salientam que as diferentes perspectivas são parcialmente sobrepostas e não são exaustivas. No entanto, fornecem uma boa caracterização dos aspectos que a mineração de processos pretende analisar. Por fim, cada tipo de mineração de processos desempenha um papel importante na gestão e no aprimoramento dos processos organizacionais, adaptando-se a diferentes necessidades e cenários.

3.2. Log de Eventos

Logs de eventos são registros detalhados e, cronologicamente, sequenciais de todas as atividades e ações realizadas em um processo, sistema ou sistema de informação. Cada

entrada em um *log* de eventos, é registrada, com informações, como a atividade realizada, o momento em que ocorreu e, muitas vezes, o identificador do caso ao qual a atividade está associada. Esses *logs* são coletados automaticamente a partir de sistemas de informação, aplicativos, dispositivos ou processos em execução, em que cada evento é registrado à medida que ocorre, criando um rastro de auditoria das operações realizadas. Os *logs* de eventos, geralmente, incluem três elementos-chave: CASE ID (um identificador exclusivo associado a cada instância do processo, que permite rastrear todas as atividades relacionadas a um caso específico), ATIVIDADE (que representa a ação ou atividade que foi executada) e TIMESTAMP (que indica o momento em que a atividade ocorreu, incluindo data e hora).

Os *logs* de eventos são a matéria-prima essencial na mineração de processos, já que capturam a realidade das operações de um processo, oferecendo *insights* valiosos sobre como o trabalho é realizado na prática. Ao analisá-los, é possível identificar padrões, sequências de atividades, desvios, gargalos e oportunidades de aprimoramento nos processos. Utilizando a mineração de processos, eles se tornam a base para a descoberta, a verificação de conformidade e a melhoria de processos, pois fornecem dados reais que podem ser usados para construir modelos de processos, verificar se as atividades estão em conformidade com as regras e regulamentos e identificar áreas para otimização.

Em resumo, os *logs* de eventos desempenham um papel central na mineração de processos, haja vista que fornecem informações cruciais para a análise e para o aprimoramento de processos organizacionais, tornando-se a base sobre a qual as técnicas de mineração de processos são aplicadas para extrair conhecimento e *insights* úteis. Tais *logs* são criados a partir de dados que estão disponíveis para a extração dessas informações do sistema, como citado na seção acima, sendo fundamentados em três necessidades: a) *case identifier* (ID): de modo que é indispensável distinguir diferentes execuções do processo, ele depende do domínio do processo, como o número do processo de um crime; b) a atividade deve conter várias etapas/mudanças do processo, como categoria e tipo de crime que o indivíduo realizou; e c) *timestamp*: é usado para colocar os eventos/datas na ordem correta, como a data da primeira ocorrência, última data da ocorrência e o relato da ocorrência, ou seja, o boletim de ocorrência (BO), como é constituído no Brasil, conforme a *Figura 5*.

Figura 5 - *Log* de evento de crime, extraído da base de dados.

CASE ID: INCIDENT_ID	**ATIVIDADE:** OFFENSE_TYPE_ID	**TIMESTAMP:** REPORTED_DATE
2018869789	theft-other	12/27/2018 4:51:00 PM
2015664356	traffic-accident	11/13/2015 8:38:00 AM
20176005213	theft-bicycle	06/12/2017 08:44
20196012240	theft-from-bldg	12/09/2019 13:35
2018861883	violation-of-restraining-order	12/22/2018 10:00:00 PM
2018264446	threats-to-injure	4/20/2018 1:33:00 PM
2016461725	traf-other	7/21/2016 7:09:00 PM
2017409119	traf-other	6/22/2017 5:20:00 PM
2018473421	public-order-crimes-other	7/13/2018 10:11:00 AM
2016829592	sex-aslt-rape	12/31/2016 4:59:00 AM
2017455505	traffic-accident	07/10/2017 18:45
2015587324	theft-shoplift	10/08/2015 18:45
20186012344	theft-parts-from-vehicle	12/20/2018 4:22:00 PM
2019738697	theft-other	11/20/2019 12:58:00 PM
2015171560	burglary-residence-by-force	3/29/2015 2:45:00 PM
201985551	burglary-business-by-force	02/07/2019 10:13
201694006	theft-shoplift	2/13/2016 7:59:00 PM
2016102596	traffic-accident	2/17/2016 6:00:00 PM
2015317996	theft-items-from-vehicle	06/08/2015 16:15
2017443179	burglary-residence-no-force	07/05/2017 21:54
2020131877	criminal-trespassing	2/29/2020 9:16:00 PM

Fonte: A autora.

3.3. Modelo de Processos

Um modelo de processos refere-se a uma representação abstrata e visual de um processo organizacional ou de negócios. Descreve a sequência de atividades, tarefas, decisões e eventos que compõem o processo, bem como de que forma esses elementos estão interconectados. Além disso, é representado por grafos com o objetivo de determinar quais atividades podem ser realizadas significativamente e em que ordem podem ser escolhidas, ou seja, busca capturar a essência do funcionamento de um processo de uma maneira compreensível e eficaz. Assim, o processo pode ser observado por atividades tendo ponto de início e fim, desde que possuam entrada e saída evidentemente reconhecidas.

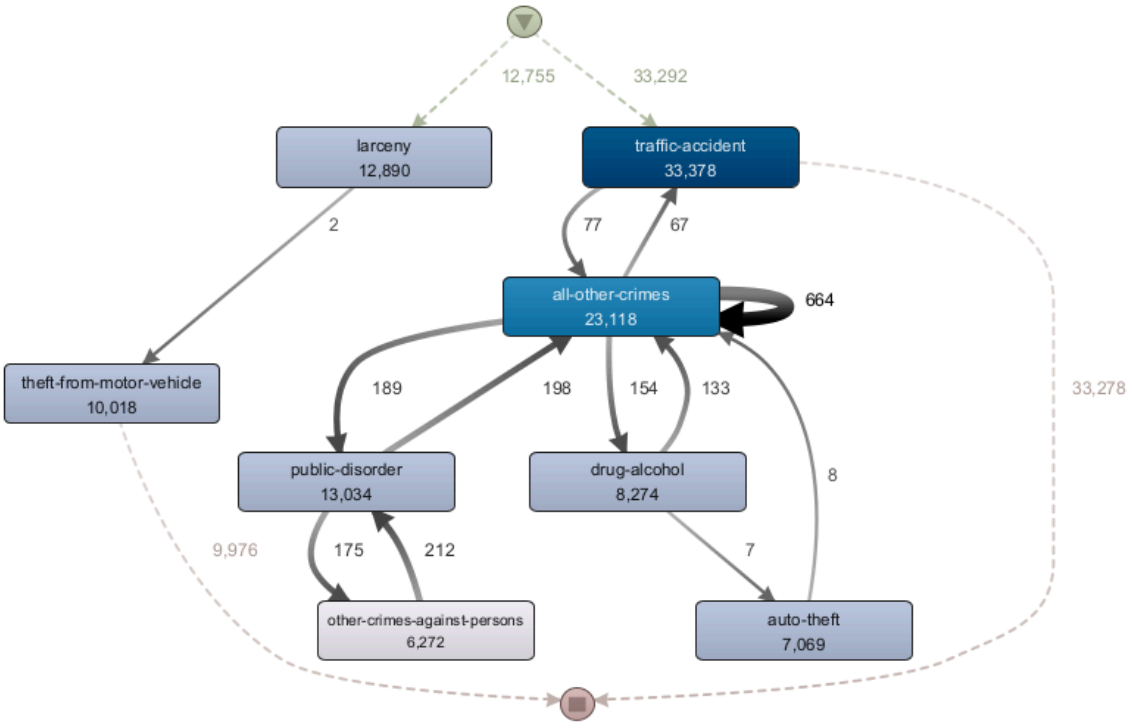
A criação de modelos de processos pode ser feita manualmente, com base no conhecimento humano e nas informações disponíveis sobre o processo, e também podem ser gerados automaticamente a partir de *logs* de eventos por meio de técnicas de descoberta de processos, como algoritmos de mineração de processos.

No contexto da mineração de processos, *Directly Follows Graph* (DFG) é um grafo de acompanhamento direto, em que cada nó representa uma atividade, e os arcos descrevem a relação entre várias atividades. Geralmente, o grafo tem uma origem e um destino, ou seja, atividades iniciais e finais; assim, um arco de sequências diretas entre qualquer das duas

atividades retrata que a atividade de origem é seguida pela atividade de coletor no *log* de evento. Uma das principais vantagens do Grafo DFG é a sua simplicidade e clareza. Em comparação com representações mais complexas, como redes de Petri ou BPMN, oferece uma visão direta das relações sequenciais entre atividades, facilitando a interpretação. Ao analisá-lo, podem-se identificar padrões de sequência de atividades comuns no processo, incluindo atividades que, frequentemente, ocorrem em uma ordem específica, bem como aquelas que são seguidas por um conjunto específico de atividades.

Para trabalhar com o Grafo DFG, pode-se usar *software* de mineração de processos especializado que permite criar, visualizar e analisar o grafo com base nos dados do *log* de eventos. Nesta dissertação, usou-se inicialmente o *software DISCO* (Process Mining and Automated Process Discovery Software for Professionals, n.d.) para a criação do Grafo DFG, conforme a *Figura 5*.

Figura 6 - Criação do grafo DFG a partir de *logs* de eventos de uma base de dados.



Fonte: A autora.

Na *Figura 6*, para gerar o Grafo DFG em questão, utilizou-se em CASE ID a coluna '*Incident_Id'*. Na coluna ATIVIDADE, utilizou-se a coluna *Offense_Category_Id'* e, na

coluna TIMESTAMP, a '*Reported_Id'*. Observa-se que a atividade dominante é a *"All-Other-Crimes"* (23,118), e o conjunto de dados contém 67 transições da atividade "Traffic-Accident" para a atividade "All-Other-Crimes" e 77 transições da atividade "All-Other-Crimes" para a atividade "Traffic-Accident", retornando para si mesmo. Portanto, "All-Other-Crimes" conta com 664 atividades.

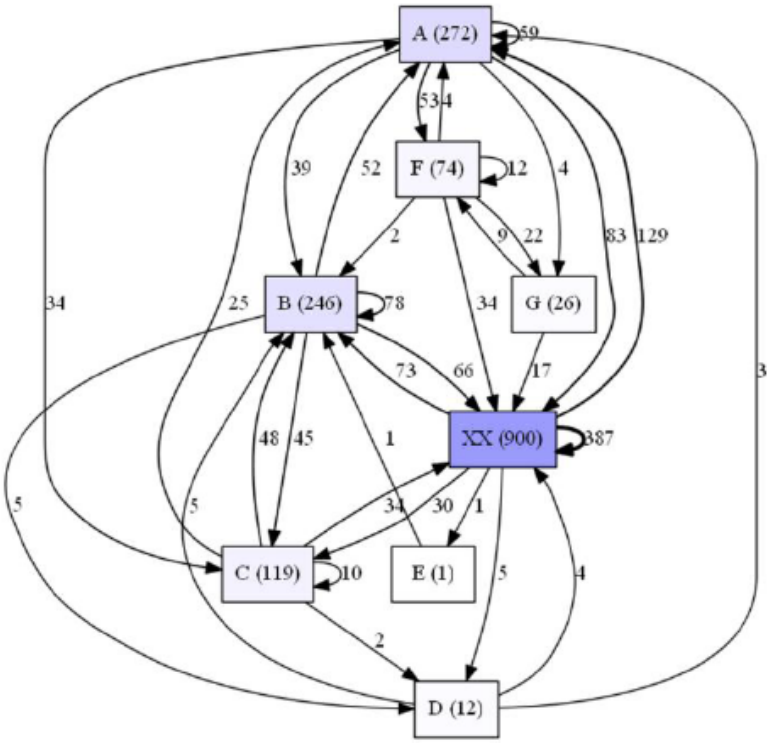
Em resumo, o Grafo DFG é uma ferramenta valiosa para a mineração de processos, pois oferece uma representação clara e direta das relações sequenciais entre atividades, sendo usado para identificar padrões, gargalos e oportunidades de melhoria em processos organizacionais e também uma parte essencial da presente dissertação.

3.4. *Directly Follows Graph* (DFG)

A ênfase do Grafo DFG está nas relações diretas entre atividades visto que não captura informações sobre *loops*, paralelismo ou outras complexidades no processo. Ele fornece uma representação visual direta das sequências de atividades em um processo, facilitando a interpretação e a análise. Para criar um Grafo DFG, normalmente, inicia-se com um *log* de eventos (um registro cronológico de todas as atividades executadas em um processo), em que cada entrada no *log* de eventos inclui informações como a atividade realizada e o momento em que ocorreu.

O Grafo DFG pode ajudar a identificar atividades que atuam como gargalos no processo, que ocorre quando uma atividade frequentemente é seguida por atrasos ou problemas, destacando áreas que podem ser otimizadas. Também pode revelar desvios em relação ao processo planejado, quando as sequências de atividades observadas não estão alinhadas às expectativas, destacando áreas de potencial melhoria, conforme o exemplo do artigo de (Dupuis, Dadouchi, & Agard, 2022).

Figura 7 - DFG com atividade 'XX'.



Fonte: (Dupuis, Dadouchi, & Agard, 2022, p. 5).

Pode-se observar, na *Figura 7,* que a atividade dominante, após a atividade 'XX', é a A (272 ocorrências), e que o conjunto de dados contém 39 transições da atividade A para a atividade B e 52 transições da atividade B para a atividade A, retornando para si mesmo (A), contando com 59 atividades. Essa estrutura do DFG é importante para ajudar na montagem da matriz de transição de Markov.

3.5. Cadeias de Markov

As cadeias de Markov são modelos matemáticos usados para descrever o comportamento dinâmico de sistemas estocásticos e modelar situações em que um sistema pode estar, sendo que, de modo crucial, a probabilidade de transição para um novo estado depende apenas do estado atual, sem levar em consideração os estados anteriores. Aplicadas em uma variedade de campos — estatística, física, economia, engenharia e ciência da computação, por exemplo —, elas são usadas para modelar a dinâmica de sistemas financeiros para prever o fluxo de tráfego de veículos, para analisar a dinâmica de populações biológicas e moldar o

comportamento de sistemas de comunicação. Além disso, são usadas para gerar texto, áudio e imagem usando sistemas de geração automática de conteúdo.

Segundo (Tomé, 2001), (Douc, Moulines, Priouret, & Soulier, 2018) e (Kemeny & Snell, 1960), as cadeias de Markov são processos estocásticos (probabilísticos), cujo estado seguinte, depende somente do estado atual, e não de estados anteriores. Assim, o estado pode ser anterior, atual e seguinte, tendo a transição e obtendo a formação de probabilidade. A definição de cadeia de Markov se dá por uma sequência de variáveis aleatórias, em que o conjunto de valores é chamado de "espaço de estado", onde X_t é chamado de processo de tempo t. Então:

Equação I:

$$Pr(X_{t+1} = X \mid X_0, X_1, X_2, \ldots, X_t) = Pr(X_{t+1} = X_t)$$

Para exemplificar a *Equação I*, suponha-se que se esteja modelando o preço da ação de uma empresa na bolsa de valores. Isso significa que o preço futuro da ação só depende do preço atual, e não das flutuações anteriores. O preço da ação é de R$ 100 por ação; $X_t = 100$. Agora, pretende-se calcular a probabilidade de que o preço da ação seja R$ 110 no próximo dia, $X_{t+1} = 110$.

$$Pr(X_{t+1} = 110 \mid X_t) = Pr(X_{t+1} = 110 \text{ e } X_t)/Pr(X_t)$$

Para calcular a probabilidade no numerador $Pr(X_{t+1} = 110 \mid X_t)$, precisa-se de dados históricos ou de um modelo que forneça essas probabilidades com base em observações passadas. Deste modo, sabe-se que:

$$Pr(X_{t+1} = 110 \text{ e } X_t = 100) = 0{,}2 \text{ (ou } 20\%)$$

Para calcular o denominador $Pr(X_t)$, usa-se a mesma lógica:

$$Pr(X_t = 100) = 0{,}5 \; (ou \; 50\%)$$

Agora, calcula-se a probabilidade condicional completa:

$$Pr(X_{t+1} = 110 \mid X_t) = (0,2) / (0,5) = 0,4$$

Diante disso, a probabilidade de que o preço da ação suba de R$100,00 para R$110,00 no próximo dia, dado que o preço atual é de R$100,00 será de um aumento de 0,4 (ou 40%) no preço da ação.

Já a matriz de transição é composta por um conjunto de estados e descreve as probabilidades de transição entre eles. Nesse aspecto, a probabilidade de transição entre dois estados é dependente apenas do estado atual, e não de estados passados. Diante disso, para obter uma matriz de transição, ou seja, uma matriz quadrada de tamanho $M_{\mathbf{kxk}}$, em que o k é o número de estados possíveis, pode-se representá-la da seguinte maneira:

Figura 8 - Matriz Quadrada.

$$M_{k\times k} = \begin{bmatrix} w_{11} & \cdots & w_{1j} \\ \vdots & \ddots & \vdots \\ w_{i1} & \cdots & w_{ij} \end{bmatrix}$$

Fonte: (Dupuis, Dadouchi, & Agard, 2022, p. 03)

Exemplificando uma matriz quadrada de ordem 3x3, na *Figura 8*, cada elemento é representado por um número, e os elementos estão dispostos em 3 linhas e em 3 colunas.

Figura 9- Exemplo de Matriz Quadrada.

$$A = \begin{bmatrix} 2 & 3 & 1 \\ 0 & 4 & 6 \\ 0 & 5 & 0 \end{bmatrix}$$

Fonte: (Novaes, n.d.)

Na *Figura 9*, uma matriz de transição é usada em teoria das probabilidades e em processos estocásticos, como as Cadeias de Markov, para representar as probabilidades de

transição de um estado para outro ao longo do tempo. Ela é definida da seguinte maneira: $P = [p(i,j)]$, onde $p(i,j)$ é a probabilidade do estado i para o estado j, dada por:

Equação II:

$$p(i,j) = P(S_{(t+1)} = j \mid S_t = i)$$

Onde S_t é o estado de tempo t e $S_{(t+1)}$ é o estado de $t + 1$. A probabilidade de transição deve ser calculada para cada par de estados (i,j).

Veja-se, a seguir, um exemplo de uma matriz de transição simples com três estados (A, B e C):

$$A = \begin{vmatrix} 0.5 & 0.2 & 0.3 \\ 0.4 & 0.3 & 0.3 \\ 0.1 & 0.5 & 0.4 \end{vmatrix}$$

Na matriz acima,

- A primeira linha representa as probabilidades de transição do estado A para os estados A, B e C. Por exemplo, a probabilidade de permanecer no estado A é 0,5, a probabilidade de ir para o estado B é 0,2 e a probabilidade de ir para o estado C é 0,3;
- A segunda linha representa as probabilidades de transição do estado B para os estados A, B e C;
- A terceira linha representa as probabilidades de transição do estado C para os estados A, B e C.

Além disso, é importante lembrar que a soma das probabilidades de transição para cada estado deve ser igual a 1.

Quando i e j representarem o estado, P_{ij} representa a probabilidade de observar a transição do estado i para o estado j. Essa afirmativa pode ser expressa matematicamente por: $P\left(X_{t+1} = j | X_t = i\right) = p_{ij}$. Se o conjunto de probabilidades, que representa as transições entre estados, é agrupado em uma matriz estocástica (chamada de matriz de transição (MT)), o vetor de estado no tempo $t + 1\left(Q_{t+1}\right)$ é capaz de prever usando o produto vetorial de estado no tempo *t*, como:

Equação III:

$$Q_{t+1} = Q_t \times MT$$

Em que

- Q_{t+1} representa o vetor de estado no tempo t+1;
- Q_t representa o vetor de estado no tempo t;
- MTrepresenta a matriz de transição transposta.

Exemplifica-se, simplificadamente, o comportamento de uma população de coelhos (R) e lobos (W) em uma ilha. Os estados possíveis são "R", para coelhos, e "W", para lobos.

$$MT = \begin{vmatrix} P(R \rightarrow R) & P(R \rightarrow W) \\ P(W \rightarrow R) & P(W \rightarrow W) \end{vmatrix}$$

Suponha-se que, no tempo t, tem-se o vetor de estado:

Q_t = [número de coelhos, número de lobos]

E a matriz de transição transposta é:

$$MT' = \begin{vmatrix} P(R \rightarrow R) & P(W \rightarrow R) \\ P(R \rightarrow W) & P(W \rightarrow W) \end{vmatrix}$$

Usando a *Equação III*, pode-se prever o vetor de estado no tempo $t + 1\left(Q_{t+1}\right)$ com base no vetor de estado atual Q_t e na matriz de transição transposta MT':

[Q_{t+1} para coelhos, Q_{t+1} para lobos] = [Q_t para coelhos, Q_t para lobos] * MT'

Isso permitirá prever que forma a população de coelhos e lobos mudará no próximo período de tempo com base nas probabilidades de transição entre os estados. Essa é a essência de como a matriz de transição e o vetor de estado são usados para modelar e prever sistemas estocásticos, como as populações de coelhos e lobos em uma ilha.

Em resumo, as cadeias de Markov são modelos matemáticos usados para descrever uma sequência de eventos, em que a probabilidade de cada evento ocorrer depende apenas do estado do sistema no momento atual, e não de sua história passada. Em outras palavras, as

cadeias de Markov são processos estocásticos sem memória que podem ser usados para prever a probabilidade de um futuro estado do sistema com base no estado atual.

A equação básica, para uma cadeia de Markov, é a equação de Chapman-Kolmogorov, que descreve a probabilidade de transição entre dois estados em um tempo *t*, com base na probabilidade de transição em um tempo anterior t-1. A equação é dada por:

Equação IV:

$$P_{(i,j)}^{(t)} = \sum_{(i,k)}^{(t-1)} \times P_{(k,j)(1)}$$

Onde:

- $P_{(i,j)}^{(t)}$ é a probabilidade de ir do estado *i* para o estado *j* em um tempo *t*;
- $P_{(i,k)}^{(t-1)}$ é a probabilidade de ir do estado *i* para o estado *k* em um tempo *t-1;*
- $P_{(k,j)}^{(1)}$ é a probabilidade de ir do estado *k* para o estado *j* em um tempo *1*.

Para exemplificar a *Equação IV*, suponha-se que se esteja modelando o tráfego em uma cidade com três estados de tráfego: Baixo (B), Médio (M) e Alto (A). Calcula-se $P_{(M,A)(2)}$, ou seja, a probabilidade de que o tráfego mude de Médio (M) para Alto (A) no tempo 2. Para tanto, imagine-se que as probabilidades de transição no tempo 1 sejam:

- $P_{(B,M)(1)} = 0,2$ (probabilidade de ir de Baixo para Médio);
- $P_{(M,A)(1)} = 0,3$ (probabilidade de ir de Médio para Alto).

E que as probabilidades de transição no tempo 2 sejam:

- $P_{(B,M)(2)} = 0,1$ (probabilidade de ir de Baixo para Médio);
- $P_{(M,A)(2)} = ?$.

Usando a *Equação IV*:

- $P_{(M,A)(2)} = \left(P_{(B,M)(1)} * P_{(M,A)(1)}\right) + \left(P_{(M,M)(1)} * P_{(M,A)(1)}\right) + \left(P_{(A,M)(1)} * P_{(M,A)(1)}\right)$

Substituindo as probabilidades conhecidas:

- $P_{(M,A)(2)} = (0,2 \ * \ 0,3) + \left(P_{(M,M)(1)} \ * \ 0,3\right) + \left(P_{(A,M)(1)} \ * \ 0,3\right)$

Para descobrir as probabilidades de transição em um tempo anterior ao *tempo 1* e calcular a probabilidade de transição no *tempo 2*, as probabilidades de transição de $P_{(M,M)}$ e de $P_{(A,M)}$ também precisam ser conhecidas, resultando, assim, em um cálculo $P_{(M,A)(2)}$ de maneira precisa.

Essa equação pode ser usada para calcular a probabilidade de transição entre dois estados em qualquer tempo *t*, desde que as probabilidades de transição em um tempo anterior *t-1* sejam conhecidas. A partir disso, é possível calcular a distribuição de probabilidade estacionária, que descreve a probabilidade de estar em cada estado em um estado de equilíbrio.

3.6. Considerações Finais

Diante do exposto, denota-se que esta seção explorou conceitos básicos e exemplificações de Mineração de Processos, *Logs* de Eventos, Modelo de Processo, DFG e Cadeias de Markov.

Nesse sentido, a primeira — mineração de processos — é uma disciplina que utiliza técnicas de análise de dados para extrair informações valiosas de *logs* de eventos em processos de negócios, os quais contêm informações sobre as ações realizadas, a ordem em que ocorrem e quando ocorrem. Seu objetivo é entender, otimizar e aprimorar os processos organizacionais para identificar gargalos, ineficiências e padrões de comportamento em processos, permitindo melhorias substanciais.

O Grafo DFG é uma ferramenta na área de mineração de processos que permite representar, graficamente, as relações sequenciais entre atividades em um processo. Nesse grafo, cada atividade é representada por um nó, e as setas direcionadas entre os nós mostram a ordem em que as atividades ocorrem. É uma maneira eficaz de visualizar padrões e relações de sequência em processos, identificando áreas de melhoria e análise de desempenho.

Finalmente, a cadeia de Markov é um modelo matemático que descreve um sistema estocástico, no qual as transições de estado do sistema dependem apenas do estado atual e não possuem memória de transições anteriores. Utilizando as propriedades de estados e transições constituídas por um conjunto finito de estados, a probabilidade de transição para o

próximo estado depende apenas do estado atual, pois são representadas as probabilidades das matrizes de transição entre estados. Essas cadeias são úteis para modelar sistemas que evoluem de maneira probabilística.

Diante disso, a presente dissertação abrange a seguinte proposta: desenvolver um método para prever e visualizar rotações de tipos de crimes, subsidiando discussões sobre predição de tipos de crimes/roubos em ambiente urbano e os métodos de predição. O modelo proposto combina técnicas de mineração de processos e cadeias de Markov, enquanto redes neurais recorrentes (RNN), LSTM e GRU foram utilizadas como ferramentas de comparação para avaliar a eficácia do método na captura de padrões temporais mais complexos. Na próxima seção, será explanado o método usado para compor a presente proposta.

4. MÉTODO

Conforme apresentado na Seção 2, do Mapeamento Sistemático da Literatura (MSL), sobre predição de tipos de crime para um ano *n,* percebeu-se que o mesmo problema pode ser amplamente encontrado nas literaturas. O objetivo do método proposto é predizer o tipo de crime que ocorrerá na próxima unidade de tempo *(n + 1)*, a partir do histórico de ocorrências policiais. Estas predições para o momento *(n+1)* e a visualização dos elementos, que levaram a essas predições, podem subsidiar discussões estratégicas.

Como mostrado na *Figura 1*, o método consiste em seis fases que levam a uma predição do tipo de crime a ocorrer no próximo momento (n + 1) e a um grafo que representa as relações entre os tipos de crime. O método se baseia em técnicas de mineração de processos e cadeias de Markov.

Para exemplificar, usou-se um conjunto de dados simplificado do artigo inspiração (Dupuis, Dadouchi, & Agard, 2022). Neste exemplo, os dados podem ser lidos da seguinte forma, conforme mostra a *Tabela 10*.

Tabela 10 - Representação da base de dados.

Case 1: (A1, 2012) → (A2, 2013) → (B, 2015) → (F, 2016)
Case 2: (G, 2015) →(F,2016) → (A1,2017)

Fonte: (Dupuis, Dadouchi, & Agard, 2022, p. 3)

Neste exemplo, os dados podem ser lidos da seguinte forma: na primeira realização do processo (Caso 1), a atividade A1 foi realizada em 2012, a atividade A2, em 2013 e, finalmente, as atividades B e F, em 2015 e 2016, respectivamente, e assim sucessivamente para os seguintes casos. Para o restante do exemplo, considera-se que duas atividades com a mesma letra (A1 e A2, por exemplo) são variantes da mesma atividade, de acordo com a *Tabela 11*.

Tabela 11 - Representação dos dados com valores omissos.

	Anos					
	2012	**2013**	**2014**	**2014**	**2016**	**1017**
Caso 1	A	A	XX	XX	B	F
Caso 2	XX	XX	XX	G	F	A
...	...	...	...	...	...	...

Fonte: (Dupuis, Dadouchi, & Agard, 2022, p. 4).

Tabela 12 - Representação dos dados preparados com L = 5 e W = 3.

	Lugar				
	Início	**1**	**2**	**3**	**Fim**
Caso 1 - W1	XX	A	A	XX	XX
Caso 1 - W1	XX	A	XX	XX	XX
Caso 1 - W1	XX	XX	XX	B	XX
...	...	...	...	...	...

Fonte: (Dupuis, Dadouchi, & Agard, 2022, p. 4).

De acordo com o artigo inspiração (Dupuis, Dadouchi, & Agard, 2022), o método apresentado na *Figura 1* foi aplicado ao histórico de cultivo de um total de 10.376 campos de 409 fazendas geolocalizadas em Quebec, Canadá, entre 2004 e 2020 para validação.

Esta seção apresenta uma contextualização sobre o método utilizado para predizer tipos de crimes, como a apresentação e preparação dos dados, criação de um grafo DFG, criação do modelo de predição, predição, avaliação do modelo e simplificação do grafo.

4.1. Fase 1: Preparação dos dados

De acordo com a metodologia de (Dupuis, Dadouchi, & Agard, 2022), a preparação de dados tratará principalmente de dados ausentes, duplicados e generalização de dados, a fim de obter uma tabela que possa ser usada em um algoritmo de mineração de processos. Nesse sentido, em primeiro lugar, os dados devem ser formatados de modo que, para cada registro, exista um identificador único do caso, da atividade e do marcador de tempo associado, pois, caso haja demasiada variabilidade nas atividades, alguma fusão pode ser realizada. Neste exemplo, notamos a existência das atividades 'A1' e 'A2', em que os dados são generalizados substituindo cada ocorrência de 'A1' e 'A2' pelo valor 'A'. Esta operação deve ser feita com

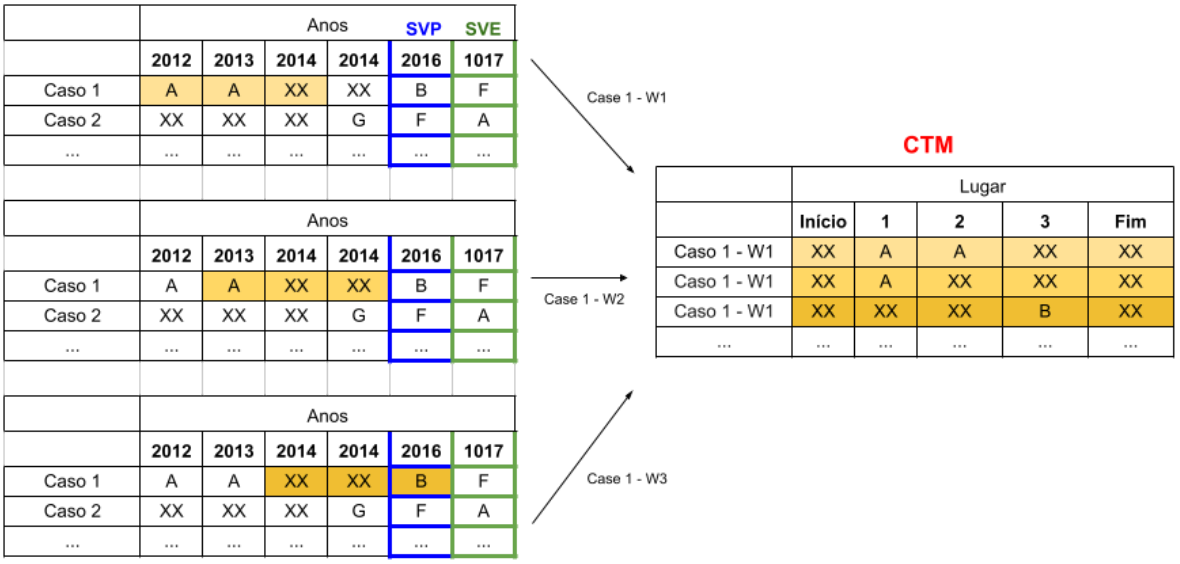
conhecimento do contexto e de acordo com as realidades que ligam as diferentes atividades (conhecimento especializado). No caso da rotação de culturas, o conhecimento agronômico permite agrupar culturas com comportamentos semelhantes.

Todas as duplicatas no conjunto de dados são removidas, e os dados são reordenados de acordo com os marcadores de tempo. A unidade de tempo utilizada é o ano, e o período estudado é de 2012 a 2017. Assim, para cada caso, as atividades devem ser associadas a cada ano do período [2012–2017]. Qualquer informação ausente será substituída por 'XX' (*Tabela 11 e Tabela 12*).

A *Tabela 11* mostra que, para o Caso 1, em 2012 e 2013, foi realizada a atividade A, enquanto faltam informações para 2014 e 2015. As atividades B e F foram realizadas em 2016 e 2017, respectivamente. Essa transformação nos permite identificar os valores ausentes.

Conforme mostrado na *Figura 1*, três conjuntos de dados são criados na fase de preparação de dados (CTM, SVP e SVE) para uso nas fases subsequentes da metodologia. O conjunto de dados SVP representa os vetores de estado do ano *n* usados para predição, enquanto o conjunto de dados SVE representa os vetores de estado do ano *n + 1* usados para avaliação. Esses dois conjuntos de dados correspondem às duas últimas colunas do conjunto de dados, conforme apresentado na *Tabela 11*. No exemplo, os dados de 2016 são usados para criar o conjunto de dados SVP, enquanto o conjunto de dados SVE é criado usando os dados de 2017 (*Figura 10*). O conjunto de dados SVE representa os valores que irão prever com as informações do SVP.

Figura 10 – Ilustração do desenvolvimento da tabela CTM.



Fonte: (Dupuis, Dadouchi, & Agard, 2022, p. 4).

O conjunto de dados CTM é usado para criar o primeiro grafo DFG, que propõe uma matriz adjacente e, em seguida, uma matriz de transição da atual inter-relações nos dados de MA. A matriz adjacente é uma matriz quadrada que resume o número de ocorrências entre dois estados no conjunto de dados. Para criar este conjunto de dados, dois hiperparâmetros devem ser definidos - sejam L o período considerado na criação do modelo e W o tamanho de uma janela móvel em L. O conjunto de dados CTM corresponde às sequências de tamanho W presentes no conjunto de dados original para um período L anterior ao ano a ser previsto (no exemplo, todos os dados de 2012 a 2016). Para marcar o início e o fim de cada sequência, duas colunas de valores 'XX' são adicionadas ao quadro de dados. Ao ser utilizado o exemplo anterior do Caso 1, com L = 5 e W = 3, então, os três primeiros registros do conjunto de treinamento obtido podem ser representados, conforme se vê na *Tabela 12*.

Tabela 13 - Representação Do Formato Final Do Caso 1.

Caso	Atividade	Tempo
Caso 1 - W1	XX	0
Caso 1 - W1	A	1
Caso 1 - W1	A	2
Caso 1 - W1	XX	3
Caso 1 - W1	XX	4
...	...	...

Fonte: (Dupuis, Dadouchi, & Agard, 2022, p. 5).

Figura 11 - Transformação da *Tabela 12* na *Tabela 13*.

	Lugar				
	Início	**1**	**2**	**3**	**Fim**
Caso 1 - W1	XX	A	A	XX	XX
Caso 1 - W1	XX	A	XX	XX	XX
Caso 1 - W1	XX	XX	XX	B	XX
...	...	...	...	...	...

↓

Caso	Atividade	Tempo
Caso 1 - W1	XX	0
Caso 1 - W1	A	1
Caso 1 - W1	A	2
Caso 1 - W1	XX	3
Caso 1 - W1	XX	4
...	...	...

Fonte: (Dupuis, Dadouchi, & Agard, 2022, p. 5).

Nesta fase, cada caso é transformado em múltiplas sequências de comprimento W representando a sequência de valores presentes na janela deslizante de tempo W, e cada sequência é instanciada por 'XX' simbolizando o início e o fim da sequência. Exemplificando, no Caso 1 (a primeira linha da *Tabela 11*), com L = 5 e W = 3, há três sequências distintas correspondentes aos valores da janela deslizante de tamanho W movendo-se sobre L. Assim, a primeira sequência corresponderá aos valores do primeiro caso nos anos 2012, 2013, 2014 (primeira linha da *Tabela 12*), a segunda sequência corresponderá aos valores do primeiro caso nos anos 2013, 2014, 2015 (segunda linha da *Tabela 12*) e a última sequência referente ao Caso 1 corresponderá aos valores dos anos 2014, 2015, 2016 (terceira linha da *Tabela 12*).

Outrossim, é importante observar que cada registro deve ser identificado exclusivamente. A etapa final, a fase de preparação de dados, consiste em reorganizar os dados preparados em um formato compatível com algoritmos de mineração de processos, ou seja, com uma representação explícita do identificador de caso, atividade e marcador de tempo, conforme mostra a *Tabela 13*. A transformação da Tabela *12* na *Tabela 13* é ilustrada na *Figura 11*.

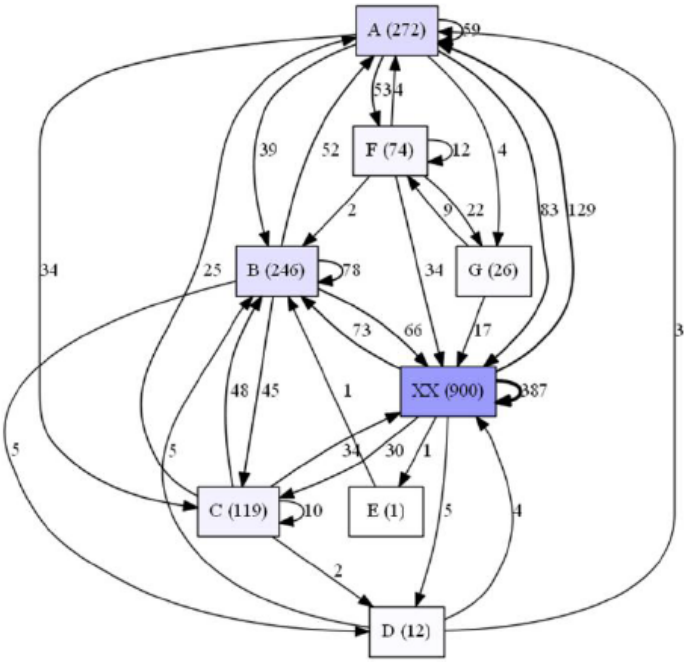
4.2. Fase 2: Criação de um grafo DFG

Uma vez que os dados estejam no formato correto e contenham todas as informações necessárias, a criação do gráfico DFG é relativamente rápida, conforme apresentado na seção anterior. Pois, a representação visual DFG de um grafo direcionado e ponderado em que cada aresta representa a relação direta entre duas atividades e o peso corresponde ao número de ocorrências dessa relação no conjunto de dados.

Sobre isso, o gráfico DFG, obtido com dados do exemplo CTM, *Figura 11*, é mostrado na *Figura 12*. A adição das colunas 'Início' e 'Fim' na fase de preparação de dados, bem como a identificação de dados ausentes, resulta em um gráfico balanceado.

No gráfico DFG, o número de transições entre duas atividades no conjunto de dados é representado pelo peso das arestas. A cor — associada a cada atividade — depende do seu número de ocorrências no conjunto de dados: quanto mais a atividade estiver presente no conjunto de dados, mais escura será sua cor. O número exato de ocorrências de cada atividade também está escrito dentro de cada retângulo, entre parênteses.

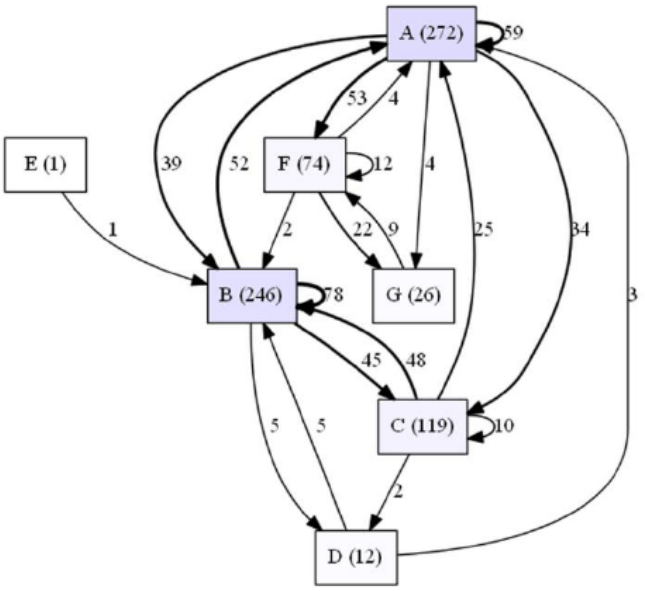
Figura 12 - DFG com atividade 'XX'.



Fonte: (Dupuis, Dadouchi, & Agard, 2022, p. 5).

Assim, no grafo da *Figura 12*, observa-se que a atividade dominante, após a XX, é a B, com 246 ocorrências, e nota-se que o conjunto de dados contém 66 transições da atividade B para a atividade XX e 73 transições da XX para a B, assim retornando para si mesmo, em que B conta com 78 atividades. No entanto, a informação XX não é útil em um contexto preditivo. Para uma análise mais aprofundada, é necessário removê-la e manter os pesos das arestas que conectam às outras atividades. A atividade XX é, portanto, filtrada apenas quando o DFG é criado (conforme *Figura 13*).

Figura 13 - DFG sem atividade 'XX'.



Fonte: (Dupuis, Dadouchi, & Agard, 2022, p. 5).

Ao comparar a *Figura 12* à *Figura 13*, pode-se notar que os pesos das arestas que conectam às diferentes atividades restantes permanecem inalterados. Assim, se na *Figura 12* a atividade A é mais frequentemente seguida pela atividade B em vez da atividade C, isso permanece verdadeiro na *Figura 13*. Desse modo, o grafo DFG é útil para entender a inter-relação entre os tipos de crimes.

Tabela 14 - Matriz Adjacente (MA) do grafo DFG *Figura 13*.

	A	**B**	**C**	**D**	**E**	**F**	**G**	
A	59	39	34	0	0	53	4	189
B	52	78	45	5	0	0	0	180
C	25	48	10	2	0	0	0	85
D	3	5	0	0	0	0	0	8
E	0	1	0	0	0	0	0	1
F	4	2	0	0	0	12	22	40
G	0	0	0	0	0	9	0	9

Fonte: (Dupuis, Dadouchi, & Agard, 2022, p.6).

A *Tabela 14* mostra as relações entre as atividades do grafo DGF em questão, em que a última coluna se configura como a soma dos valores das linhas. A tabela, nesta

configuração, será útil para montar a matriz de transição para o método de predição markoviano.

4.3. Fase 3 : Criação do modelo de predição

O modelo de predição, utilizado neste método, é baseado nos princípios de Markov. A matriz de transição pode ser estimada usando as probabilidades condicionais entre cada atividade como uma probabilidade de transição, ou seja, $P_{(j-i)}$ = probabilidade de obter o evento *j*, sabendo que o evento anterior é *i*. Estima-se, então, pelo método da máxima verossimilhança, $P_{(j-i)}$ como:

Equação V:

$$\widehat{P(j|i)} = \frac{W_{ij}}{\Sigma_j^n W_{ij}}$$

Onde:

$W_{i \to j}$ é o número de relações de *i* para *j*.

Como as frequências das arestas do grafo DFG representam o número de ocorrências de cada relação no conjunto de dados, a matriz adjacente pode ser criada extraindo cada peso em uma matriz quadrada por *n*, em que *n* é o número de atividades únicas presentes no conjunto de dados. A matriz adjacente pode ser expressa da seguinte forma:

Dicionário de Codificação
{A:0. B:1, C:2, D:3, E:4, F:5, G:6}

Para *n* atividades:

Equação VI:

$$MA_{nxn} = \begin{bmatrix} W_{11} & \cdots & W_{1i} \\ \cdots & \cdots & \cdots \\ W_{i1} & \cdots & W_{ii} \end{bmatrix}$$

Em que:

$$W_{ij} = \begin{cases} peso\ da\ aresta'\ i \rightarrow j \\ \forall i, j \in n \\ 0\ se\ não\ houver\ aresta \end{cases}$$

Essa matriz adjacente pode ser usada para estimar a matriz de transição usando a *Equação V.* Então:

Seja $MT_{(nxn)}$ a matriz de transição de *n* atividades,

Equação VII:

$$MT_{(nxn)} = \begin{bmatrix} \frac{W_{11}}{W_1} & \cdots & \frac{W_{1i}}{W_1} \\ \cdots & \cdots & \cdots \\ \frac{W_{i1}}{W_1} & \cdots & \frac{W_{ii}}{W_1} \end{bmatrix}$$

Onde:

$$P_{ij} = \begin{cases} P(j|i) = \frac{Wij}{\Sigma_j^n w_{ij}\ \forall i,j \in n} \\ 0\ então \end{cases}$$

O grafo DFG (*Figura 12*) é utilizado para gerar a matriz adjacente associada (*Tabela 14*) e, consequentemente, estimar sua matriz de transição (*Tabela 15*).

Tabela 15 - Matriz de Transição (MT) do grafo DFG *Figura 13*.

	A	B	C	D	E	F	G	
A	31,22%	20,63%	17,99,%	0,00%	0,00%	28,04%	21,20%	100,00%
B	28,89%	43,33%	25,00%	2,78%	0,00%	0,00%	0,00%	100,00%
C	29,41%	56,47%	11,76%	2,35%	0,00%	0,00%	0,00%	100,00%
D	37,50%	62,50%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%
E	0,00%	100,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%
F	100,00%	50,00%	0,00%	0,00%	0,00%	30,00%	55,00%	100,00%
G	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%	0,00%	100,00%

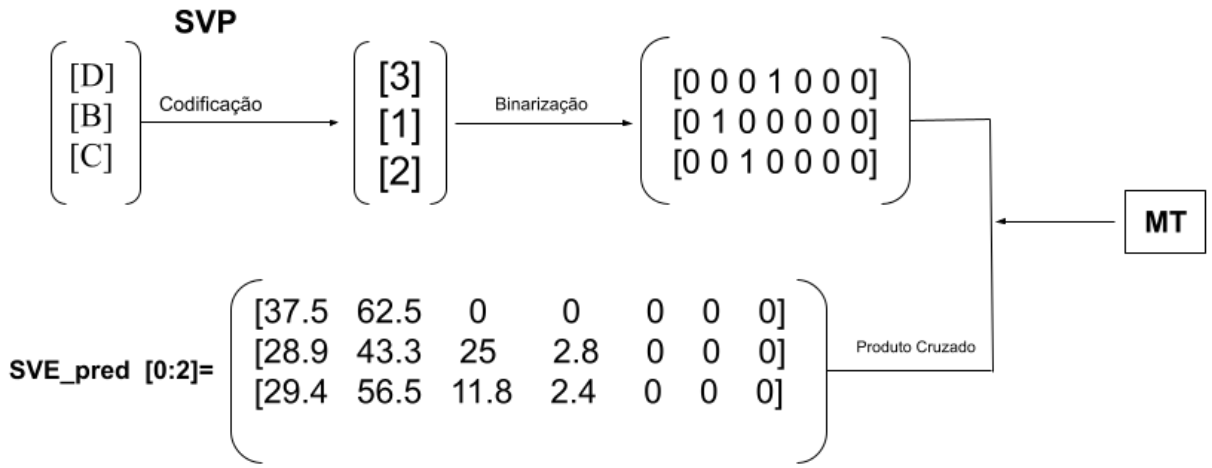
Fonte: (Dupuis, Dadouchi, & Agard, 2022, p. 6).

De acordo com a matriz de transição da *Tabela 15*, a probabilidade de A ser seguida pela atividade B é de 20,63%. Da mesma forma, a probabilidade de C ser seguida pela atividade D é de 2,35%. A matriz de transição também permite representar "transições impossíveis". Por exemplo, a probabilidade de que a atividade E seja seguida por qualquer outra atividade além de F é nula.

4.4. Fase 4: Predição

Para a fase de predição, uma matriz de transição previamente criada (*Tabela 15*) é usada com o conjunto de dados SVP e com um dicionário de codificação. Este, por sua vez, pode ser criado associando um número a cada atividade presente no conjunto de dados, sendo que, para usos práticos, as atividades podem ser classificadas em ordem crescente. No exemplo proposto, cinco atividades únicas são representadas e podem ser associadas ao dicionário de codificação mostrado acima, em que o objetivo é obter um vetor probabilístico usando o vetor SVP e a matriz de transição. Para isso, o conjunto de dados SVP deve ser transformado como um conjunto de dados de vetores binários.

Figura 14 - Ilustração do Processo de Predição.



Fonte: (Dupuis, Dadouchi, & Agard, 2022, p. 6).

A primeira etapa trata da codificação dos valores presentes no conjunto de dados SVP usando o dicionário de codificação criado durante as etapas de preparação. Os dados codificados são, então, transformados em um vetor binário de tamanho *n*, no qual o valor 1 se encontra indicado pelo valor codificado.

Finalmente, o vetor binário é multiplicado pela matriz de transição MT, usando o produto vetorial. O resultado obtido é o vetor probabilístico SVE-pred contendo a probabilidade de pertencimento de cada classe no tempo *n + 1*. As etapas de predição são ilustradas para os três primeiros registros de SVP na *Figura 14*.

O resultado obtido pode ser entendido da seguinte forma: se no tempo *t* a atividade D for realizada, então, no tempo *n + 1,* a atividade A será realizada com uma probabilidade de 37,5% e a atividade B, de 62,5%. Da mesma forma, se a atividade B for realizada no tempo *t*, as probabilidades das atividades A, B, C ou D serem realizadas no tempo *t + 1* são de, respectivamente, 28,9%, 43,4%, 25% e, finalmente, 2,8%, e assim por diante.

4.5. Fase 5: Avaliação do modelo

A predição resultante é um vetor de probabilidade que associa cada atividade possível à probabilidade de que ela ocorra no tempo *n + 1*. As atividades podem, portanto, ser classificadas de acordo com sua probabilidade de ocorrência; por isso, é crucial a obtenção de uma lista das *N* atividades mais prováveis (*Top-N Activities*).

Equação VIII:

$$Recall\ (N)\ =\ \frac{\sum_{k}^{|SVEprd|} TP_k}{|SVEpred|}$$

Equação IX:

$$Precision\ (N)\ =\ \frac{Recall\ (N)}{N}$$

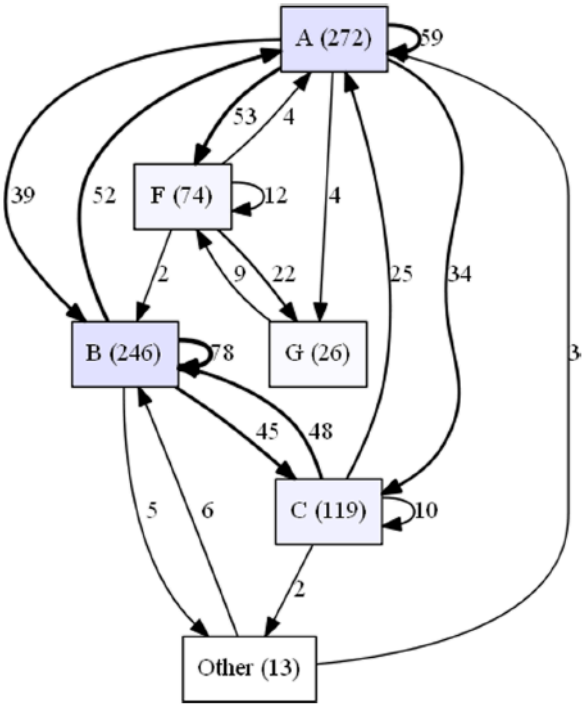
Onde:

$|SVEpred|\ =\ número\ de\ registros\ no\ conjunto\ de\ dados\ SVE.$

$$TP_k = \begin{cases} 1\ se\ o\ SVEpred_k \in Top \\ 0\ então; \end{cases}$$

N = número de atividades recomendadas na lista *Top - N*.

Figura 15 - Transformação do grafo DFG *Figura 13* com o agrupamento de operações (SG = 2%).



Fonte: (Dupuis, Dadouchi, & Agard, 2022, p. 7).

Um modelo de predição apropriado coloca, em correspondência, as métricas de *Recall(N)* e *Precision(N)*. À proporção que *N* aumenta, *Recall(N)* também aumenta, enquanto a métrica *Precisão (N)* diminui. Quando *N* for muito alto, a medida *Recall(N)* sempre estará em valor superior, enquanto a *Precisão (N)* será menor. Assim, à medida que *N* aumenta, o limite superior da precisão *(N)* diminui. Dada esta definição, o limite superior do A *precisão (N)* para $N = 1$, 2 ou 3 é, respectivamente, 100%, 50% e 33%.

4.6. Fase 6: Simplificação do grafo

Para melhorar a utilização dos grafos DFG na análise de dados complexos, é essencial desenvolver métodos que aprimorem sua legibilidade sem comprometer a integridade das informações representadas. Pois, a operação de agrupamento visa simplificar a visualização ao combinar nós semelhantes ou sequências de atividades, reduzindo assim o número de elementos no gráfico e permitindo uma interpretação mais direta das tendências e padrões predominantes.

Já a operação de filtragem tem o intuito de remover elementos menos significativos ou ruídos que podem obscurecer as relações essenciais, destacando as conexões mais fortes e relevantes. A combinação dessas duas operações tem o potencial de transformar os grafos DFG de ferramentas abstratas e, muitas vezes, intransponíveis, em representações claras e acessíveis, facilitando a tomada de decisões baseada em dados e em extração de *insights* cruciais para diversas aplicações.

4.6.1. Operação de agrupamento

A finalidade da operação de agrupamento é reunir atividades minoritárias em uma única categoria, cuja ação permite manter os fluxos existentes entre as atividades, limitando a diversidade das atividades representadas. No entanto, caso seja impossível utilizar agrupamento baseado no conhecimento especializado, é possível definir um limite de agrupamento que permita agrupar as atividades de acordo com suas representações na base de dados. Assim, qualquer atividade, cuja taxa de ocorrência na base de dados seja inferior a um limite fixo de SG, agrupa-se na mesma categoria — por exemplo, a categoria “outros”.

No conjunto de dados de exemplo, são apresentadas 750 atividades. Se o limite de agrupamento SG for fixado em 2%, todas as atividades com menos de 15 ocorrências serão

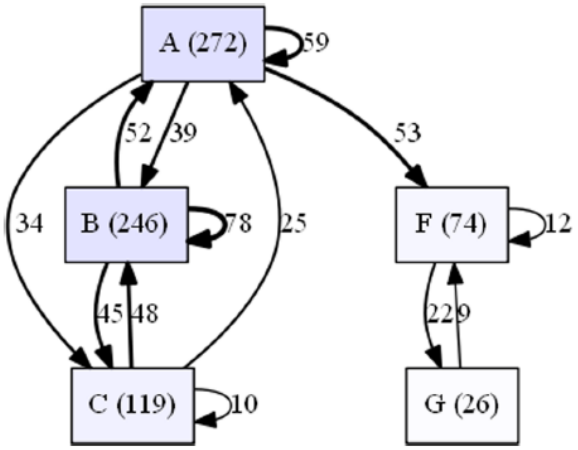
agrupadas como “outras”; daí o agrupamento das atividades E e G como “outras” (*Figura 15*). Ao comparar o gráfico da *Figura 13* com o gráfico obtido, após as operações de agrupamento (*Figura 15*), nota-se uma simplificação na forma como os gráficos podem ser lidos.

Observa-se também que os fluxos e o número de ocorrências das atividades desagrupadas permanecem idênticos; já os fluxos e o número de ocorrências das atividades agrupadas são somados e permitem ampliar o número de ocorrências das transições entre as atividades da base de dados e o agrupamento das atividades menos representadas.

4.6.2. Operação de filtragem

A segunda operação de simplificação consiste em filtrar as arestas de peso relativamente baixo para manter apenas as relações mais frequentes. Para tanto, um limiar de filtragem, SF, deve ser escolhido para remover todas as arestas para as quais a taxa de ocorrência na base de dados é menor do que o SF fixo do grafo DFG. No conjunto de dados do exemplo, 512 transições são apresentadas. Se o limite de filtragem SF for definido em 1%, todas as arestas com menos de 6 ocorrências serão removidas (*Figura 16*).

Figura 16 - Transformação do grafo DFG *Figura 13* com o filtragem de operações (SF = 1%).



Fonte: (Dupuis, Dadouchi, & Agard, 2022, p. 8).

Essa operação de filtragem causa uma perda de informações que pode afetar a qualidade da predição. Por isso, a escolha do limiar de filtragem deve ser feita de acordo com um dado objetivo, o qual, conforme mencionado no início da Seção 1, busca propor uma predição, mas também explicar a origem do resultado previsto. Necessita-se, portanto,

encontrar um compromisso entre a simplificação dos grafos (pelos limites SF e SG) e as medidas de desempenho do modelo *Recall(N)* e *Precision(N)* definidas nas *Equações VIII e IX*.

Ademais, a escolha de um limite de filtragem de 1% permite que o gráfico DFG seja bastante simplificado, comparando as *Figuras 13 e 16*, sem causar uma perda significativa de informações, em que a medida *Recall Top 3* diminui cerca de 3%. A escolha do limite de filtragem pode, portanto, ser selecionada de acordo com o objetivo.

4.7. Considerações Finais

Nesta seção, explicou-se de que forma o método será empregado. A primeira fase — preparação dos dados — tratará de dados que possam ser duplicados, dados ausentes e uma generalização dos dados, a fim de obter uma tabela que possa ser utilizada em um algoritmo de mineração de processos. Já na segunda fase, os dados dispõem no formato correto, obtêm as informações necessárias e contemplam a criação do grafo DFG. Assim, a representação visual é dada por grafo DFG direcionado, em que cada aresta representa a relação entre duas atividades, e seu peso corresponde ao número de ocorrências dessa relação no conjunto de dados. A terceira fase, de criação de um modelo de predição, se dá pelo método aplicado, seguindo os princípios de uma cadeia de Markov e gerando uma matriz de transição para representar a probabilidade de transição entre os tipos de crimes. Na quarta fase de predição, a matriz, previamente criada a partir da matriz de transição, é usada como um conjunto de dados e um dicionário de codificação, associando um número a cada atividade presente no conjunto de dados. A quinta fase, a avaliação do modelo, consiste em um vetor de probabilidade associado a cada atividade com a probabilidade de que ocorra no tempo $(n + 1)$, sendo que as atividades podem ser classificadas de acordo com a probabilidade da ocorrência. Na sexta e última fase, a simplificação do grafo fornece *insights* para explicar a origem dos resultados previstos: a utilização desse grafo DFG permite destacar relações mais frequentes entre as atividades.

Para uma análise comparativa mais robusta, redes neurais, incluindo RNN, GRU e LSTM, foram aplicadas ao estudo. Métricas como *Precision e Recall* também foram calculadas para avaliar o desempenho dos modelos e garantir resultados comparativos consistentes.

5. CASO DE ESTUDO

Neste estudo, apresenta uma análise detalhada de um conjunto de dados criminais da cidade de Denver, obtido pela plataforma *Kaggle* (Kumar, n.d.). Este conjunto de dados contém 512.657 registros de ocorrências entre 2015 e 2019 e fornece informações sobre a natureza dos crimes, incluindo data, horário e localização exata (ver *Tabela 16*). Por outro âmbito, a estrutura e a riqueza dos dados permitem uma investigação sistemática das dinâmicas criminais em diferentes regiões da cidade, viabilizando a identificação de áreas de alta incidência criminal e padrões temporais associados. Adicionalmente, a inclusão de múltiplas categorias de crimes no conjunto de dados enriquece a análise, fornecendo subsídios para o planejamento estratégico e a alocação eficiente de recursos por parte das autoridades responsáveis. Essa abordagem também contribui para a geração de *insights* sobre fatores relacionados à criminalidade, permitindo explorar suas possíveis causas e impactos no contexto urbano. O estudo, portanto, reforça a importância de dados abrangentes e organizados para o desenvolvimento de políticas públicas baseadas em evidências e a formulação de estratégias eficazes de prevenção e de combate ao crime.

Sob esse viés, a integridade e a qualidade dos dados foram asseguradas por meio de um processo rigoroso e abrangente de preparação. Este processo incluiu a padronização dos registros, abrangendo a remoção de atributos irrelevantes, o tratamento de valores ausentes e a eliminação de duplicatas. Como parte desse procedimento, 47.886 registros duplicados foram identificados e removidos, resultando em uma redução do conjunto de dados para 464.771 registros únicos. Para manter a completude dos dados, valores ausentes foram substituídos por um *placeholder* específico ("XX"), o que preserva a consistência e permite a continuidade analítica em etapas posteriores. Adicionalmente, as variáveis de data foram uniformizadas no formato AM/PM, visando aprimorar a manipulação e a interpretação temporal das informações. Em suma, este processo assegura um conjunto de dados robusto e confiável para análises subsequentes.

Nesse sentido, o *dataset,* utilizado neste estudo, foi estruturado com 12 atributos específicos que facilitam a identificação e a classificação das ocorrências criminais dentro do conjunto de dados. Cada atributo representa uma característica relevante dos crimes registrados, como o tipo de crime e o tempo de ocorrência, permitindo uma categorização e

análise detalhada dos dados. Na configuração dos *hiperparâmetros*, utilizamos L para representar o período analisado (mês, porcentagem (25%, 50%, 75%), estações do ano, dias da semana/finais de semana, W para a janela deslizante, podendo variar, conforme a necessidade de cálculo dos dados.

Tabela 16 - Descrição detalhada dos atributos do conjunto de dados de crimes da cidade de Denver.

Atributo	Descrição
INCIDENT_ID	Um identificador único para cada incidente de crime registrado;
OFFENSE_ID	Número único para cada infração, diferenciando entre várias ofensas no mesmo incidente;
OFFENSE_CODE	Código numérico designado para a ofensa/crime específico.
OFFENSE_CODE_EXTENSION	Fornece mais especificidade;
OFFENSE_TYPE_ ID	Tipo específico de crime;
OFFENSE_CATEGORY_ID	Classifica os crimes em categorias gerais;
FIRST_OCCURRENCE_DATE e LAST_OCCURRENCE_DATE	Indicam quando o crime começou e terminou, respectivamente;
REPORTED_DATE	Data em que o crime foi reportado;
INCIDENT_ADDRESS	Fornece o endereço do incidente;
GEO_X e GEO_Y	Representa números inteiros, de coordenadas X e Y geográficas em um sistema de referência específico.
GEO_LON e GEO_LAT	Representa números decimais, em formato decimal;
DMS_LONGITUDE e DMS_LATITUDE	Representa longitudes em formato de graus, minutos e segundos.
DISTRICT_ID e PRECINCT_ID	Identificam o distrito e o setor policial responsáveis pela área do incidente;
NEIGHBORHOOD_ID	Indica o bairro onde ocorreu o incidente;
IS_CRIME	Indica se o incidente é considerado um crime, com 1 para ocorrido, e 0 para não ocorrido;
IS_TRAFFIC	Indica se está relacionado ao trânsito, com 1 para ocorrido, e 0 para não ocorrido.

Fonte: a Autora.

A *Tabela 17* apresenta um resumo das configurações dos conjuntos de dados utilizados nos experimentos, especificando o número de registros e as granularidades temporais aplicadas. Cada conjunto foi estruturado para tentar capturar aspectos temporais dos dados de crimes e permitir uma análise preditiva detalhada e contextualizada.

Tabela 17 - Descrição dos conjuntos de dados utilizados nos experimentos.

Dataset	**Descrição**	**N. registros de dados limpos**	**SVP - 2018**	**SVE - 2019**	**CTM (conjunto de dados derivado)**	**Granularidade**
D-0	Dados originais	464.772	84.493	81.045	1.048.576	mês
D-1	25% de D-0	116.193	21.133	20.239	722.100	mês
D-2	50% de D-0	232.386	42.234	40.373	1.048.576	mês
D-3	75% de D-0	348.579	63.288	60.852	1.048.576	mês
D-4	Dados originais	464.772	84.493	81.045	413.000	Estações do ano
D-5	Dados originais	464.772	84.493	81.045	413.000	Dias da semana e finais de semana.

Fonte: a Autora.

Para suportar uma análise preditiva eficaz e para garantir a qualidade dos resultados, os conjuntos de dados foram segmentados em diversas granularidades temporais e tamanhos de amostra. Os principais conjuntos de dados incluem:

Esta configuração resultou no conjunto de dados CTM (***C****urrent* ***T****ransition* ***M****atrix*), criado especificamente durante a etapa de preparação para modelar transições entre tipos de crimes, com o objetivo de permitir uma análise inicial preditiva. Nesse aspecto, cada registro do CTM inclui dois atributos essenciais, *"start"* e *"end"*, que representam o ponto inicial e final de cada transição, além da janela deslizante que facilita a análise temporal e categórica dos crimes. Como se apresenta, esse conjunto é útil para algoritmos de modelagem sequencial, uma vez que permite capturar padrões temporais, recorrentes nas ocorrências criminais. Adicionalmente, foi incorporada uma janela deslizante configurável, permitindo uma análise detalhada das dinâmicas temporais e das categorias criminais.

O conjunto CTM é adequado para algoritmos de modelagem sequencial, devido à sua capacidade de identificar padrões temporais recorrentes em séries históricas de crimes. Para potencializar essa capacidade, foi aplicada a granularidade mensal aos conjuntos de dados

D-0 a D-3, os quais serviram como base para a estruturação do CTM. Essa granularidade permite ao modelo observar e aprender padrões periódicos e sazonais, mesmo em volumes de dados variáveis, fortalecendo a robustez das análises e predições realizadas.

Os conjuntos SVP (***S**tate **V**ectors for Prediction*) e SVE (***S**tate **V**ectors for **E**valuation*) foram preparados com dados específicos dos anos de 2018 e 2019, respectivamente. Essa divisão temporal possibilita o treinamento do modelo com os dados de um ano (SVP) e sua validação com os dados do ano subsequente (SVE), capturando variações temporais e garantindo uma avaliação balanceada. Essa abordagem é para medir a consistência e a precisão do modelo ao longo do tempo, refletindo sua capacidade de generalização.

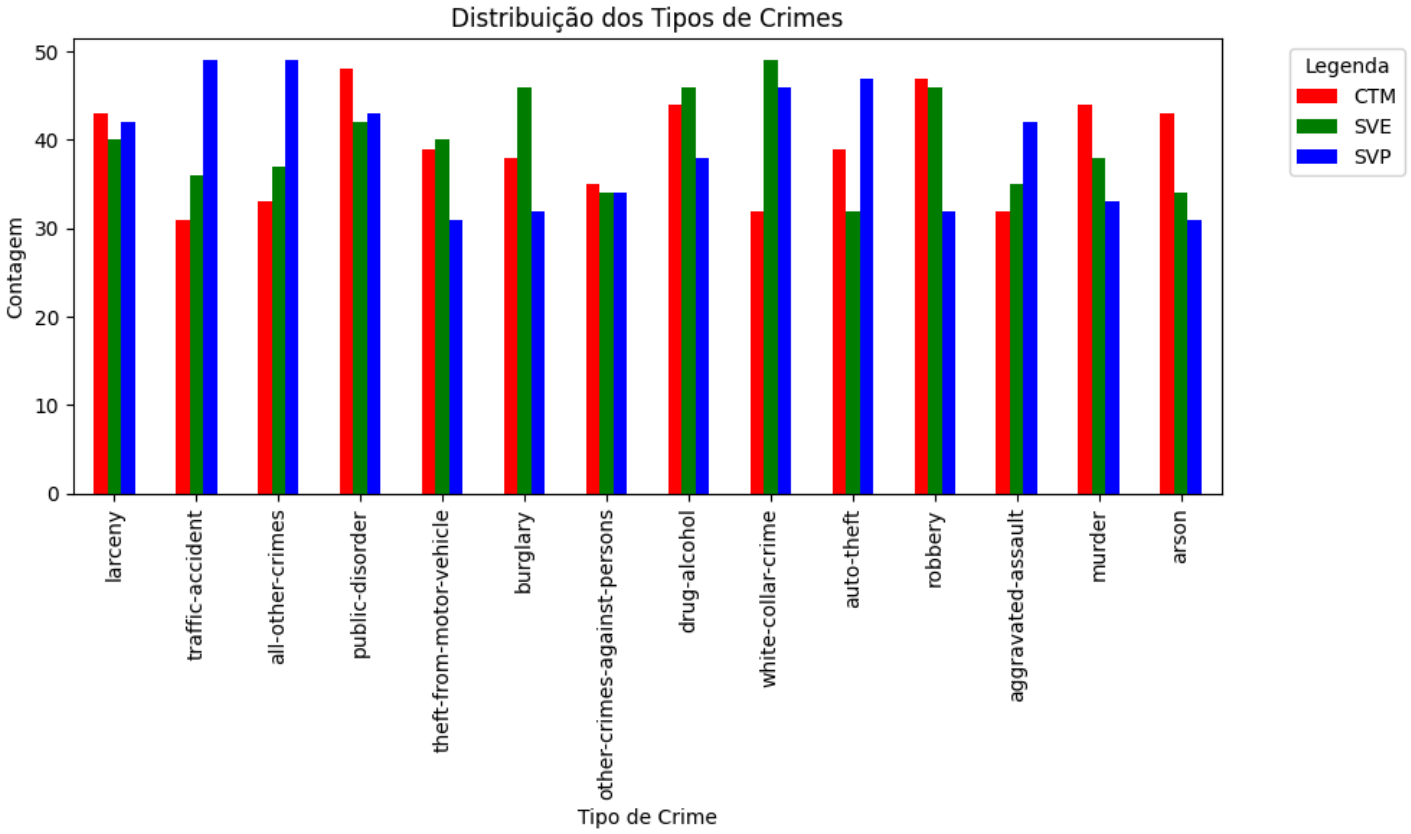
Adicionalmente, para investigar o impacto de diferentes granularidades temporais nos padrões criminais, o conjunto **D-0** foi reorganizado em diversas segmentações: sazonal (**D-4**, estações do ano) e semanal (**D-5**, dias úteis e finais de semana). Essas segmentações fornecem *insights* sobre variações cíclicas nos padrões criminais, permitindo identificar flutuações sazonais e comportamentais específicas.

Outrossim, as segmentações **D-1** a **D-3**, com 25%, 50%, e 75% dos dados originais, foram incluídas para avaliar a resiliência do modelo diante de amostras menores, enquanto **D-4** e **D-5** oferecem uma análise aprofundada das variações sazonais e de curto prazo. Essa estrutura possibilita avaliar a influência de fatores externos, como sazonalidade e padrões comportamentais, fortalecendo a robustez dos modelos preditivos.

Ao capturar nuances temporais e contextuais, as granularidades aplicadas permitem que os modelos identifiquem padrões recorrentes que seriam negligenciados em abordagens mais simplificadas. Consequentemente, essa metodologia aprimora a capacidade dos modelos de prever a criminalidade em diferentes contextos temporais, aumentando sua aplicabilidade em cenários reais de segurança pública.

A *Figura 17* apresenta um gráfico de barras que ilustra a distribuição dos tipos de crimes nos três conjuntos de dados principais.

Figura 17. Distribuição de tipos de crimes dos três conjuntos de dados.



Fonte: a Autora.

Observa-se uma forte consistência entre os três conjuntos de dados para determinadas categorias de crimes, incluindo *larceny* (furto), *traffic-accident* (acidente de trânsito), *public-disorder* (desordem pública), *auto-theft* (roubo de automóvel) e *aggravated-assault* (agressão agravada). Ademais, a estabilidade nas contagens para esses tipos de crimes sugere que eles mantêm uma taxa de ocorrência relativamente constante entre os anos de 2018 e 2019, indicando que esses crimes podem ter padrões temporais ou sociais estáveis na cidade de Denver.

Para certas categorias de crimes, como *theft-from-motor-vehicle* (roubo de itens de veículos), *burglary* (invasão domiciliar), *robbery* (roubo), *white-collar-crime* (crime de colarinho branco) e *murder* (assassinato), observam-se variações mais significativas nas contagens entre os conjuntos SVP (2018) e SVE (2019), indicando flutuações temporais entre os anos analisados. Essas diferenças podem refletir tanto mudanças nos comportamentos criminais quanto variações nos métodos de registro ou aplicação da lei. Em particular, as contagens mais altas de *traffic-accident* e *public-disorder* no conjunto SVP em relação ao

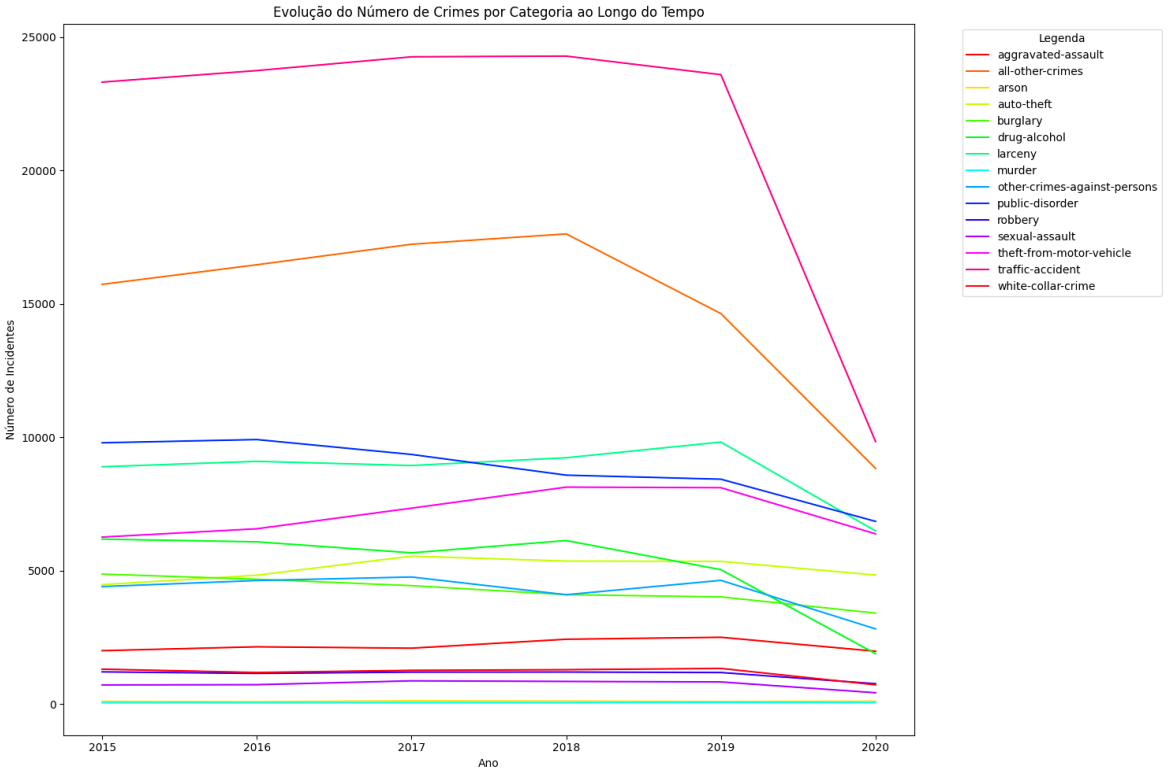
SVE sugerem uma redução na incidência desses crimes em 2019, possivelmente devido a políticas de segurança pública introduzidas ou a eventos específicos que influenciaram suas ocorrências.

Nas categorias *aggravated-assault* (agressão agravada) e *auto-theft* (roubo de automóvel), o conjunto CTM apresenta valores intermediários em comparação aos conjuntos SVP e SVE, pois, sugere que ele capture uma média das transições de ocorrências entre os anos de 2018 e 2019, atenuando as variações anuais. Essa característica posiciona o CTM como uma ferramenta para o treinamento de modelos que buscam capturar padrões temporais dinâmicos e transições de estado, uma vez que oferece uma base de dados mais equilibrada e robusta para a modelagem de sequência. Ao suavizar as flutuações entre períodos específicos, o CTM permite que o modelo identifique tendências de longo prazo e transições entre tipos de crimes de maneira mais estável e previsível.

A análise da distribuição dos tipos de crimes nos conjuntos SVP, SVE e CTM proporciona ajudar no desenvolvimento de modelos preditivos robustos. A consistência observada nas categorias mais frequentes, como *larceny* (furto) e *traffic-accident* (acidente de trânsito), indica que essas categorias possuem uma taxa de ocorrência estável ao longo do tempo, o que permite previsões com alta precisão. Em contrapartida, as variações detectadas em categorias menos comuns, como *white-collar-crime* (crime de colarinho branco) e *murder* (assassinato), ressaltam a necessidade de estratégias de ajuste dinâmico no modelo, permitindo que ele se adapte a flutuações sazonais e a mudanças anuais nas ocorrências criminais.

Para complementar essa análise, o gráfico *Figura 18* apresenta a evolução do número de crimes por categoria ao longo do tempo, cobrindo o período de 2015 a 2020. Este gráfico mostra as tendências de incidência para cada tipo de atividade criminal, permitindo observar padrões temporais e variações nas ocorrências, elementos essenciais para ajustes e validação dos modelos preditivos desenvolvidos.

Figura 18 - Evolução dos tipos de crimes.



Fonte: a Autora.

Observa-se que as categorias de crimes com maior frequência de ocorrência ao longo do período analisado — "traffic accident" "all other crimes" — mantêm altos níveis de incidentes até 2019, seguidos por uma queda acentuada em 2020. Esse declínio do ano de 2020 está relacionado a alguns valores no *dataset* original. Por outro aspecto, algumas categorias *"public disorder", "drug-alcohol"* e *"larceny",* mostram estabilidade relativa, com variações moderadas, indicando um padrão de ocorrência previsível e, possivelmente, associado a fatores sociais e econômicos constantes.

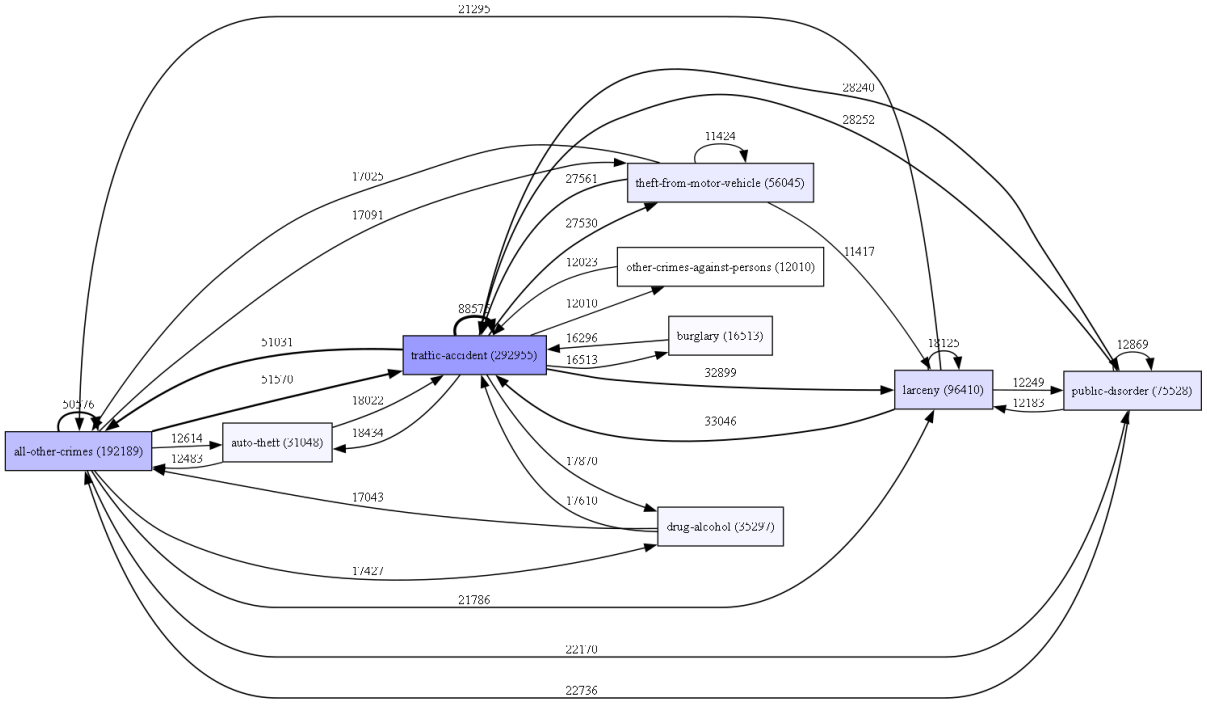
Em contrapartida, categorias, como *"public disorder", "drug-alcohol"* e *"larceny"*, mostram uma estabilidade relativa ao longo dos anos, com flutuações moderadas, o que sugere um padrão de ocorrência previsível, possivelmente associado a fatores sociais constantes. Essas categorias, por sua natureza, não parecem ser fortemente afetadas por fatores pontuais, mantendo uma taxa de ocorrência mais regular. Outras categorias, como *"sexual assault"* e *"aggravated assault"*, apresentam um crescimento leve, sugerindo uma

possível tendência emergente ou maior eficiência no registro e resposta a esses crimes. A partir de 2019, diversas categorias registram uma redução significativa, particularmente em crimes relacionados a atividades públicas e interações sociais, indicando uma mudança estrutural nos padrões de criminalidade, possivelmente devido à mobilidade reduzida.

Essas observações sobre a estabilidade e o crescimento de determinadas categorias de crimes, bem como a redução em outras, oferecem um panorama essencial para compreender a distribuição dos tipos de crimes nos conjuntos CTM, SVP e SVE. A análise da frequência e da evolução temporal de cada categoria ao longo de diferentes períodos fornece percepções sobre a dinâmica temporal desses fenômenos criminais. Complementarmente, a investigação das interações entre as atividades criminosas, por meio do grafo de seguimento direto (*Directly Follows Graph* – DFG), aprofunda essa análise ao identificar padrões sequenciais de ocorrência. Essa abordagem não apenas caracteriza a frequência das atividades, mas também elucida as relações dinâmicas entre elas, evidenciando como determinados crimes apresentam maior propensão a suceder outros. Esse detalhamento contribui para uma compreensão mais robusta e fundamentada das estruturas e comportamentos subjacentes ao fenômeno criminal.

No DFG, cada nó corresponde a uma atividade criminal específica, enquanto as arestas direcionadas representam transições diretas entre essas atividades. Os valores numéricos associados às arestas indicam a frequência com que essas transições ocorrem, conforme ilustrado na *Figura 19*. Essas conexões são fundamentais para a análise de tendências de sucessão entre crimes, permitindo identificar padrões recorrentes e inferir relações comportamentais subjacentes entre diferentes categorias de delitos. A intensidade das transições expressa não apenas a prevalência de determinadas sequências de crimes, mas também fornece *insights* sobre dinâmicas temporais e contextuais que podem auxiliar na compreensão e predição de fenômenos criminais.

Figura 19 - Grafo de crimes urbanos destacando a centralidade da categoria desordem pública ("*public disorder*") como um nó intermediário estratégico, com alta conectividade e influência nas transições entre diferentes tipos de crimes, incluindo acidentes de trânsito ("*traffic accident*") e 'incidentes relacionados a drogas e a álcool' ("*drug-alcohol*"). O grafo foi gerado a partir de uma **redução de 85%** do dataset D-0.



Fonte: a Autora.

O grafo de crimes urbanos, apresentado na *Figura 19*, revela a presença de categorias criminais que desempenham papéis centrais devido à elevada conectividade e à função de intermediação no fluxo de transições entre diferentes tipos de crimes. Dentre essas categorias, destaca-se a desordem pública ("*public disorder"*), identificada como um nó crítico no grafo. Essa categoria atua como ponto de convergência para múltiplas infrações, desempenhando tanto o papel de origem para crimes subsequentes quanto o de destino de eventos precursores.

A análise estrutural do grafo evidencia que a desordem pública possui um número significativo de conexões de entrada e de saída, consolidando-se como um evento de transição altamente influente. Essa centralidade funcional permite que a categoria module e amplifique fluxos criminais, promovendo interações que, frequentemente, resultam em crimes associados, como acidentes de trânsito ("*traffic accident"*) e incidentes relacionados ao consumo de drogas e de álcool ("*drug-alcohol"*).

A posição estratégica da desordem pública no grafo sublinha sua relevância como um nó intermediário essencial, com potencial para impactar a dinâmica de ocorrência e a

propagação de outros crimes. Esse papel intermediário destaca a necessidade de atenção especial a essa categoria no desenvolvimento de estratégias preventivas e no planejamento de políticas públicas voltadas à mitigação de atividades criminosas interconectadas.

Atividades centrais, como acidentes de trânsito e incidentes relacionados a drogas e a álcool, destacam-se por sua elevada conectividade nos grafos analisados, evidenciando uma relação intrínseca com contextos, nos quais crimes de menor ou média gravidade tendem a ocorrer em sequência. Em particular, a alta frequência de conexões, associadas a crimes relacionados a drogas e a álcool, sugere que tais eventos desempenham um papel de intermediação essencial em cadeias criminais, atuando como predecessores e sucessores de outros crimes, como agressão ou desordem pública. Esse comportamento coloca tais atividades no centro das dinâmicas criminais, funcionando como *hubs* no sistema de transição de eventos.

Em vista dessa premissa, a análise das múltiplas transições de entrada e saída, associadas a essas atividades, revela que elas desempenham um papel crucial na manutenção do fluxo dentro do sistema criminal. *Do ponto de vista do sistema criminoso*, ao distribuir conexões para uma ampla gama de categorias criminais, essas atividades promovem maior diversificação e flexibilidade nos fluxos de eventos, mitigando a formação de gargalos e reduzindo a concentração em pontos críticos. Essa característica de alta conectividade facilita não apenas a propagação dos fluxos entre diferentes tipos de crimes, mas também sustenta a dinâmica de eventos criminais complexos e interconectados. *Sob a perspectiva do sistema de controle e prevenção*, a compreensão dessas estruturas dinâmicas permite o desenvolvimento de estratégias mais eficazes de intervenção, prevenindo a escalada de atividades criminais e promovendo a alocação precisa de recursos.

Outrossim, os experimentos realizados neste estudo, descritos na *Tabela 18*, forneceram uma base quantitativa para avaliar a capacidade de predição de diferentes modelos. Conforme mencionado anteriormente, os conjuntos de dados utilizados foram segmentados em granularidades temporais distintas — mensal (D-0 a D-3), sazonal (D-4, baseado nas estações do ano) e categórica (dias úteis versus finais de semana, D-5) — com o objetivo de testar a resiliência e robustez dos modelos em diferentes tamanhos de amostras e escalas temporais.

As tabelas a seguir apresentam os resultados de desempenho dos modelos de *Cadeias de Markov* (MC), *Redes Neurais Recorrentes* (RNN), *Long Short-Term Memory* (LSTM) e *Gated Recurrent Unit* (GRU), analisados em diferentes granularidades temporais. O objetivo principal é investigar a eficácia desses modelos na predição de crimes e, considerando os conjuntos de dados utilizados, foram segmentados em granularidades temporais distintas — mensal (D-0 a D-3), sazonal (D-4, baseado nas estações do ano) e categórica (dias úteis versus finais de semana, D-5).

Para a análise, foram utilizadas métricas amplamente empregadas em tarefas de classificação: Recall, Precision e Acurácia, avaliadas em três categorias (Top-1, Top-2 e Top-3). Essas métricas oferecem uma visão abrangente sobre a capacidade dos modelos em capturar padrões relevantes nos dados e classificar corretamente as ocorrências previstas.

Tabela 18 - Descrição dos dados de *Recall* e *Precision* de GRU x LSTM.

	GRU							**LSTM**						
	Recall			*Precision*				*Recall*			*Precision*			
Conjunto de Dados	*Top - 1*	*Top - 2*	*Top -3*	*Top - 1*	*Top - 2*	*Top -3*	*Accuracy*	*Top - 1*	*Top - 2*	*Top -3*	*Top - 1*	*Top - 2*	*Top -3*	*Accuracy*
D-0 (100%, mês)	93.93 %	95.45 %	96.57 %	93.93 %	47.73 %	32.19 %	93.92 %	93.93 %	95.45 %	96.57 %	93.93 %	47.73 %	32.19 %	93.92 %
D-1 (25%,mês)	93.89 %	95.43 %	96.55 %	93.89 %	47.72 %	32.18 %	93.89 %	93.89 %	95.43 %	96.55 %	93.89 %	47.72 %	32.18 %	93.89 %
D-2 (50%,mês)	93.91 %	95.43 %	96.55 %	93.91 %	47.71 %	32.18 %	93.91 %	93.91 %	95.43 %	96.55 %	93.91 %	47.72 %	32.18 %	93.91 %
D-3 (75%,mês)	93.92 %	95.44 %	96.56 %	93.92 %	47.72 %	32.19 %	93.92 %	93.92 %	95.44 %	96.56 %	93.92 %	47.72 %	32.19 %	93.92 %
D-4 (100%, estações do ano)	83.31 %	87.57 %	90.68 %	83.31 %	43.79 %	30.23 %	83.30 %	83.31 %	87.58 %	90.68 %	83.31 %	43.79 %	30.23 %	83.30 %
D-5 (100%, Dias da semana e finais de semana)	81.55 %	85.96 %	88.67 %	81.55 %	42.98 %	29.56 %	81.55 %	81.56 %	85.96 %	88.68 %	81.56 %	42.98 %	29.56 %	81.56 %

Fonte: a Autora.

Na *Tabela 18,* a granularidade mensal (D-0 a D-3), ambos os modelos apresentaram desempenho consistente e elevado. O Recall variou de **93% a 96%** entre as categorias analisadas, enquanto a acurácia permaneceu estável em torno de **93%**. A precisão para **Top-1** foi superior a **93%**, mas apresentou uma tendência na precisão decrescente nas categorias

Top-2 e Top-3, alcançando valores mínimos de **32%**. Esses resultados indicam que os modelos foram capazes de identificar padrões predominantes de crimes com elevada confiança, apresentando consistência nos diferentes conjuntos de dados mensais avaliados.

Para granularidades mais amplas, como estações do ano (D-4) e dias da semana/fins de semana (D-5), o desempenho foi inferior em comparação com a granularidade mensal. No caso da granularidade sazonal (D-4), o Recall variou entre **83% no Top-1 e 90% no Top-3**, enquanto a precisão apresentou valores entre **83% no Top-1 e 30% no Top-3**. A acurácia observada foi de aproximadamente **83%**. Na granularidade semanal (D-5), o desempenho seguiu uma tendência semelhante, com Recall entre **81% e 88%**, precisão entre **81% e 29%**, e acurácia de **81%**.

Esses resultados mostram que, embora os modelos GRU e LSTM apresentem alta eficiência na predição de crimes em granularidades temporais mais detalhadas, como dados mensais, seu desempenho é reduzido ao lidar com granularidades mais amplas. Pois, a dificuldade em capturar padrões em contextos temporais sazonais ou semanais pode estar relacionada a características específicas desses períodos, que agregam informações menos frequentes e mais dispersas ao longo do tempo. Assim, a análise destaca a importância de considerar a granularidade temporal como um fator determinante na modelagem preditiva de crimes, evidenciando que granularidades mais detalhadas resultam em maior desempenho preditivo.

Tabela 19 - Descrição dos dados de *Recall* e *Precision* de GRU x RNN

	GRU							**RNN**						
	Recall			*Precision*				*Recall*			*Precision*			
Conjunto de Dados	*Top - 1*	*Top - 2*	*Top -3*	*Top - 1*	*Top - 2*	*Top -3*	*Accuracy*	*Top - 1*	*Top - 2*	*Top -3*	*Top - 1*	*Top - 2*	*Top -3*	*Accuracy*
D-0 (100%, mês)	93.93 %	95.45 %	96.57 %	93.93 %	47.73 %	32.19 %	93.92 %	93.93 %	95.45 %	96.57 %	93.93 %	47.73 %	32.19 %	93.92 %
D-1 (25%,mês)	93.89 %	95.43 %	96.55 %	93.89 %	47.72 %	32.18 %	93.89 %	93.89 %	95.43 %	96.55 %	93.89 %	47.72 %	32.18 %	93.89 %
D-2 (50%,mês)	93.91 %	95.43 %	96.55 %	93.91 %	47.71 %	32.18 %	93.91 %	93.91 %	95.43 %	96.55 %	93.91 %	47.72 %	32.18 %	93.91 %
D-3 (75%,mês)	93.92 %	95.44 %	96.56 %	93.92 %	47.72 %	32.19 %	93.92 %	93.92 %	95.44 %	96.56 %	93.92 %	47.72 %	32.19 %	93.92 %
D-4 (100%, estações do	83.31 %	87.57 %	90.68 %	83.31 %	43.79 %	30.23 %	83.30 %	83.31 %	87.57 %	90.68 %	83.31 %	43.79 %	30.23 %	83.30 %

ano)														
D-5 (100%, Dias da semana e finais de semana)	81.55 %	85.96 %	88.67 %	81.55 %	42.98 %	29.56 %	81.55 %	81.55 %	85.96 %	88.63 %	81.55 %	42.98 %	29.54 %	81.54 %

.Fonte: a Autora.

Para a *Tabela 19,* as granularidades mensais (D-0 a D-3), ambos os modelos apresentaram desempenho elevado e consistente. O Recall variou entre **93% e 96%** nas categorias analisadas, enquanto a precisão no **Top-1** se manteve acima de **93%**, diminuindo gradativamente nas categorias **Top-2 e Top-3,** onde atingiu valores mínimos de **32%**. A acurácia permaneceu constante, com valores próximos de **93%** para ambos os modelos. Esses resultados indicam que os modelos demonstraram elevada capacidade de identificação dos padrões predominantes de crimes em granularidades temporais mensais, refletindo uma capacidade robusta de generalização para crimes recorrentes.

Nas granularidades sazonais (D-4) e semanais (D-5), observou-se uma redução no desempenho de ambos os modelos. Na granularidade sazonal (D-4), o Recall variou entre **83% no Top-1 e 90% no Top-3**, enquanto a precisão apresentou valores entre **83% no Top-1 e 30% no Top-3**. A acurácia observada foi de aproximadamente **83%**. Na granularidade semanal (D-5), o desempenho seguiu uma tendência semelhante, com o Recall variando entre **81% e 88%**, precisão entre **81% e 29%**, e acurácia de **81%**.

Esses resultados evidenciam que, enquanto os modelos GRU e RNN alcançam alta eficiência em granularidades temporais mais detalhadas, como dados mensais, seu desempenho diminui à medida que a granularidade temporal se torna mais ampla. Essa redução pode ser atribuída à maior variabilidade temporal e à complexidade na identificação de padrões menos frequentes ou específicos associados a períodos sazonais ou semanais. Assim, a análise reforça a importância de considerar a granularidade temporal como um fator crítico no desenvolvimento de modelos preditivos, destacando que granularidades mais detalhadas favorecem um desempenho superior na predição de crimes.

Tabela 20 - Descrição dos dados de *Recall* e *Precision* de GRU x MC.

	GRU							Markov Chain						
	Recall			*Precision*				*Recall*			*Precision*			
Conjunto de Dados	*Top - 1*	*Top - 2*	*Top -3*	*Top - 1*	*Top - 2*	*Top -3*	*Accuracy*	*Top - 1*	*Top - 2*	*Top -3*	*Top - 1*	*Top - 2*	*Top -3*	*Accuracy*
D-0 (100%, mês)	93.93 %	95.45 %	96.57 %	93.93 %	47.73 %	32.19 %	93.92 %	26%	43%	79%	26%	21%	26%	26%
D-1 (25%,mês)	93.89 %	95.43 %	96.55 %	93.89 %	47.72 %	32.18 %	93.89 %	26%	43%	60%	26%	21%	20%	26%
D-2 (50%,mês)	93.91 %	95.43 %	96.55 %	93.91 %	47.71 %	32.18 %	93.91 %	26%	43%	60%	26%	21%	20%	26%
D-3 (75%,mês)	93.92 %	95.44 %	96.56 %	93.92 %	47.72 %	32.19 %	93.92 %	26%	43%	54%	26%	21%	18%	26%
D-4 (100%, estações do ano)	83.31 %	87.57 %	90.68 %	83.31 %	43.79 %	30.23 %	83.30 %	26%	43%	69%	26%	21%	23%	26%
D-5 (100%, Dias da semana e finais de semana)	81.55 %	85.96 %	88.67 %	81.55 %	42.98 %	29.56 %	81.55 %	26%	43%	69%	26%	21%	23%	26%

Fonte: a Autora.

A análise da *Tabela 20* apresenta as granularidades mensais (D-0 a D-3) e observou-se que o GRU obteve desempenho consistentemente elevado. O *Recall* variou entre **93,89% e 96,57%**, e a *Precisão* apresentou valores entre **93,93% no Top-1** e **32,19% no Top-3**. Já a acurácia permaneceu estável, em torno de **93,89% a 93,92%**. Esses resultados indicam que o GRU é capaz de capturar padrões temporais e sequenciais com alta precisão, fornecendo previsões robustas em dados mensais. Em contrapartida, as Cadeias de Markov apresentaram resultados significativamente inferiores, com *Recall* variando entre **26% a 79%** e *Precisão* entre **26% no Top-1** e **43% Top-2 a 79% no Top-3**. A acurácia para a Cadeia de Markov esteve em constância em **26%**, evidenciando suas limitações na modelagem de padrões temporais mais complexos.

Na granularidade sazonal (D-4), correspondente às estações do ano, o GRU manteve um desempenho robusto, com *Recall* entre **83,31% no Top-1** e **90,68% no Top-3**. A *Precisão* apresentou valores de **43,79% no Top-1** e **30,23% no Top-3**, com acurácia em torno de **83,30%**. Esses resultados demonstram que o GRU consegue capturar variações

sazonais nos dados, mesmo com a maior dispersão temporal característica desse tipo de granularidade. Por outra perspectiva, as Cadeias de Markov apresentaram desempenho inferior, com *Recall* variando entre **26% e 69%** e *Precisão* entre **26% no Top-1** e **69% no Top-3**. A acurácia das Cadeias de Markov permaneceu em torno de **26%**, indicando limitações em capturar padrões temporais sazonais mais complexos.

Na granularidade semanal (D-5), que considera a diferenciação entre dias da semana e finais de semana, o GRU manteve resultados superiores. O *Recall* variou entre **81,55% no Top-1** e **88,67% no Top-3**, enquanto a *Precisão* apresentou valores entre **42,98% no Top-1** e **29,56% no Top-3**. A acurácia manteve-se estável em **81,55%**. Em contraste, as Cadeias de Markov apresentaram *Recall* entre **26% e 69%** e *Precisão* entre **26% no Top-1** e **79% no Top-3**, com acurácia em torno de **26%**. Esses resultados sugerem que o GRU é capaz de capturar padrões cíclicos de curta duração, como os presentes em granularidades semanais, com desempenho significativamente superior.

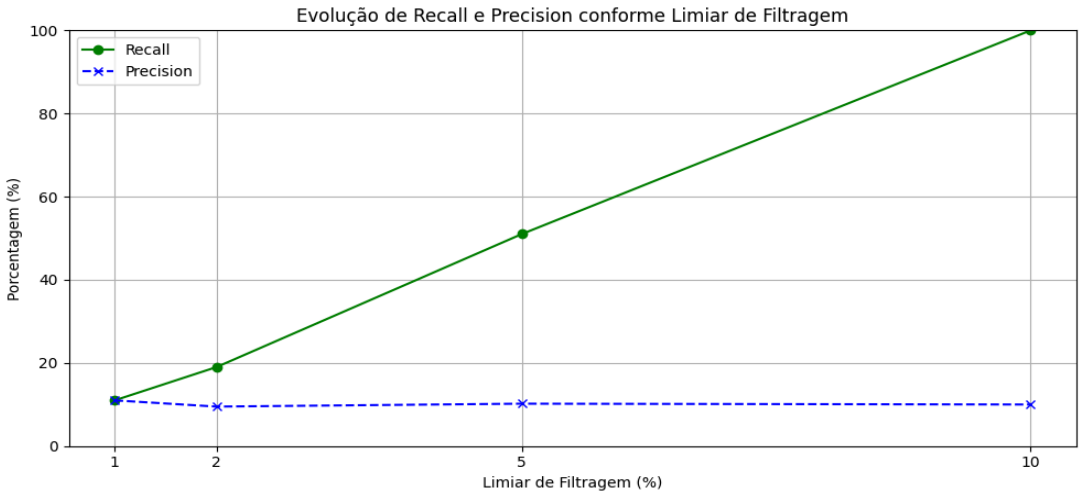
O desempenho consistente do modelo GRU, ao longo das diferentes granularidades temporais, demonstra sua superioridade em lidar com dados sequenciais e com eventos temporais complexos. Sua capacidade de generalização, mesmo em contextos de maior dispersão temporal, reflete sua adequação para aplicações relacionadas à predição de crimes. Já as Cadeias de Markov, embora apresentem resultados mais modestos, podem ser úteis em cenários de menor complexidade ou como base para análises comparativas.

O gráfico apresenta a evolução dos valores de *recall* e *precision* em função do limiar de filtragem, evidenciando o impacto desse parâmetro nos desempenhos das métricas preditivas. No eixo horizontal, temos o limiar de filtragem (%) variando de 1% a 10%, enquanto o eixo vertical mostra os valores percentuais de *recall* e *precision*.

O gráfico da *Figura 20* ilustra a evolução das métricas de *recall* e precisão em função do limiar de filtragem, destacando a influência direta desse parâmetro no desempenho preditivo. No eixo horizontal, é representada a variação do limiar de filtragem, expressa em percentuais de 1% a 10%. Já no eixo vertical, encontram-se os valores percentuais de *recall* e precisão, permitindo observar como diferentes níveis de filtragem afetam a capacidade dos modelos em identificar corretamente os eventos relevantes (*recall*) e a proporção de predições corretas em relação às realizadas (precisão). A análise conjunta das curvas oferece

uma visão detalhada do *trade-off* entre essas métricas à medida que o limiar de filtragem é ajustado.

Figura 20 - Evolução do *recall* e *precision* conforme o limiar de filtragem.



Fonte: a Autora.

A análise dos resultados revela que o recall aumenta linearmente com o incremento do limiar de filtragem. Em um limiar de 1%, o recall é baixo, indicando que apenas uma pequena proporção dos eventos relevantes é capturada. No entanto, ao atingir o limiar de 10%, o recall alcança 100%, demonstrando que, nesse ponto, todos os eventos esperados são identificados. Esse comportamento é característico de cenários em que a sensibilidade do modelo é ajustada para captar uma maior proporção de eventos relevantes, geralmente ao custo de um aumento na ocorrência de falsos positivos.

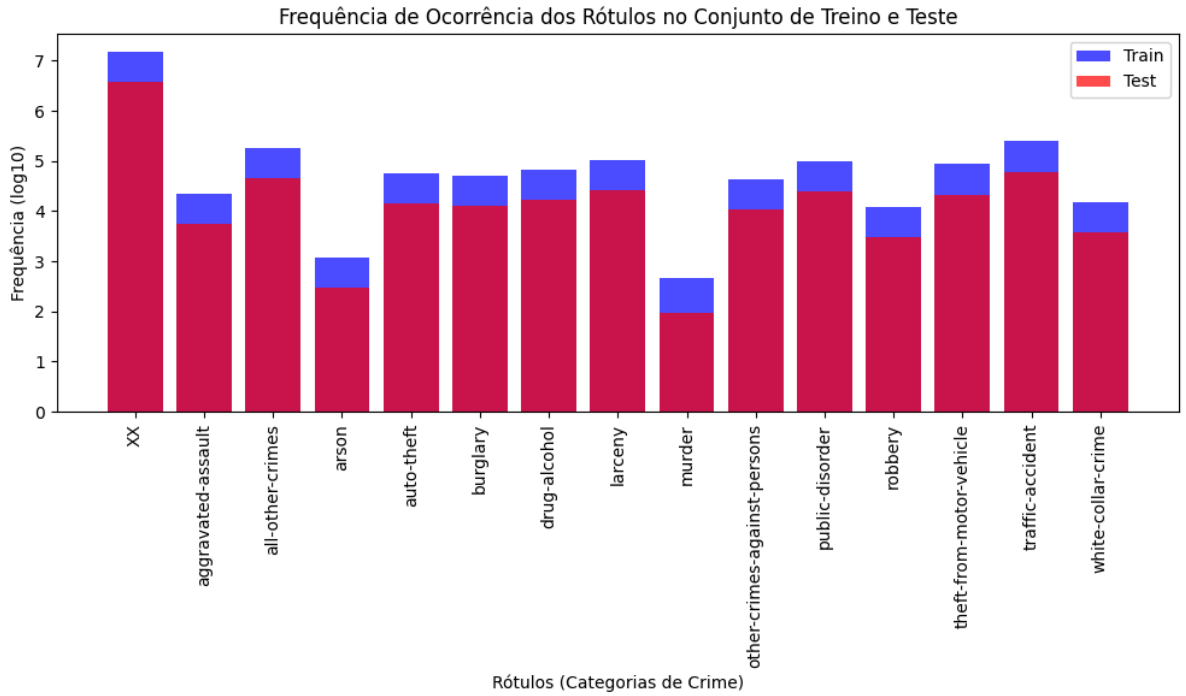
Por outro aspecto, a precisão mantém-se constante em torno de um valor relativamente baixo (~10%) em todos os limites analisados. Isso indica que, mesmo com o aumento do *recall*, a proporção de previsões relevantes entre as classificações positivas permanece limitada. Esse comportamento sugere que o modelo apresenta uma alta taxa de falsos positivos, evidenciando que o crescimento no *recall* não se traduz em uma melhora proporcional na precisão.

A interação distinta entre *recall* e precisão evidencia uma característica fundamental do modelo: enquanto a capacidade de identificar eventos aumenta com o relaxamento do

limiar de filtragem, a habilidade do modelo em discriminar eventos relevantes permanece restrita. Esse desbalanceamento é comum em problemas envolvendo dados desbalanceados, como em cenários de predição de eventos raros. Tal situação destaca a importância de uma análise cuidadosa da distribuição dos dados, fundamental para entender as limitações do modelo.

Complementando essa análise, o gráfico da *Figura 21* apresentado a seguir ilustra a distribuição das frequências de ocorrência dos rótulos nos conjuntos de treino e teste em escala logarítmica. Essa visualização revela o impacto direto do desbalanceamento dos dados no desempenho preditivo, reforçando a necessidade de estratégias que não apenas otimizem o equilíbrio entre *recall* e precisão, mas que também considerem a estrutura subjacente dos dados nos quatro conjuntos acarretados (*train_X, train_y, test_X e test_y*).

Figura 21 - Frequência de treinamento e frequência de testes.



Fonte: a Autora.

A *Figura 21* apresenta um gráfico que detalha a distribuição de frequência dos rótulos criminais nos conjuntos de treino e teste, utilizando uma escala logarítmica para capturar adequadamente a ampla variação nas ocorrências. Essa abordagem permite uma representação mais clara tanto dos crimes mais frequentes quanto daqueles com menor

incidência, evidenciando a predominância de determinadas categorias criminais. Além disso, a visualização expõe, de forma explícita, o desbalanceamento de classes presente nos dados, um aspecto crucial que impõe desafios significativos ao desenvolvimento de modelos preditivos robustos, especialmente em contextos urbanos complexos, onde a diversidade e a disparidade dos fenômenos criminais são amplificadas.

Conforme observado anteriormente, entre as categorias analisadas, destaca-se uma prevalência significativa de rótulos genéricos e de alta frequência, como "XX" e "*traffic accident*", que aparecem de forma consistente tanto no conjunto de treino quanto no de teste. Em particular, o rótulo "XX" registra a maior frequência, evidenciando a ocorrência de dados incompletos ou não especificados. Essa observação ressalta a necessidade de estratégias robustas para o tratamento de valores ausentes durante o processo de modelagem, uma etapa crucial para garantir a qualidade e a confiabilidade das predições.

Além disso, categorias como "*aggravated assault*" e "*public disorder*" também exibem alta representatividade, refletindo padrões recorrentes de criminalidade característicos de áreas urbanas. Esses padrões não apenas fornecem uma visão sobre a dinâmica criminal nesses contextos, mas também indicam sua relevância para análises preditivas e a identificação de tendências, auxiliando na formulação de estratégias de segurança pública mais direcionadas e eficazes.

Além disso, a predominância de categorias de alta frequência no conjunto de dados é essencial para identificar tendências gerais e padrões consistentes, mas também apresenta um desafio significativo: o risco de enviesamento dos modelos preditivos, que podem priorizar crimes comuns em detrimento de categorias menos frequentes. Esse cenário destaca a necessidade de abordar o desbalanceamento de classes de forma estratégica, utilizando técnicas como a reamostragem, ajuste de pesos ou algoritmos especializados em tratar disparidades de distribuição. Desse modo, o aprimoramento desses modelos exige que eles sejam capazes de generalizar adequadamente, ao mesmo tempo que capturam padrões relevantes em crimes menos recorrentes.

Por outra perspectiva, categorias de baixa frequência, como "*murder*" e "*arson*", apresentam uma representatividade reduzida, o que pode impactar negativamente o desempenho preditivo, especialmente em casos de crimes raros, mas de alta relevância social.

A escala logarítmica, utilizada para análise, reforça a percepção do desbalanceamento de classes e sua potencial influência nos resultados. No entanto, a consistência observada entre os conjuntos de treino e teste é encorajadora, pois garante uma separação adequada dos dados, permitindo avaliações mais equilibradas do desempenho dos modelos.

Apesar disso, a alta frequência de categorias dominantes em ambos os conjuntos reforça a necessidade de estratégias que evitem negligenciar crimes menos frequentes. Soluções como o ajuste de pesos para categorias sub-representadas e o uso de métricas mais abrangentes, como a *F1-Score* — que considera tanto precisão quanto *recall* —, podem ser relevantes para um treinamento mais equilibrado.

Adicionalmente, a visualização das distribuições de frequência em escalas logarítmicas oferece uma visão das disparidades entre os tipos de crimes, evidenciando a importância de incluir tanto crimes frequentes quanto raros no processo de modelagem preditiva. Esses resultados enfatizam a necessidade de abordagens adaptativas e balanceadas, que permitam lidar com os desafios inerentes aos dados criminais urbanos. Dessa forma, os modelos podem suscitar previsões mais precisas e, consequentemente, apoiar de maneira eficaz a formulação de políticas públicas e estratégias de segurança.

Os resultados completos desta pesquisa podem ser encontrados no **Anexo A**, no qual estão organizados de acordo com as pastas e subpastas, permitindo fácil acesso aos dados, gráficos e imagens gerados durante o estudo.

6. DISCUSSÃO

Os resultados deste estudo destacaram a eficácia dos modelos preditivos aplicados aos dados criminais da cidade de Denver, revelando diferenças significativas entre as abordagens em termos de capacidade preditiva e eficiência computacional. As Redes Neurais Recorrentes, especialmente as arquiteturas LSTM e GRU, mostraram desempenho superior em comparação às Cadeias de *Markov*, evidenciando sua habilidade em capturar padrões temporais complexos e dependências de longo prazo, principalmente em granularidades mais amplas, como estações do ano (D-4) e dias úteis/fins de semana (D-5).

Entre os modelos avaliados, o GRU destacou-se como a abordagem mais equilibrada, combinando precisão preditiva com eficiência computacional. Em cenários como a granularidade mensal (D-0 a D-3), o GRU manteve métricas elevadas de *recall* (acima de 95% no Top-3) e acurácia (93%), enquanto em configurações sazonais (D-4) e semanais (D-5) sua performance foi consistente, com *recall* superior a 85% e tempos de execução reduzidos em relação ao LSTM. Essa eficiência pode ser atribuída à sua arquitetura simplificada, que exige menos parâmetros e reduz significativamente o tempo de processamento sem comprometer o desempenho.

As Cadeias de *Markov*, por sua vez, apresentaram limitações em granularidades reduzidas e configurações mais desafiadoras, com *recall* e precisão inferiores às obtidas pelos modelos baseados em redes neurais. Embora sejam simples e interpretáveis, essas abordagens mostraram incapacidade de capturar padrões temporais complexos, alcançando precisão no Top-3 de apenas 79% na granularidade mensal (D-0).

A análise revelou que granularidades temporais densas, como D-0, são mais adequadas para o desempenho preditivo consistente dos modelos, permitindo a identificação de padrões recorrentes. Já em granularidades mais amplas, como D-4 e D-5, os desafios de predição são maiores, exigindo ajustes nos modelos para capturar padrões sazonais e dispersos.

Os achados sugerem que o GRU é a escolha mais eficiente e robusta para a tarefa de predição de crimes, equilibrando sensibilidade e especificidade mesmo em cenários desafiadores.

Os resultados indicam caminhos promissores para estudos futuros, incluindo a aplicação do método em outras regiões e a inclusão de variáveis contextuais, como dados socioeconômicos e climáticos, para aprimorar a capacidade preditiva e oferecer informações mais profundas sobre as dinâmicas criminais. Tais avanços podem contribuir para estratégias mais eficazes de prevenção e alocação de recursos na segurança pública.

Benefícios para a Segurança Pública

Os resultados desta dissertação têm implicações diretas para a segurança pública, oferecendo ferramentas baseadas em aprendizado de máquina que podem otimizar estratégias de prevenção e alocação de recursos. A capacidade de prever a ocorrência de crimes com maior precisão permite que forças policiais concentrem esforços em áreas ou períodos de maior risco, reduzindo a incidência de crimes e melhorando a eficiência operacional. Modelos preditivos como GRU e LSTM podem ser integrados a sistemas de segurança para fornecer informações em tempo real sobre tendências criminais, possibilitando ações proativas que minimizem os impactos sociais e econômicos do crime.

Além disso, as redes neurais recorrentes mostraram potencial para identificar padrões temporais complexos e dependências contextuais, fornecendo insights valiosos para o planejamento estratégico. Por exemplo, a identificação de padrões sazonais ou específicos de determinados bairros pode ajudar a polícia a alocar patrulhas de maneira mais eficiente, enquanto dados preditivos sobre categorias específicas de crimes podem informar campanhas de conscientização pública e políticas preventivas.

A integração dessas ferramentas nos sistemas de segurança pública também pode promover uma abordagem mais justa e informada, utilizando dados para melhorar a transparência na tomada de decisões. A implementação desses métodos no contexto brasileiro, por exemplo, poderia contribuir para enfrentar desafios específicos, como a separação de dados para policiamento preventivo e investigativo, alinhando esforços das polícias civil e militar. Esse tipo de abordagem integrada fortalece a segurança pública ao possibilitar o uso de análises baseadas em evidências para a formulação de políticas mais eficazes e equitativas.

Reflexões sobre os Resultados Obtidos

Os resultados deste estudo destacam reflexões críticas ao aplicar aprendizado de máquina a dados reais e complexos, como aqueles relacionados à criminalidade urbana. Um dos desafios mais evidentes é o impacto do desbalanceamento de classes, especialmente em categorias menos frequentes, como "*murder*" e "*arson*". Apesar do desempenho reduzido nessas classes, esse comportamento é inerente às características intrínsecas de dados desbalanceados e reflete os desafios reais enfrentados no tratamento de dados do mundo real. Ao contrário de *datasets* artificialmente balanceados, os dados reais carregam nuances, inconsistências e desproporções que tornam a tarefa preditiva mais desafiadora, mas também oferecem oportunidades únicas de aprendizado e inovação.

Assim, os resultados enfatizam que trabalhar com dados reais não apenas expõe limitações, mas também oferece um terreno rico para inovações. As soluções desenvolvidas neste contexto precisam ser robustas, práticas e adaptáveis às condições do mundo real. Isso não apenas fortalece a relevância do aprendizado de máquina em aplicações críticas, como a predição de crimes, mas também aponta para a importância de abordar os desafios de dados desbalanceados com estratégias inovadoras que atendam às complexidades e às demandas práticas. Essa abordagem tem o potencial de contribuir para políticas públicas mais informadas e para estratégias de segurança mais eficazes, impactando diretamente a formulação de soluções práticas e escaláveis para a gestão da criminalidade.

Por fim, a escolha pelo uso de redes neurais como GRU, LSTM e RNN neste estudo fundamenta-se em sua habilidade comprovada para modelar eventos temporais e prever padrões em sequências complexas. Os resultados obtidos demonstram o potencial dessas abordagens para aplicações práticas, especialmente em tarefas que envolvem a análise de padrões temporais de alta complexidade, como a predição de crimes em diferentes escalas temporais.

7. CONCLUSÃO E TRABALHOS FUTUROS

Esta dissertação propôs um método híbrido para a predição de crimes urbanos, integrando mineração de processo, cadeia de *Markov* e redes neurais recorrentes (RNNs), incluindo suas variantes *Long Short-Term Memory* (LSTM) e *Gated Recurrent Unit* (GRU). A aplicação desse método ao contexto da cidade de Denver mostrou sua eficácia ao identificar padrões temporais e sequenciais de ocorrências criminais, oferecendo uma base promissora para análises preditivas e suporte estratégico às forças de segurança pública.

Os resultados obtidos destacam contribuições para a segurança pública, ao fornecer uma abordagem prática e baseada em dados para a alocação de recursos e a formulação de estratégias preventivas. O método desenvolvido permite que órgãos responsáveis pela segurança pública visualizem padrões criminais com maior clareza, identifiquem *hotspots* de criminalidade e antecipem comportamentos delituosos, o que pode levar à redução da incidência de crimes e à otimização de operações policiais. Além disso, a capacidade de modelar transições entre diferentes tipos de crimes com precisão e de prever ocorrências futuras possibilita intervenções mais direcionadas e eficientes.

A mineração de processo possibilitou a construção de mapas, na forma de DFGs, os quais foram fundamentais para modelar as relações dinâmicas entre diferentes tipos de crimes. Por meio de operações de simplificação, como agrupamento e filtragem, esses grafos foram otimizados para garantir maior eficiência computacional e interpretabilidade, facilitando a análise de padrões complexos. Complementarmente, as Cadeias de Markov forneceram uma estrutura probabilística para modelar as transições entre crimes, permitindo a geração de estimativas sobre eventos futuros.

Além dessas técnicas, as RNNs e suas variantes (LSTM e GRU) foram aplicadas para lidar com a complexidade e com as dependências temporais nos dados criminais. Estas redes mostraram elevado desempenho na predição de crimes frequentes, com métricas de avaliação como precisão e *recall,* refletindo a capacidade de prever corretamente os tipos de crimes em diferentes granularidades temporais. Contudo, as análises também indicaram que a predição de tipos de crimes menos frequentes permanece um desafio, apontando para a necessidade de refinamento do modelo e tratamento do desbalanceamento de classes.

Ao considerar a aplicabilidade prática, os benefícios para a segurança pública se estendem à formulação de políticas públicas baseadas em evidências. A capacidade de prever crimes com maior precisão pode subsidiar a criação de programas de prevenção mais eficazes, além de fortalecer a integração entre diferentes esferas de segurança, como as polícias militar e civil, promovendo uma atuação mais coordenada e eficiente.

Em síntese, esta pesquisa contribui para o avanço da ciência aplicada à predição criminal, ao propor um método eficaz que combina técnicas de mineração, modelagem estatística e aprendizado profundo. O trabalho estabelece uma base metodológica relevante, com implicações práticas e científicas, oferecendo caminhos promissores para futuros estudos e aplicações na formulação de políticas públicas orientadas por dados.

Trabalhos Futuros

Para ampliar o impacto e a aplicabilidade do método proposto, são sugeridas as seguintes direções para trabalhos futuros:

1. *Otimização do código existente:* Expandir as análises para incluir janelas temporais mais específicas, como ciclos anuais (52 semanas) e mensais (4 semanas), otimizando a granularidade da previsão.
2. *Refinamento do modelo preditivo:* Melhorar o equilíbrio entre precisão e *recall* para crimes frequentes e raros, além de incorporar redes neurais ao modelo híbrido de mineração de processo e cadeia de *Markov*, visando à maior robustez e precisão.
3. *Exploração de técnicas de balanceamento de classes:* Investigar e implementar estratégias para lidar com o problema de desbalanceamento nas classes de crimes, utilizando técnicas como subamostragem e superamostragem.
4. *Aplicação em diferentes contextos:* Aplicar o método em outras cidades para comparar padrões criminais e avaliar sua aplicabilidade em diferentes cenários. Além disso, explorar a aplicabilidade multissetorial das técnicas, como: previsão de surtos de doenças (*Saúde*); otimização de rotas de transporte (*Logística*); identificação de padrões de evasão escolar (*Educação*); e antecipação de congestionamentos (*Transportes*); e detecção de fraudes (*Finanças*).

Essas propostas reforçam o compromisso em refinar e em expandir a aplicabilidade do método, assegurando sua relevância tanto no campo científico quanto nas aplicações práticas.

References

Aalst, W. van der. (2011). *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer.

Aalst, W. M. P. van der. (2016). *Process Mining: Data Science in Action*. Springer.

Aziz, R. M., Sharma, P., & Hussain, A. (2022). Machine learning algorithms for crime prediction under Indian Penal Code. *Annals of Data Science*. https://doi.org/10.1007/s40745-022-00424-6

Blumstein, A., Cohen, J., & Farrington, D. P. (1988). CRIMINAL CAREER RESEARCH: ITS VALUE FOR CRIMINOLOGY*. *Criminology*, *26*(1), 1–35. https://doi.org/10.1111/j.1745-9125.1988.tb00829.x

Budgen, D., & Brereton, P. (2006). Performing systematic literature reviews in software engineering. *Proceedings of the 28th International Conference on Software Engineering*. New York, NY, USA: ACM. Retrieved from http://dx.doi.org/10.1145/1134285.1134500

Dargan, S., Kumar, M., Garg, A., & Thakur, K. (2019). Writer identification system for pre-segmented offline handwritten Devanagari characters using k-NN and SVM. *Soft Computing*, *24*(13), 10111–10122. https://doi.org/10.1007/s00500-019-04525-y

Ding, N., & Zhai, Y. (2019). Crime prevention of bus pickpocketing in Beijing, China: Does air quality affect crime? *Security Journal*, *34*(2), 262–277. https://doi.org/10.1057/s41284-019-00226-1

Donato, H., & Donato, M. (2019). Etapas na Condução de uma Revisão Sistemática. *Acta Médica Portuguesa*, *32*(3), 227–235. https://doi.org/10.20344/amp.11923

Douc, R., Moulines, E., Priouret, P., & Soulier, P. (2018). *Markov Chains*. Springer.

Dupuis, A., Dadouchi, C., & Agard, B. (2022). Predicting crop rotations using process

mining techniques and Markov principals. *Computers and Electronics in Agriculture*, *194*, 106686. https://doi.org/10.1016/j.compag.2022.106686

Heidensohn, F. (1989). *Crime and society*. london: Macmillan education ltd.

Jonathan, O. E., Olusola, A. J., Bernadin, T. C. A., & Inoussa, T. M. (2021). Impacts of crime on socio-economic development. *Mediterranean Journal of Social Sciences*, *12*(5), 71. https://doi.org/10.36941/mjss-2021-0045

Kemeny, J. G., & Snell, J. L. (1960). *Finite Markov Chains*.

Kumar, P. (n.d.). Denver crime data. Retrieved November 13, 2023, from Kaggle website: https://www.kaggle.com/datasets/penchalaiah123/denver-crime-data?select=offense_codes.csv

Lazar, J., Feng, J. H., & Hochheiser, H. (2010). *Research methods in human-computer interaction*. John Wiley & Sons.

Morshed, A., Forkan, A. R. M., Tsai, P.-W., Jayaraman, P. P., Sellis, T., Georgakopoulos, D., … Ranjan, R. (2019). VisCrimePredict. *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*. New York, NY, USA: ACM. Retrieved from http://dx.doi.org/10.1145/3297280.3297388

Orong, M. Y., Sison, A. M., & Hernandez, A. A. (2018). Mitigating vulnerabilities through forecasting and crime trend analysis. *2018 5th International Conference on Business and Industrial Research (ICBIR)*. IEEE. Retrieved from http://dx.doi.org/10.1109/icbir.2018.8391166

Paul Walker. (2009). Paul Walker. Retrieved February 21, 2023, from Pensador website: https://www.pensador.com/frase/MTQ1MjA2Mw/

Pradhan, I., Potika, K., Eirinaki, M., & Potikas, P. (2019). Exploratory data analysis and crime prediction for smart cities. *Proceedings of the 23rd International Database*

Applications & Engineering Symposium on - IDEAS '19. New York, New York, USA: ACM Press. Retrieved from http://dx.doi.org/10.1145/3331076.3331114

Process mining and automated process discovery software for professionals. (n.d.). Retrieved November 6, 2023, from Fluxicon Disco. website: https://fluxicon.com/disco/

Rumi, S. K., Deng, K., & Salim, F. D. (2018). Crime event prediction with dynamic features. *EPJ Data Science*, *7*(1). https://doi.org/10.1140/epjds/s13688-018-0171-7

Rummens, A., Hardyns, W., & Pauwels, L. (2017). The use of predictive analysis in spatiotemporal crime forecasting: Building and testing a model in an urban context. *Applied Geography*, *86*, 255–261. https://doi.org/10.1016/j.apgeog.2017.06.011

Sullivan, C. J., & Piquero, A. R. (2016). The Criminal Career Concept. *Journal of Research in Crime and Delinquency*, *53*(3), 420–442. https://doi.org/10.1177/0022427815627313

Thomas, A., & Sobhana, N. V. (2022). A survey on crime analysis and prediction. *Materials Today: Proceedings*, *58*, 310–315. https://doi.org/10.1016/j.matpr.2022.02.170

Tomé, T. (2001). *Dinâmica Estocástica e Irreversibilidade Vol. 35*. EdUSP.

Wu, J., Abrar, S. M., Awasthi, N., Frias-Martinez, E., & Frias-Martinez, V. (2022). Enhancing short-term crime prediction with human mobility flows and deep learning architectures. *EPJ Data Science*, *11*(1). https://doi.org/10.1140/epjds/s13688-022-00366-2

Yoo, Y., & Wheeler, A. P. (2019). *Using risk terrain modeling to predict homeless related crime in los angeles, california*. Center for Open Science. Retrieved from Center for Open Science website: http://dx.doi.org/10.31235/osf.io/swfpn

Zhang, Y., Siriaraya, P., Kawai, Y., & Jatowt, A. (2020). Predicting time and location of future crimes with recommendation methods. *Knowledge-Based Systems*, *210*,

106503. https://doi.org/10.1016/j.knosys.2020.106503

Zhao, X., & Tang, J. (2018). Crime in urban areas: *ACM SIGKDD Explorations Newsletter*, *20*(1), 1–12. https://doi.org/10.1145/3229329.3229331

ANEXOS

Anexo A - Descrição dos Resultados por Pasta

Os resultados desta pesquisa estão organizados em diferentes pastas, acessíveis por intermédio do seguinte link:

https://mega.nz/folder/40AQyDJZ#r-6ZFU9OvCXUVWv1Ctp6MQ.

- **Pasta "Denver_crimes_100%"**: Contém os dados completos, com arquivos em formato .csv armazenados na pasta **data**. Na pasta **image**, encontram-se as imagens geradas durante a análise, enquanto os gráficos salvos estão na pasta **graph**.
- **Pasta "period"**: Abriga as subpastas **seasons** e **work_weekend**, cada uma delas estruturada da mesma forma que a pasta "Denver_crimes_100%", com subpastas **data**, **image** e **graph**, contendo os resultados específicos para cada período analisado.
- **Pasta "precints"**: Contém as subpastas **precint_25**, **precint_50** e **precint_75**, que correspondem aos resultados obtidos para 25%, 50% e 75% da distribuição de dados. Cada uma dessas pastas também segue a mesma organização, com subpastas **data**, **image** e **graph**, permitindo o acesso aos dados, imagens e gráficos gerados para essas proporções.