UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

JEAN GOMES TURET

METODOLOGIA HÍBRIDA DE ANÁLISE DE DADOS ESTRUTURADOS E DADOS NÃO ESTRUTURADOS PARA APOIO AO PROCESSO DECISÓRIO EM SEGURANÇA PÚBLICA.

Recife
2023

JEAN GOMES TURET

METODOLOGIA HÍBRIDA DE ANÁLISE DE DADOS ESTRUTURADOS E DADOS NÃO ESTRUTURADOS PARA APOIO AO PROCESSO DECISÓRIO EM SEGURANÇA PÚBLICA.

Tese apresentada ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Engenharia de Produção.

Área de concentração: Gerência da Produção.

Orientadora Prof. Drª. Ana Paula Cabral Seixas Costa

Recife

2023

Catalogação na fonte:
Bibliotecário Carlos Moura, CRB-4/1502

T935m Turet, Jean Gomes.
Metodologia híbrida de análise de dados estruturados e dados não estruturados para apoio ao processo decisório em segurança pública. / Jean Gomes Turet. – 2023.
92 f.: il.

Orientadora: Profª. Drª. Ana Paula Cabral Seixas Costa.
Tese (doutorado) – Universidade Federal de Pernambuco. CTG. Programa de Pós-Graduação em Engenharia de Produção, 2023.
Inclui referências.

1. Engenharia de produção. 2. Aprendizagem de máquina. 3. Dados estruturados. 4. Dados não estruturados. 5. Segurança pública. 6. Big data. I. Costa, Ana Paula Cabral Seixas (orientadora). II. Título.

658.5 (22. ed.)

UFPE
BCTG/2023-85

JEAN GOMES TURET

METODOLOGIA HÍBRIDA DE ANÁLISE DE DADOS ESTRUTURADOS E DADOS NÃO ESTRUTURADOS PARA APOIO AO PROCESSO DECISÓRIO EM SEGURANÇA PÚBLICA

Tese apresentada ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Pernambuco, Centro de Tecnologia e Geociências, como requisito parcial para a obtenção do título de Doutor em Engenharia de Produção. Área de concentração: Gerência da Produção.

Aprovado em: 01 / 03 / 2023.

BANCA EXAMINADORA

_____
Profª. Drª. Ana Paula Cabral Seixas Costa (Orientadora)
Universidade Federal de Pernambuco

_____
Profª. Drª. Caroline Maria de Miranda Mota (Examinadora Interna)
Universidade Federal de Pernambuco

_____
Profª. Drª. Danielle Costa Morais (Examinadora Interna)
Universidade Federal de Pernambuco

_____
Prof. Dr. Ricardo Augusto Cassel (Examinador Externo)
Universidade Federal do Rio Grande do Sul

_____
Prof. Dr. Guido Lemos de Souza Filho (Examinador Externo)
Universidade Federal da Paraíba

AGRADECIMENTOS

À minha mãe, Eliene, pelo amor, carinho e dedicação a mim ao longo da vida. Agradeço aos meus familiares que sempre estiveram do meu lado em todas as circunstâncias.

Agradeço aos meus amigos, que adquiri ao longo da vida, em especial ao meu grupo de pesquisa GPSID-UFPE pelo suporte à minha pesquisa.

Meus sinceros agradecimentos à professora e minha orientadora, Ana Paula Cabral Seixas Costa pelo suporte dado durante meu doutorado, quando não mediu esforços para mostrar caminhos e oportunidades. Uma profissional que inspira diariamente a me tornar uma pessoa e um profissional melhor.

Ao Programa de Pós-graduação em Engenharia de Produção (PPGEP/UFPE), agradeço por todas as oportunidades e pela confiança em mim depositada, assim como pelos grandes ensinamentos e conhecimentos que obtive durante minha jornada.

RESUMO

A segurança pública deve ser estrategicamente concebida para minimizar a criminalidade e garantir um nível de segurança mais elevado. Ao considerar esse contexto, verifica-se que a análise de dados de crime contribui para o estabelecimento de ações que devem ser realizadas para garantir este nível mais elevado de segurança. Nestas análises são considerados dados estruturados de agências governamentais e dados não estruturados de redes sociais, como o *Twitter*, mas não os dois em conjunto. Comprova-se que não há incorporação e integração desses dados, o que permitiria aumentar a precisão de métodos de análise de dados, como algoritmos de aprendizagem de máquinas. Assim, este trabalho propõe uma metodologia híbrida de análise de dados, considerando a integração de dados estruturados e dados não estruturados, que permita que algoritmos de análise de dados, como aprendizagem de máquina, por exemplo, possam realizar as análises necessárias em segurança pública. A integração acontece em duas esferas principais: a primeira, a partir da absorção e análise de dados estruturados disponibilizados por agências governamentais; e a segunda, a partir da absorção, classificação e análise de dados não estruturados, provenientes de plataformas digitais, como é o caso do *Twitter,* plataforma *Onde Fui Roubado* e do *CityCopy*. Com isto, torna-se possível transformar e incorporar esses dados em uma base de repositório única. A partir dessa metodologia híbrida foi construído um sistema de apoio à decisão (*DS.Security*) para que as análises possam ocorrer em segurança pública. O *DS.Security* possui três modelos para ilustrar a aplicabilidade desta metodologia: o primeiro, de classificação de bairros; o segundo, de perfil criminal; e o terceiro, de classificação de bairros levando em consideração a violência contra a mulher. Como resultado, a taxa de precisão dos algoritmos de aprendizagem de máquina, como é o caso do J.48, *ExtraTrees, Naive Bayes* e *Support Vector Machine* aumentou cerca de 10% em comparação com a aplicação desses algoritmos, apenas, em dados estruturados nos modelos associados ao sistema de apoio à decisão.

Palavras-chave: aprendizagem de máquina; dados estruturados; dados não estruturados; segurança pública; *big data.*

ABSTRACT

Public safety should be strategically designed to minimize crime and ensure a higher level of security. It is found that crime data analysis contributes to the establishment of actions that should be taken to ensure this higher level of security. In these analyses, structured data from government agencies and unstructured data from social networks such as Twitter are considered, but not both together. It is found that there is no incorporation and integration of this data, which would allow us to increase in the accuracy of data analysis methods such as machine learning algorithms. Thus, this paper proposes a hybrid data analysis methodology, considering the integration of structured data and unstructured data, which would allow data analysis algorithms, such as machine learning, for example, to perform the necessary analysis for public safety. The integration happens in two main spheres, the first from the absorption and analysis of structured data made available by government agencies, and the second from the absorption, classification, and analysis of unstructured data, coming from digital platforms, as is the case of Twitter, the Where Was I Robbed platform and CityCop. With this, it becomes possible to transform and incorporate this data into a single repository base. Based on this hybrid methodology, a decision support system (DS.Security) was built so that analysis can occur in public security. DS.Security has three models to illustrate the applicability of this methodology: the first is for neighborhood classification; the second is for criminal profile; and the third is for neighborhood classification, taking violence against women into consideration. As a result, the accuracy rate of machine learning algorithms such as J.48, ExtraTrees, Naive Bayes, and Support Vector Machine increased by about 10% compared to applying these algorithms to only structured data in the models associated with the decision support system.

Keywords: machine learning; structured data; unstructured data; public safety; big data

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE EQUAÇÕES

SUMÁRIO

1 INTRODUÇÃO

Garantir a segurança dos cidadãos é um desafio diário para diversas entidades governamentais. Ações para minimização da ocorrência de crimes devem ser eficientes no combate à criminalidade, no entanto, vários problemas interferem na efetividade destas ações e contribuem para soluções não eficazes, como é o caso das tecnologias desatualizadas, integração ineficiente entre os setores responsáveis, qualidade dos dados criminais adquiridos e diversos fatores culturais. Esses são alguns exemplos de problemas que impactam no combate ao crime em países em desenvolvimento, como é o caso do Brasil (FRANCISCO & MARINHO, 2017; MURRAY et al., 2013).

No Brasil, ao considerar este contexto, a falta de segurança impacta profundamente o cotidiano de seus moradores, como demonstram diversos índices, entre eles o índice de Crimes Violentos Letais Intencionais (CVLI), que identifica o número diário de homicídios no país, interligado com violência contra a vida. Este tipo de crime tem o maior efeito prejudicial sobre a sensação de segurança individual (MURRAY et al., 2013). A maioria desses crimes envolve algum tipo de furto acompanhado de homicídio, o que tem repercussão profunda tanto social quanto culturalmente. No que tange aos impactos sociais, um novo crime pode incitar um sentimento de medo na população que passa a considerar que outros indivíduos da comunidade possam sofrer um crime semelhante (MURRAY et al., 2013; STEEVES et al., 2015). Assim, os programas de promoção da segurança pública devem ter como objetivo minimizar a criminalidade e aumentar a sensação de segurança na população.

Para tornar o processo de promoção da segurança pública mais eficiente, é necessário identificar formas de garantir decisões efetivas, uma vez que tais decisões impactarão diretamente a população. No entanto, garantir decisões eficientes requer tecnologias de suporte e uma análise robusta dos dados sobre segurança pública – tanto de grande quantidade de dados produzidos pelos órgãos de segurança pública, quanto dos dados reportados pela própria população (FRANCISCO & MARINHO, 2017; MURRAY et al., 2013; STEEVES et al., 2015).

Na literatura, existem dois tipos de análise de segurança pública: apenas com dados estruturados e apenas com dados não estruturados. Ao se referir a dados

estruturados, os trabalhos se reportam à construção de sistemas de apoio à decisão (COLLADOS & LIBERATORES, 2015), a incorporação de sistemas de informações geográficas (ZHONG, 2020), análise de informações gerais (HASSIO et al., 2020) e métodos para detectar parâmetros envolvidos na segurança (SUN, 2020). Com dados não estruturados, os trabalhos se concentram na análise de previsão e monitoramento em segurança (JAVED et al., 2019; WANG et al., 2020 e YANG et al., 2020). Em geral, tanto os estudos que trabalham com dados estruturados quanto os que trabalham com dados não estruturados fornecem modelos matemáticos que permitem uma análise para estabelecer uma sequência de ações que pode ser realizada para trazer melhorias nessa área. No entanto, verifica-se que não existe uma combinação de dados estruturados e dados não estruturados que possibilite ampliar a análise sobre a segurança pública.

A integração de dados não estruturados com dados estruturados oferece a possibilidade de expansão do banco de dados, incorporando conteúdo gerado por usuários UGC em redes sociais e outras plataformas, permitindo que as análises possam ser mais precisas e com melhores índices de precisão dos algoritmos de aprendizado de máquina.

1.1 OBJETIVOS

Nesta seção serão apresentados o objetivo geral do trabalho, bem como os objetivos específicos.

1.1.1 Objetivo Geral

O objetivo geral deste trabalho é a proposição de uma metodologia híbrida de análise de dados em segurança pública que permita a análise integrada e simultânea de dados estruturados (dados de órgãos de segurança pública) e dados não estruturados (conteúdos gerados por usuários em mídias sociais) contribuindo para melhorias nos desempenhos de algoritmos de aprendizagem de máquina.

1.1.2 Objetivos Específicos

a) Capturar dados estruturados de base de dados oficiais de órgãos de segurança pública.
b) Capturar dados não estruturados a partir de *User Generated Content* (UGC) – conteúdo gerado pelo usuário.
c) Elaborar um sistema de apoio à decisão que embarque a metodologia híbrida proposta.
d) Ilustrar e testar a metodologia proposta a partir de aplicações com dados reais.

1.2 JUSTIFICATIVA

Este estudo justifica-se em três pilares principais: governamental, tecnológico e sociedade. Ao trazer um tema de forte impacto no cotidiano destes três pilares, este estudo se propõe ir além da proposição de uma metodologia e da construção de um sistema de apoio à decisão, objetivando, também, a discussão em torno de soluções eficazes no combate à criminalidade. Ao propor a metodologia e o sistema, o presente estudo busca, portanto, fortalecer a importância da análise de dados e a qualidade destes dados na execução de ações que contribuem para o processo decisório no cotidiano de órgãos governamentais em segurança pública.

No campo governamental, este estudo propõe meios eficazes para analisar a ocorrência de crimes, permitindo que entidades envolvidas na segurança pública obtenham algoritmos mais eficazes que apoiem o processo decisório.

No campo tecnológico este trabalho providencia avanços na integração de dados estruturados e não estruturados em um único repositório, permitindo que algoritmos de aprendizagem de máquina (e outros tipos de algoritmos) possam atuar para contribuir com a segurança pública.

No que se refere ao último pilar, a sociedade, fica evidente a importância deste trabalho e a proposição de uma metodologia híbrida de análise de dados incorporada a um sistema de apoio à decisão para combate à criminalidade. Garantir que ações sejam eficientes contra o crime é fator de importância para o bem-estar da sociedade e para a sua sensação de segurança.

Este trabalho se justifica, portanto, ao propor uma metodologia de integração de dados que difere do que é apresentado na literatura, providenciando, também, um sistema de apoio à decisão exclusivamente para utilização na segurança pública, permitindo que os pilares governo, tecnologia e sociedade estejam cada mais integrados na busca constante de segurança.

Por outro lado, ao providenciar novos meios de integração de dados a partir de uma metodologia híbrida que considera dados estruturados e dados não estruturados, busca-se contribuir com conteúdo inédito para a literatura. A literatura, como será explanado no capítulo 2, detém meios que permitem a análise de dados estruturados e instrumentos que permitem a análise de dados não estruturados, contudo não há uma metodologia capaz de analisar estes dados simultaneamente. Esta obra vem para suprir essa necessidade, permitindo uma nova arquitetura de dados que pode ser empregada em qualquer cenário, de acordo com suas características e particularidades.

1.3 MATERIAIS E MÉTODOS

Os procedimentos metodológicos de uma pesquisa podem adotar diversas características, entre elas: descritivas, exploratórias e aplicada (MARCONI & LAKATOS, 2003). Esta tese tendo em conta os seus objetivos, adota a característica metodológica aplicada, através da qual aborda e desenvolve uma metodologia híbrida de análise de dados dentro do ambiente de segurança pública, buscando analisar dados para descrever fatores associados à segurança pública. A abordagem da pesquisa será quantitativa a partir da absorção de dados estruturados e não estruturados.

Para a condução metodológica, inicialmente, foi realizado um levantamento bibliográfico em fontes especializadas, em consonância com os objetivos da pesquisa, bem como visitas a departamento de segurança pública do estado de Pernambuco, providenciando, assim, uma pesquisa de caráter descritivo. Com o conhecimento adquirido sobre os aspectos envolvidos na segurança pública e seus princípios, bem como a importância de integrar dados estruturados e não estruturados no processo de análise, verificou-se a relevância de produzir uma metodologia híbrida de análise de dados para segurança pública e um sistema de apoio à decisão que embarque

essa metodologia, adotando a relação dessas ferramentas com os pontos estudados, no escopo desta produção acadêmica, e considerando duas etapas: desenvolvimento da metodologia híbrida de análise de dados e desenvolvimento de um sistema de apoio à decisão.

Na primeira etapa, quanto ao desenvolvimento da metodologia híbrida de análise de dados, os estudos em bases literárias, bem como visitas a órgãos governamentais envolvidos em segurança pública foram os pontos iniciais para o início do desenvolvimento da metodologia, que possui duas partes principais: a primeira relacionada a dados estruturados e a segunda a dados não estruturados. Para realizar a primeira parte estudos sobre a análise de dados estruturados foram fundamentais. Tornou-se necessária, nesta parte, a inclusão de softwares para refinamento de dados, como é o caso do Open Refine, bem como o Hadoop Framework, que permitem contribuir com a arquitetura de grande volume de dados, formação de clusters com cerca de dez computadores, bem como armazenamento de dados em sistemas de gerenciamento de banco de dados MySql que finalizam a etapa de construção desse processo.

No que se refere à construção da segunda parte, esta tem em conta a absorção de dados não estruturados que foram capturados de plataformas on-line, como é o caso do Twitter. A obtenção de acesso à versão de pesquisa acadêmica, que permite alcance a mais de 10 milhões de tweets por mês, bem como a arquivos de busca, possibilitou extrair esses dados, definir os domínios que serão levados em consideração e estabelecer os dicionários para a transformação desses dados. Estes dicionários estão em constante evolução e sempre há adição de novas palavras que facilitem o processo de análise.

As etapas acima mencionadas foram disponibilizadas em artigos científicos e congressos. No trabalho intitulado Big Data Analytics to Improve the Decision-Making Process in Public Safety: A Case Study in Northeast Brazil, os autores Turet e Costa, 2018, apresentaram a primeira parte da metodologia, um framework capaz de analisar grande volume de dados levando em consideração dados estruturados. Turet, Costa & Zaraté, 2020, propuseram no trabalho Selection of an Integrated Security Area for locating a State Military Organization (SMO) based on group decision system: a multicriteria approach meios que permitem analisar, dentre grande volume de dados, áreas integradas de segurança (AIS) que são mais propícias a receber um batalhão policial. Para esse trabalho, os autores utilizaram a metodologia de análise de dados

estruturados. Ao avançar nos estudos, Turet e Costa, 2022, desenvolveram o trabalho intitulado Hybrid methodology for analysis of structured and unstructured data to support decision-making in public security em que apresentam a metodologia híbrida de análise de dados, considerando dados estruturados e dados não estruturados para suporte à segurança pública.

Na segunda etapa, o desenvolvimento do sistema de apoio à decisão, denominado DS.Security, agrega práticas de aprendizado de máquina e inteligência artificial disponíveis na literatura, como o algoritmo ExtraTrees e outros, que compreendem uma gama de algoritmos capazes de prever e classificar de acordo com um repositório único de dados. O DS.Security permite a inovação na captura, processamento e análise de dados, direcionando a análise de grandes volumes de dados em tempo real e, ao mesmo tempo, aumentando a precisão dos algoritmos de aprendizado de máquina.

O sistema, primeiramente, recebeu o registro de software do Instituto Nacional da Propriedade Intelectual (INPI), com o número BR512020001467-4, e Turet e Costa, 2020, no trabalho Ds.security v.2 (sistema de apoio à decisão para gerenciamento da segurança pública apresentado no Encontro Nacional dos Programas de Pós-Graduação em Engenharia de Produção (EPPGEP), evidenciaram a estrutura do sistema de apoio à decisão, divulgando-o para a sociedade em âmbito nacional.

Como resultado dessa apresentação e concepção do sistema, o DS.security recebeu o prêmio de Melhor Produto Tecnológico do Ano de 2020 pela Associação Nacional dos Programas de Pós-Graduação em Engenharia de Produção (Anpepro). Em seguida, Turet e Costa, 2021, no trabalho Decision Support System to Brazilian Public Security Management: a hybrid methodology (unstructured and structured data) avançam no sistema de apoio à decisão, com novos modelos e melhorias na metodologia de integração de dados, apresentando-o para a sociedade em âmbito internacional.

Com isso, em 2021, o *DS.Security* e a metodologia híbrida de análise de dados recebem um segundo prêmio, o Best Ph.D. Student paper awards (EWG-DSS). Esse prêmio internacional é concedido a partir do International Conference on Decision Support System Technology (ICDSST). Esta premiação considera o melhor e mais inovador dentre os trabalhos submetidos e apresentados por estudantes ou jovens investigadores nos workshops organizados e nas jornadas de conferências que ocorreram no evento. Por fim, Turet e Costa, 2022, no trabalho Knowledge Discovery

Database Model to classify neighborhoods according to violence against women evidenciam um novo modelo que faz parte do DS.Security e que utiliza a metodologia híbrida de análise de dados para identificar potenciais bairros com criminalidade contra a mulher.

A concepção do *DS.Security*, aliada às bases literárias, permitiu o planejamento e o desenvolvimento de três modelos, que visam ilustrar a aplicabilidade do sistema. Este desenvolvimento ocorreu, também, com apoio da Université Toulouse 1 Capitole, Toulouse – França, a partir de um estágio doutoral na referida universidade. Para construir esses modelos, foram necessárias visitas à secretaria de defesa social do estado de Pernambuco, a fim de analisar as reais necessidades dessa secretaria quanto à análise dos dados. Os dados estruturados foram disponibilizados pela Secretaria de Defesa Social. Quanto aos dados não estruturados, foram coletados em plataformas de internet já existentes, como Twitter e CityCopy.

A primeira aplicação permite a classificação dos bairros em tempo real, de acordo com sua intensidade (hot spots). Os hot spots são regiões que, a partir de uma análise detalhada, mostram-se mais vulneráveis aos crimes e, consequentemente, necessitam de melhores ações para minimizá-los. O sistema permite que essas análises aconteçam em tempo real, considerando grandes volumes de dados, tanto estruturados quanto não estruturados. A aplicação é realizada inteiramente a partir do *DS.Security*, quando é feito o tratamento dos dados, e os resultados são apresentados ao decisor, que executará as ações necessárias. Na análise de dados do sistema, os dados são coletados diariamente, estruturados e não estruturados, de forma a garantir o processamento em tempo real, permitindo assim treinar os algoritmos de aprendizado de máquina para aumentar sua precisão para as análises que precisam ser realizadas.

A segunda aplicação envolve um problema associado à população carcerária de uma região. Como analisar o perfil criminal? Esta pergunta é respondida a partir de um modelo que analisa o perfil criminal de um indivíduo utilizando os dados estruturados disponibilizados por órgãos governamentais e dados não estruturados sociais, que levam em consideração a localidade que o criminoso nasceu e as características do lugar. Com isso, torna-se possível descrever tendências para que ações sejam efetuadas, inclusive, nessas localidades.

A terceira aplicação permite identificar, probabilisticamente, bairros e cidades em que ocorrerão crimes contra a mulher. Os dados são coletados de duas fontes, tanto da Secretaria de Defesa Social, quanto de plataformas como o Twitter, para dados não estruturados. O sistema permite, nesse processo, que essas análises aconteçam em tempo real, tendo como base grandes volumes de dados. A aplicação dos modelos será realizada a partir do DS.Security, onde o usuário bastará selecionar o modelo desejado e verificar os resultados.

1.4 ESTRUTURA DA TESE

Este trabalho está estruturado em sete capítulos, conforme segue:

Capítulo 1: introdução sobre o tema e os conceitos da tese que serão abordados em seu desenvolvimento.

Capítulo 2: embasamento teórico sobre Machine Learning, Big Data, User Generated Content, Algoritmos de Classificação/Predição e estudos relacionados.

Capítulo 3: apresentação da metodologia híbrida de análises de dados, bem como do sistema de apoio à decisão desenvolvido para realizar as aplicações requeridas para esta pesquisa, demonstrando as abordagens utilizadas, o modelo conceitual adotado e as principais funcionalidades.

Capítulo 4: mostra a primeira aplicação do trabalho que analisa e classifica os bairros considerados hot spots.

Capítulo 5: apresenta a segunda aplicação que considera o perfil criminal da população carcerária de uma região.

Capítulo 6: expõe a terceira aplicação que analisa e realiza a previsão de crimes contra a mulher em determinadas regiões.

Capítulo 7: considerações finais sobre a pesquisa, bem como algumas sugestões para trabalhos futuros.

2 REFERENCIAL TEÓRICO E ESTUDOS RELACIONADOS

Este capítulo apresenta definições sobre tópicos fundamentais, como Segurança Pública, Big Data/Conteúdo Gerado pelo Usuário (UGC) e Aprendizado de Máquina. Além destes tópicos, estudos relacionados serão apontados para fundamentar o objetivo geral.

2.1 SEGURANÇA PÚBLICA BRASILEIRA

O contexto da segurança pública norteia o desenvolvimento desta obra. Pode-se compreender segurança pública como o conjunto de ações que tem como finalidade garantir a proteção do público em geral, ou seja, da população (LIMA & MARINHO, 2017). Entre estas ações estão incluídas medidas de precaução que permitam que a população esteja livre de quaisquer tipos de danos e riscos. Estão associadas a estas ações a prevenção e a repressão qualificadas, que têm como finalidade permitir a equidade, a dignidade humana dentro do contexto dos diretos humanos e do estado democrático de direito.

À luz da Constituição Federal, no Art. 144, tem-se:

> Art. 144. A segurança pública, dever do Estado, direito e responsabilidade de todos, é exercida para a preservação da ordem pública e da incolumidade das pessoas e do patrimônio, através dos seguintes órgãos: I- polícia federal; II - polícia rodoviária federal; III - polícia ferroviária federal; IV - polícias civis; V - polícias militares e corpos de bombeiros militares (BRASIL, 1988).

No Brasil, o envolvimento das entidades governamentais, como as citadas no Art. 144 está associado a diversos desafios, considerando a extensão territorial brasileira, a corrupção, a desigualdade social e a tecnologia da informação.

No tocante à extensão territorial brasileiro, verifica-se que o Brasil é um país que detém inúmeras fronteiras, regiões de difíceis acessos e um território com diversas variedades geográficas, o que facilita a inserção do crime em várias

localidades. Associado a este fator está o baixo investimento em policiamento e políticas de prevenção (MURRAY, et,al, 2019).

Quanto ao segundo desafio, a corrupção está ligada ao cidadão convencional, uma vez que a partir da prática corruptiva o bem-estar da população é afetado diretamente, seja em infraestrutura, segurança, habitação, transporte e tecnologia. Recursos são desviados de fundos públicos envolvendo superfaturamento em obras e outros bens públicos. Neste contexto, o investimento em segurança fica prejudicado, permitindo que a criminalidade avance (MURRAY, et al., 2019).

O terceiro desafio está associado diretamente com a desigualdade social que, no Brasil, está interligada com um problema sistêmico intrínseco à relação social, onde há questões econômicas, de gênero, cor, crença ou grupo social. Dentro deste contexto, evidencia-se a grande dificuldade para a população ter seus direitos básicos respeitados. Isto acontece por falta de oportunidade e acesso a esses direitos, como é o caso da educação, fazendo com que pessoas possam se envolver com a criminalidade por se tornar um caminho mais rápido e fácil de ganhar dinheiro (MURRAY, et al., 2019; de LIMA & MARINHO, 2017).

Por fim e não menos importante está o desafio da tecnologia da informação. Compreende-se que a Tecnologia da Informação (TI) é a aplicação da tecnologia para resolver problemas de negócios ou organizacionais em larga escala. Não importa a função, um membro de um departamento de TI trabalha com outras pessoas para resolver problemas de tecnologia. A tecnologia da informação tem como finalidade permitir que órgãos governamentais envolvidos em segurança pública possam identificar ações que visem minorar a incidência de crimes, permitindo que a população tenha a sensação de segurança em sua localidade. Estas ações estão envolvidas diretamente com tecnologias de ponta, com o uso de rastreadores, câmeras de vigilância, softwares de simulação e análise de dados.

A segurança pública é um direito de todos e dever do Estado. Permitir que a sociedade tenha assegurado esse direito é disponibilizar ações que objetivem minimizar a ocorrência de crimes. É vital, nesse sentido, que ações devam efetivamente providenciar resultados positivos, o que necessita de investimento em diversas vertentes, como em sistemas de apoio à decisão que busquem, cotidianamente, analisar grande volume de dados. A análise de dados possibilita direcionar, com maior eficácia, as próximas ações, permitindo gerar novos caminhos no combate à criminalidade.

2.2 *BIG DATA E USER GENARETED CONTENT (UGC)*

Nos últimos anos, repartições públicas e privadas tiveram que trabalhar com diferentes quantidades de dados de diversas fontes. A análise desses dados tem se tornado cada vez mais complexa devido à necessidade de ferramentas robustas e alto poder de processamento. Esta seção discute, nesse contexto, conceitos importantes que são necessários para alcançar uma compreensão mais completa de Big Data e suas ferramentas.

O Big Data e suas ferramentas estão associados a conjuntos com milhões de dados. Pode-se compreender que o Big Data pode se tornar um problema, uma vez que as organizações precisam analisar, constantemente, grandes quantidades de dados e armazená-los, o que requer alto poder computacional e diferentes ferramentas de análises, contudo novas formas de captação de análise destes dados surgem, possibilitando melhorias nesse processo, como é o caso de frameworks (Spark e Hadoop) (RENU, et al., 2013 & GRILLENBERGER, et al., 2014).

A resolução de problemas em volume de dados, neste caso em grande quantidade de dados, baseia-se em cinco características principais: velocidade, volume, valor, variedade e veracidade. São os chamados 5 Vs de Big Data que estão relacionados, diretamente, às diferentes ferramentas e à forma como uma organização lida com seus dados. A seguir, os significados de cada V (figura 2.1) (RENU, et al., 2013 & GRILLENBERGER, et al., 2014):

1. Volume – Big Data é uma grande quantidade de dados gerados a cada segundo. E-mails, mensagens do Twitter, fotos e vídeos que circulam na rede a cada segundo e carregam uma grande quantidade de dados. Em segurança pública, diariamente, dados são coletados, em tempo real, seja para análise constante da criminalidade, como para roteirização de viatura, dentre outros. À medida que estes dados vão sendo coletados surgem diversos desafios, a exemplo da garantia da qualidade; da identificação dos dados; e da sua mensuração e armazenamento.

2. Variedade – No passado, a maioria dos dados era estruturada e podia ser colocada em tabelas e relacionamentos. Atualmente, 80% dos dados mundiais não se comportam dessa forma. Com Big Data, mensagens, fotos, vídeos e sons, que são dados não estruturados, podem ser gerenciados com os dados tradicionais. Os dados não estruturados permitem a ampliação da base com dados que não seriam identificados. Algoritmos de análise de texto, análise de imagem e sons fazem parte

desta arquitetura para mensuração destes dados. Em segurança pública torna-se comum a incorporação de grande variedade de dados que precisa ser analisada e mensurada em tempo real.

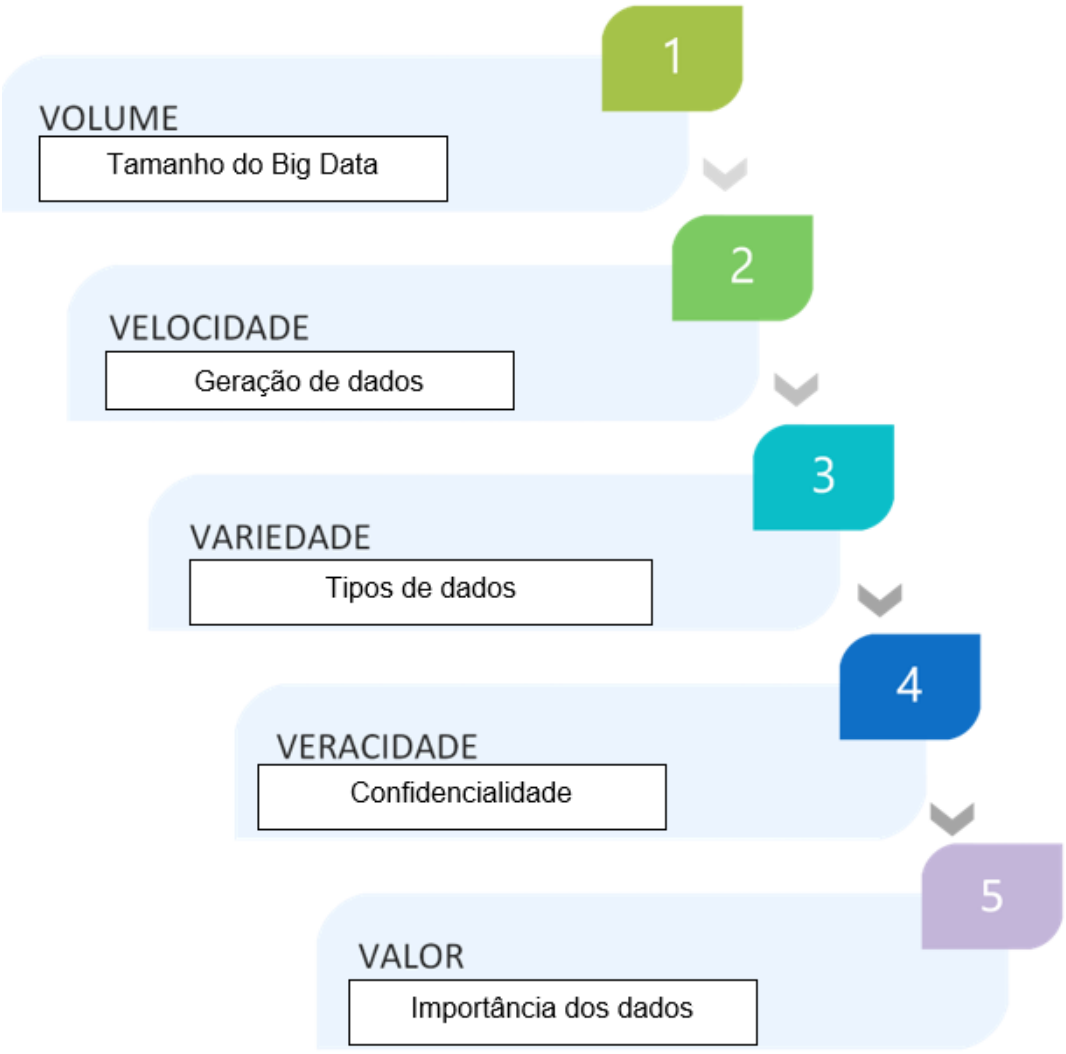
3. Veracidade – Uma das características mais importantes de qualquer informação é que seja verdadeira. Com o Big Data não é possível controlar todas as hashtags do Twitter ou notícia falsa na internet, mas com a análise e estatísticas de grandes volumes de dados é possível compensar informações incorretas, principalmente, quando estes dados estão inteiramente relacionados com segurança pública.

4. Velocidade – Refere-se à velocidade na qual os dados são criados. Isso inclui mensagens de redes sociais virais em segundos; transações de cartão de crédito sendo feitas a qualquer momento, ou os milissegundos necessários para calcular o valor de compra e venda de ações. O Big Data serve para analisar os dados quando são criados, sem a necessidade de armazená-los em bancos de dados.

5. Valor – O último V é o que torna o Big Data relevante: é normal ter acesso a uma grande quantidade de informações a cada segundo, mas isso é inútil, a menos que possa gerar valor. É importante que as empresas entrem no negócio de Big Data, sendo sempre importante lembrar os custos e os benefícios e tentar agregar valor ao que está sendo feito.

Na figura 1 consegue-se identificar a relação entre os 5 Vs de Big Data, considerando o volume de dados, bem como os tipos de dados existentes. Cada V se comporta de uma maneira diferente, providenciando maneiras distintas de análise de dados e integração de algoritmos. Assim, torna-se importante providenciar dados que detenham qualidade, minimizando os erros da base, refinando-a e permitindo que as soluções encontradas estejam de acordo com a realidade vivenciada.

Figura 1 – Relação entre os 5 Vs de Big Data



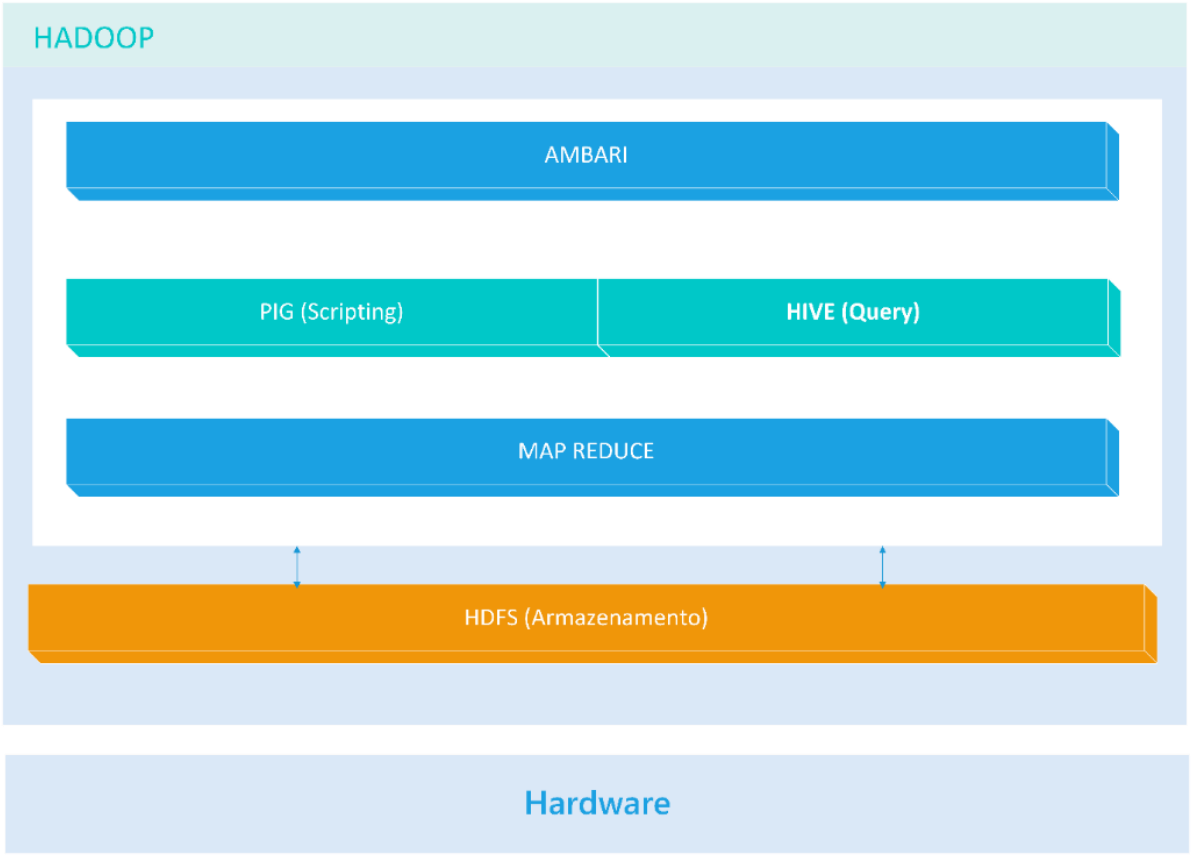
Fonte: Esta Pesquisa (2022)

Por outro lado, *o Big Data Analytics* são técnicas analíticas aplicadas à *Big Data e* o *Big Data Analytics* são as ferramentas analíticas e o trabalho inteligente realizado em grande volume de dados, estruturados ou não estruturados, que são coletados, armazenados e interpretados por softwares de alto desempenho. Envolve a verificação cruzada de uma matriz quase ilimitada de dados do ambiente interno e externo (ELGENDY, *et al.*, 2016; WHITE, 2015).

Entre as ferramentas de *Big Data Analytics,* tem-se *Hadoop, Map Reduce* e *Open Refine*. A biblioteca *Hadoop* (figura 2) é um *framework* que permite trabalhar com grandes conjuntos de dados através de *clusters* de computadores, utilizando modelos de programação (WHITE, 2015). Ele foi projetado para garantir ampla escalabilidade de um único servidor a um *cluster* com milhares de máquinas, cada

uma oferecendo capacidade de armazenamento e computação local. Em vez de depender de *hardware* para fornecer maior disponibilidade, a própria biblioteca é projetada para detectar e tratar falhas da camada de aplicação de modo a fornecer um serviço de alta disponibilidade baseado em uma grade de computadores.

Figura 2– Apache Hadoop Ecossistema



Fonte: Esta pesquisa (2022)

Já o *Map Reduce* é uma etapa para o processamento de grandes quantidades de dados, em paralelo, utilizando todo o *hardware* disponível no *cluster Hadoop.* O modelo de programação *Map Reduce* consiste na construção de um programa formado por duas operações básicas: mapear e reduzir, enquanto o *Open Refine* limpa os dados e os transforma em outras formas (WHITE, 2015).

Nesse contexto, um grande volume de dados também está relacionado com a absorção de dados não estruturados em que estão relacionados com imagens, textos e frases disponibilizados em diversas plataformas. Por definição, este tipo de dado está relacionado com o *User Generated Content (UGC).* Pode-se compreender que *UGC* significa conteúdo gerado pelo usuário. Por definição, conteúdo gerado pelo usuário é qualquer forma de conteúdo – texto, postagens, imagens, vídeos, resenhas

etc. – criado por pessoas, individualmente (não marcas) e publicado em uma rede social ou *on-line* (WANG, et.al, 2019).

O uso do *UGC* é benéfico para empresas, órgãos governamentais e usuários. Empresas e órgãos governamentais se beneficiam de uma quantidade elevada de novas ideias e conteúdo exclusivo. Além disso, na maioria das vezes, o conteúdo do usuário não exige um grande investimento, uma vez que este está disponível em diversas redes sociais. Para os usuários, a criação do *UGC* permite que colaborem com empresas, expressem-se, permitindo que seus desejos sejam direcionados para empresas e entidades governamentais.

2.3 *MACHINE LEARNING (APRENDIZAGEM DE MÁQUINA)*

Machine Learning (ML) é um ramo da estatística e da computação que reúne uma série de métodos que possui dois objetivos principais: o primeiro é o desempenho preditivo dos modelos e o segundo é automatizar o processo de modelagem de bancos de dados observados ou aprendizado com dados observados. Além disso, grande parte do ML envolve métodos para calcular o erro de validação (erro de previsão fora da amostra) e selecionar ou ponderar modelos com base nesses erros de previsão. Ao se tratar do desempenho preditivo de modelos, a precisão das estimativas ou previsões é um aspecto importante. Uma parte importante da literatura de ML, nesse sentido, é dedicada a reduzir a variância das estimativas ou a parte redutível da variância das estimativas. A parte não redutível da variância não pode ser trabalhada, por exemplo, por omissão de variáveis do modelo (QUINLAN, 1986; PETER, 2018).

Autores têm identificado que o aprendizado de máquina providencia um processo de automação que permite que sistemas possam:

> [...] permitir o desenvolvimento de técnicas computacionais de aprendizagem, bem como a construção de sistemas capazes de adquirir conhecimento automaticamente. Um sistema de aprendizagem é um programa de computador que faz decisões baseadas em experiências acumuladas através da solução bem-sucedida de problemas anteriores. (MORONEY, 2020, p .12

O aprendizado de máquina tornou-se uma ferramenta de resposta importante para diversas áreas, entre elas a segurança pública, e está sendo usado em uma

variedade de tecnologias de ponta, fazendo parte de um aglomerado de ferramentas pertencentes à inteligência artificial que inclui redes neurais, percepção, robótica, gestão do conhecimento, processamento de linguagem natural e a própria aprendizagem de máquina (figura 3).

Figura 3 - Componentes da Inteligência Artificial

Fonte: Adaptado de White (2015)

A aprendizagem de máquina possui uma divisão clara quanto aos seus tipos de algoritmos. A divisão acontece em algoritmos supervisionados, não supervisionados e reforçado, onde (PETER, 2018):

Aprendizagem de máquina supervisionada

Tem como base a regressão básica, que pode ser a linear e a logística, por exemplo, bem como a classificação de comportamento dos dados. Neste tipo, o humano participa do processo oferecendo um banco de dados com o objetivo de treinar esse algoritmo para que se possa reconhecer padrões e semelhanças. A título de exemplificação, pode-se indicar a aplicação em tipos de bicicletas. Com um banco de dados de treinamento que possui várias informações sobre bicicletas, o algoritmo irá absorver como conhecimento os padrões dessas bicicletas, como pedais, rodas, guidões e tantos outros elementos-chave. Ao realizar a classificação de futuras bicicletas, ou até mesmo, classificar o que é bicicleta e o que não é bicicleta, será fácil e intuitivo (MORONEY, 2020).

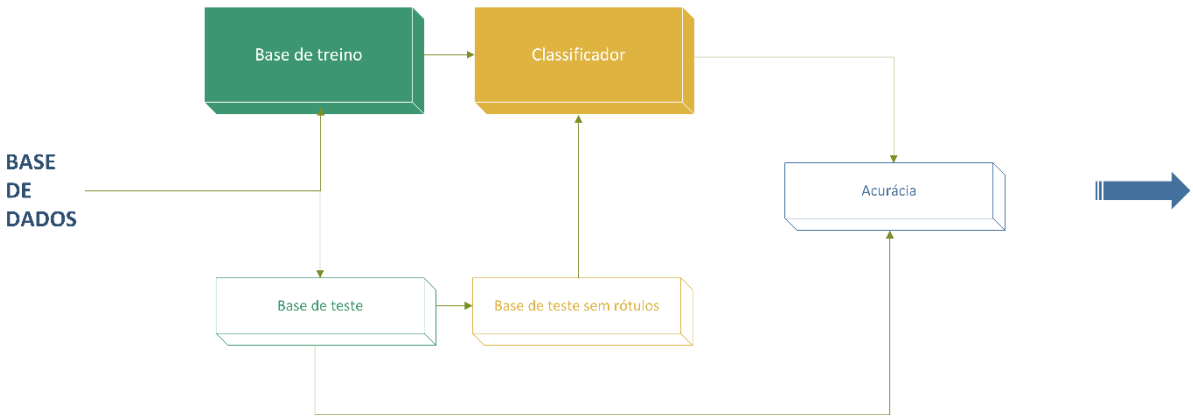
Aprendizado não supervisionado

Verifica-se que há um aprendizado a partir de dados que não foram rotulados, ou seja, não foram classificados ou categorizados inicialmente. Este tipo de aprendizagem de máquina identifica semelhanças em determinado volume de dados e acaba reagindo com a presença ou ausência de semelhanças em novos dados. Há pouca intervenção humana. Algoritmos de agrupamento, como o K-means, estão entre estes tipos de algoritmos (MORONEY, 2020).

Aprendizado reforçado

Neste tipo de algoritmo há o ensinamento baseado na experiência, com isso, a máquina deve lidar com seus próprios erros e acertos. Basicamente, há uma tentativa e erro que permitem lidar com o que errou a fim de buscar uma abordagem mais correta para entregar resultados precisos (MORONEY, 2020).

Figura 4 – Funcionamento de um algoritmo de aprendizagem de máquina



Fonte: Esta pesquisa (2022)

Quanto ao funcionamento de um algoritmo de aprendizagem de máquina (figura. 4), pode-se compreender que são executadas cinco etapas principais (MORONEY, 2020):

No primeiro momento, tem-se a aquisição dos dados, que podem ter duas categorias predefinidas: a primeira, associada a dados estruturados, ou seja, que possuem uma estruturação mais rígida e são previamente definidos. São dados categorizados como quantitativos, enquanto os dados não estruturados não são organizados e não possuem um formato predefinido.

No segundo momento, a base de dados é dividida de duas formas: dados para treino e dados para teste. No primeiro, o algoritmo de aprendizagem de máquina realiza o treinamento considerando os dados e as suas características.

Há o reconhecimento de padrões e a geração de conhecimento da base de dados (terceira e quarta etapas). Em seguida, o algoritmo realiza a classificação/predição dentro da base de testes para identificar o resultado final (quinta etapa). Ao fim, testes de acurácias são realizados para verificar o desempenho do algoritmo no problema supracitado.

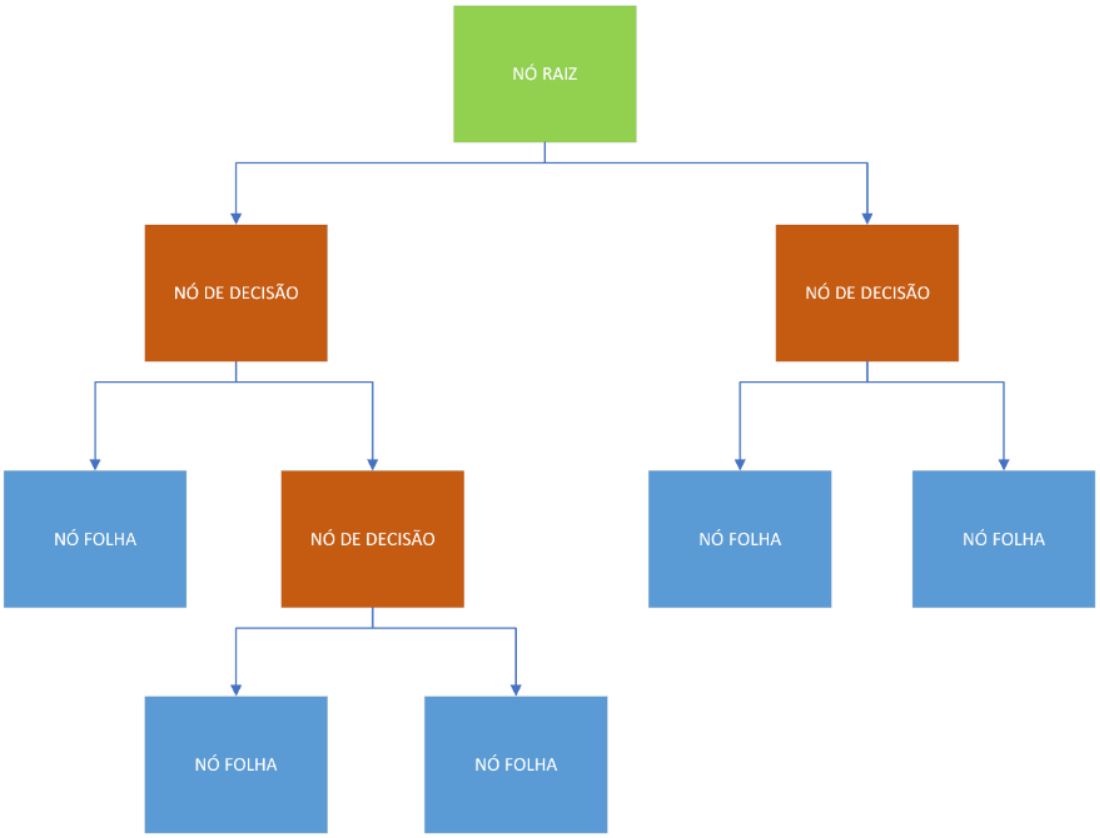
No funcionamento de um algoritmo de aprendizagem de máquina, deve-se observar que os algoritmos estão relacionados com problemas básicos do cotidiano, como os relativos à classificação, predição e previsão. São problemas associados com diversos tipos de algoritmos existentes na literatura, como é o caso dos

algoritmos J.48, ExtraTrees, Árvores de Decisão, Support Vector Machine (SVM) e Naive Bayes que serão utilizados no desenvolvimento da tese.

2.3.1 Decision Trees (Árvore de Decisão)

A primeira abordagem que será retratada neste capítulo está associada às Árvores de Decisão (figura 2.5). A Árvore de Decisão é uma estrutura de dados hierárquica que implementa a estratégia de dividir e conquistar. É um método não paramétrico eficiente que pode ser usado tanto para a classificação quanto para a regressão (QUINLAN, 1986; MORONEY, 2020).

Figura 5– Estrutura de Árvore de Decisão



Fonte: Esta pesquisa (2022)

A Árvore de Decisão é composta por nós de decisão. Cada nó de decisão implementa uma função de teste fm(x) com um valor de saída discreto que se torna a entrada para o próximo nó de decisão, e assim por diante (MORONEY, 2017).

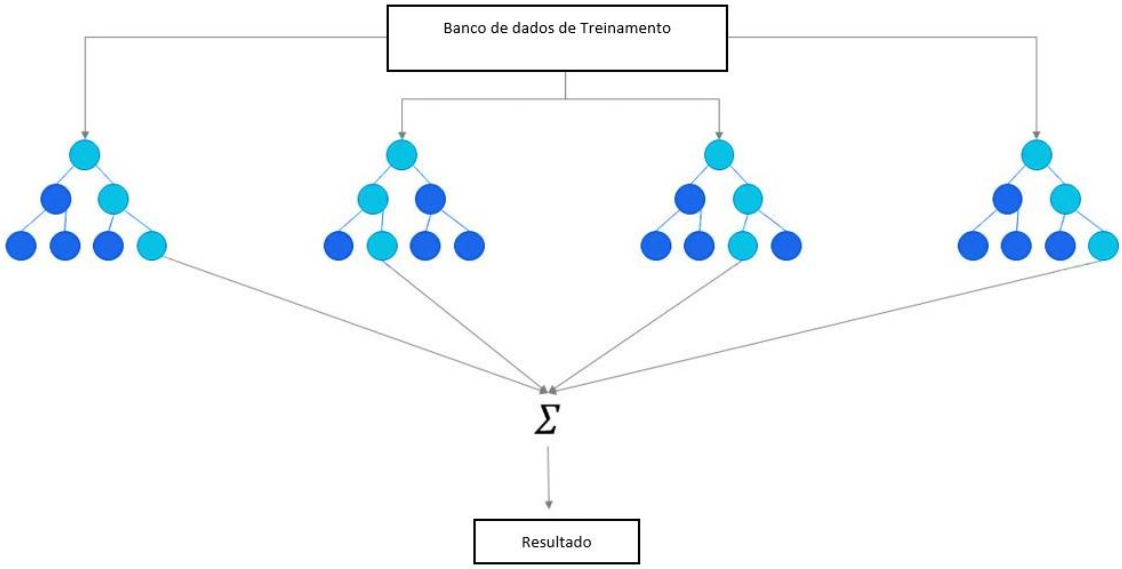
Cada função fm(x) implementa um domínio das variáveis de entrada. O nó raiz geralmente divide o domínio de entrada em duas partes similares, os próximos nós dividem novamente o domínio de entrada em duas partes similares, e dessa forma sucessivamente. Com esta estratégia, é possível mapear os valores de entrada e descobrir o(s) valor(es) de saída com n interações.

2.3.2 Florestas Aleatórias

Algoritmos de Florestas Aleatórias, também chamados de Random Forest, são uma tecnologia de aprendizado de máquina que tem como finalidade resolver problemas envolvidos com regressão e classificação, utilizando técnicas que combinam classificadores e pode oferecer resoluções para problemas de grande complexidade (QUINLAN, 1986; MORONEY, 2020).

São algoritmos que possuem três hiperparâmetros que necessitam de definição antes do treinamento com a base de dados. Estes hiperparâmetros são: tamanho do nó, número de árvores e número de recursos de amostras. Com estes valores, o algoritmo realiza a classificação ou a regressão.

Figura 6– Estrutura de Florestas Aleatórias



Fonte: Esta pesquisa (2022)

Como visto na figura 6, uma Floresta Aleatória é composta de diversas árvores, ou seja, uma coleção de árvores e cada árvore é composta de uma determinada amostra que é extraída de um conjunto de dados, considerando essa amostra, um terço é reservado para o treinamento do algoritmo e a parte restante para realizar a classificação.

Dentro do contexto de Árvores Aleatórias, tem-se o *ExtraTrees*. O *ExtraTrees*, também conhecida como Árvore Randomizada Extrema, gera modelos preditivos para problemas de classificação e regressão. É semelhante a outros métodos, como Árvores de Decisão e Florestas Aleatórias, mas usa informações extras sobre os dados para melhorar a precisão preditiva. Além disso, o algoritmo de árvore *extratrees* é mais rápido e fácil de implementar do que outros métodos. Como resultado, é uma ferramenta para mineração de dados e modelagem preditiva (QUINLAN, 1986; MORONEY, 2020).

2.3.3 Naive Bayes

É um tipo de classificador que se baseia no teorema de *Bayes.* O Teorema de *Bayes* é uma fórmula matemática simples usada para calcular probabilidades condicionais. A probabilidade condicional é uma medida da probabilidade de um evento ocorrer, dado que outro evento, por suposição, presunção, afirmação ou evidência aconteceu (MORONEY, 2020).

Equação 1 – Equação *Naive Bayes*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

A fórmula (equação 1) tem como objetivo identificar com que frequência A acontece, dado que B acontece, escrito P(A|B) também chamada de probabilidade posterior, quando sabe-se: com que frequência B acontece, dado que A acontece, escrito P(B|A) e qual a probabilidade de A por si só, escrito P(A) e qual a probabilidade de B estar sozinho, escrito P(B).

Este é um classificador que tem como premissa a independência das variáveis do problema. O algoritmo, neste contexto, realiza uma classificação probabilística a partir de classes predefinidas.

Tipos de Naive Bayes:

1. Classificador multinomial Naive Bayes

Os vetores de características representam as frequências com que certos eventos foram gerados por uma distribuição multinominal. Este é o modelo de evento normalmente utilizado para a classificação de documentos.

2. Classificador Bernoulli Naive Bayes

No modelo de evento multivariado de Bernoulli, os recursos são booleanos independentes (variáveis binárias) que descrevem as entradas. Como modelo multinominal, este modelo é popular para tarefas de classificação de documentos em que os recursos de ocorrência de termos binários – ou seja, uma palavra, acontece em um documento ou não –, são usados, em vez de frequências de termos – ou seja, a frequência de uma palavra no documento.

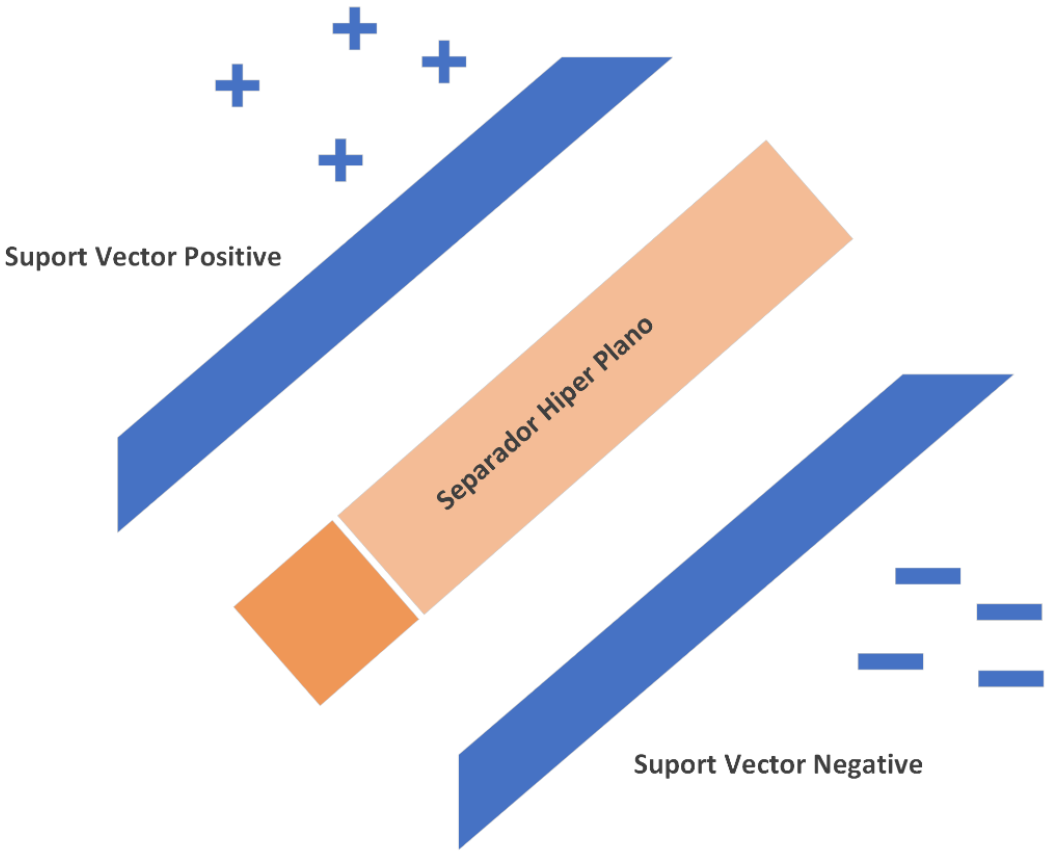
3. Classificador Gaussiano Naive Bayes

Em Gaussian Naive Bayes, assume-se que os valores contínuos associados a cada feição são distribuídos de acordo com uma distribuição Gaussiana (distribuição normal). Ao ser plotado, fornece uma curva em forma de sino que é simétrica em relação à média dos valores de recursos.

2.3.4 *Support Vector Machine (SVM)*

O *SVM* é um algoritmo de aprendizagem de máquina que estabelece um modelo a partir de entradas de treinamento, mapeando-as em espaço multidimensional e usando regressão para encontrar um hiperplano, que é uma superfície de espaço n-dimensional que separa o espaço multidimensional em duas metades, que melhor separa duas classes de entradas. Uma vez que o *SVM* tenha sido treinado, pode avaliar novas entradas em relação ao hiperplano divisor e classificá-las em uma das duas categorias (INGO & ANDREAS, 2008).

Figura 7– Estrutura de Support Vector Machine (SVM)



Fonte: Esta pesquisa (2022)

Um SVM (figura 7) é essencialmente uma máquina de entrada/saída. Um usuário pode inserir uma entrada e, com base no modelo desenvolvido através do treinamento, retornará uma saída. O número de entradas para qualquer SVM

especificado varia teoricamente de um a infinito. No entanto, em termos práticos, a capacidade computacional limita o número de entradas que pode ser utilizado. Se, por exemplo, N entradas são usadas para um SVM específico – o valor inteiro de N pode variar de um a infinito –, a máquina deve mapear cada conjunto de entradas no espaço de N dimensões e encontrar um hiperplano de N −1 dimensões que melhor separam os dados de treinamento (INGO & ANDREAS, 2008).

É importante ressaltar que esses atributos podem e devem ser modificados de acordo com a realidade enfrentada pela região de aplicação.

2.4 ESTUDOS RELACIONADOS

Nesta seção temos estudos que estão envolvidos com o tema de segurança pública e sua relação com as diversas tecnologias, como é o caso de Big Data, aprendizagem de máquina e user ganareted content, considerando os dados estruturados e os dados não estruturados.

Dentre os trabalhos no contexto de segurança pública utilizando dados estruturados, Collados & Liberatores (2015) elaborou um SAD e verificou que ele poderia apoiar o uso correto de recursos humanos (otimização) em delegacias. O estudo apresentou um SAD que, com base em um modelo matemático, pode fornecer sugestões para o patrulhamento preditivo na Espanha. Cohen, et.al, (2020) desenvolveram um modelo de série temporal para a previsão de crimes em uma área selecionada, utilizando reuniões com as partes envolvidas para coleta de dados e posterior análise. No entanto, esse estudo apresentou um SAD baseado apenas em dados estruturados e não utilizou algoritmos de ML para realizar a análise de predição.

Figueiredo & Mota, (2020) apresentaram um modelo multicritério de apoio à decisão que tinha como finalidade realizar a classificação e a análise de locais de acordo com diversos aspectos da violência nessas áreas. Este modelo buscou considerar múltiplos objetivos, utilizando visualizações gráficas que foram usadas para identificar áreas que precisavam de mais recursos para combater a violência.

Os estudos de Gong, et al., (2020); Williams, et al., (2017); Poleto, et al., (2017) estão em um fluxo de pesquisa semelhante. Todos esses trabalhos consideraram a possibilidade de desenvolver um sistema de segurança pública que incorpore tecnologias de Big Data. Esses trabalhos demonstraram a dificuldade de padronização

de grandes volumes de dados em que há textos, imagens e vídeos que exigem tecnologias diferentes no processo de análise. Elgendy & Elragal, (2016); Renu, et al., (2013) mostraram a importância do Big Data no desenvolvimento de SADs. Com o alto volume de dados disponível para um SAD, as ferramentas de Big Data podem ser um meio importante de analisar diferentes tipos e grandes quantidades de dados. Anusha, et al., (2022) realizam um processo de aprendizado profundo para a detecção de navios, a partir de imagens oriundas de satélites. De acordo com os autores, a detecção de navios proporciona maior segurança marítima, bem como, controle de poluição e vazamento de óleos. Como resultado, o modelo é capaz de detectar a região em que o navio se encontra e a quantidade de navios.

Hassio, et al., (2020) examinaram a natureza das demandas de informação sensível ao contexto, com foco nas articulações da necessidade de informação disnormativa entre usuários de drogas. Zhang, et al., (2020) desenvolveram um SAD geográfico que analisa dados estruturados para planejamento de segurança urbana, dividindo-os em três partes: gestão de sistemas, mecanismos operacionais e plataformas tecnológicas. Lima, et al., (2020) propuseram uma investigação do potencial dos classificadores de ML para identificar trabalhadores de segurança propensos ao absenteísmo de longo prazo. Os preditores puderam tomar decisões com base no histórico profissional de cada agente, que pode ser extraído dos bancos de dados das instituições de segurança pública.

Rupa, et al., (2020) avaliam ataques cibernéticos com base em técnicas de aprendizado de máquina. Este modelo computacional permite analisar e classificar ofensas a partir de dados estruturados ou dados não estruturados. Como resultado, a abordagem classifica as infrações com precisão acima de 90%. Rupa, et al., (2021) seguem temática parecida ao providenciar um estudo que explora as diferentes maneiras para a detecção de links maliciosos na internet. Para isso, os autores utilizam recursos baseados em host e lexicais da própria URL. Chiramdasu, et al., (2021) estabelecem uma abordagem de aprendizagem de máquina com a finalidade de identificar usuários maliciosos a partir da URL. Como resultado tem-se uma nova estrutura utilizando de regressão logística que detecta essas URLS.

Trabalhos que estão associados à análise de dados estruturados trazem resultados interessantes ao usar modelos de aprendizagem de máquina dentro do contexto de segurança. Contudo, uma abordagem híbrida, com a inclusão de dados não estruturados, poderia contribuir para otimizar os modelos construídos,

incorporando, para exemplificar, dados gerados pelo usuário, User Generated Content, como é o caso de imagens e textos. Estes dados permitem potencializar a realidade vivenciada pela população em localidade com incidências de crimes ou outros cenários.

No contexto de segurança pública utilizando dados não estruturados, Xue & Brown, (2006) desenvolveram um modelo para prever o comportamento criminoso usando escolha espacial com técnicas de mineração de dados. Primeiramente, foram estabelecidos padrões de comportamento nas localidades estudadas. Depois, o trabalho elaborou modelos de análise espacial para comparar o desempenho da metodologia proposta com os métodos mais tradicionais de previsão de crimes. Os autores, no entanto, focaram na análise de mineração de dados usando classificação estruturada e não incluíram a análise de dados não estruturados. Wang, et al., (2019) propuseram um modelo de previsão de segurança com base na previsão de segurança espaço-temporal capaz de capturar a dependência espacial e temporal. O estudo mostrou a aplicabilidade do modelo proposto por meio de um exemplo envolvendo previsão de segurança no trânsito. Javed, et al., (2019) propuseram um algoritmo de ML para prever cibercriminosos no Twitter, considerando downloads drive-by.

Soares, et al., (2020) sugeriram uma estrutura genérica para extração de conhecimento, que aplicaram à previsão de taxas de criminalidade na cidade de São Paulo, Brasil. O modelo utilizou uma metodologia híbrida que incorporou redes neurais artificiais (RNAs) e modelos matemáticos. A metodologia foi aplicada a um banco de dados e obteve índices de assertividade em torno de 83,12% na previsão de crimes. Silva, et al., (2017) apresentaram um sistema de análise de crimes na cidade do Rio de Janeiro, Brasil, que visava extrair informações importantes para a projeção e a previsão de crimes por meio da extração de conhecimento de bancos de dados. Segundo os autores, o sistema desenvolvido pode contribuir para fornecer informações para a polícia e demais órgãos envolvidos na segurança pública para melhorar os processos de tomada de decisão.

Wawrzyniak, et al., (2020) desenvolveram modelos para a previsão de crimes usando Redes Neurais Artificiais (RNAs), uma forma de aprendizado profundo. Os autores utilizaram duas bases de dados como estudos de caso e obtiveram resultados úteis em termos de sucesso dos algoritmos. Os algoritmos alcançaram maiores índices de precisão, indicando que sua previsão de crime estava de acordo com a realidade vivenciada. Luong, (2011) analisa se a avaliação de risco e a necessidade estavam

relacionadas à gestão de casos de jovens que cometeram algum tipo de crime e se a gestão estava interligada com casos de reincidência. Com o uso de algoritmos, identificaram que alguns pontos estavam relacionados à redução da reincidência, entre eles o princípio da necessidade.

Rummens & Hardyns, (2020) analisaram o impacto de várias técnicas de previsão de crimes e seu desempenho em tarefas de classificação e previsão usando dados sobre arrombamentos domésticos em uma cidade na Bélgica e realizaram a previsão de crimes utilizando parâmetros predefinidos. McFadzien & Sherman, (2021) identificaram como a via de manutenção permite baixas taxas de falso-negativos em investigações. O modelo obteve uma taxa de precisão maior em comparação com os modelos de três anos anteriores. Summers & Rossmo, (2021) examinaram como os infratores aquisitivos crônicos contribuem para o patrulhamento policial usando dados quantitativos e qualitativos. Os resultados indicaram que, quando há patrulhamento policial, os criminosos simplesmente se deslocam para outra região e não desistem de cometer nenhum tipo de crime.

Lamari, et al., (2020) apresentaram uma estrutura de ML capaz de prever a ocorrência de crimes espaciais sem usar crimes passados como preditor. A estrutura foi baseada em uma profunda literatura multidisciplinar que permitiu a seleção de 188 preditores de crime mais adequados, incluindo aspectos sociais, demográficos, espaciais e ambientais. Um estudo de caso foi realizado nos Estados Unidos e o framework obteve uma taxa de acerto de 73%. Olver, et al., (2009) realizaram uma meta-análise para identificar a precisão preditiva de instrumentos forenses. Incluídos nestas análises estavam: reincidência geral, não-violenta, violenta e sexual. As correlações encontradas em seus resultados foram significativas.

As obras que estão associadas à análise de dados não estruturados, portanto, produzem análises robustas e eficientes em segurança pública. Nestes cenários evidencia-se também a importância de incorporação de dados estruturados. Dados que, baseando-se em fontes oficiais, permitem que algoritmos detenham de uma base histórica elevada e obtenham uma base de treinamento mais robusta, o que demonstra a importância de abordagem híbrida.

Trabalhos em segurança pública envolvendo UGC, uma forma de dados não estruturados, também analisaram plataformas de redes sociais como o Twitter. Li, et al., (2010) produziram um framework que utilizou tags espaço-temporais do Twitter para a análise e a previsão de crimes. Essa estrutura possibilita coletar e analisar

dados e prever crimes em uma determinada região. Gerber, (2014) apresenta uma pesquisa investigativa usando tags espaço-temporais do Twitter para prever crimes. O estudo realizou testes estatísticos e análises linguísticas com o objetivo de identificar automaticamente as discussões entre usuários do Twitter em uma cidade dos Estados Unidos.

Win, et al., (2020) discutiram a importância de incluir as impressões digitais no processo de investigação criminal. Os autores relataram a importância dos algoritmos de ML na identificação e mensuração desses dados que são considerados não estruturados. Estudo que transformou impressões digitais em código binário e mostrou que algoritmos de ML podem usar essas representações binárias para classificar automaticamente comportamentos e tendências nesses processos, facilitando a análise de dados.

Rajesh, et al., (2012) preocupam-se no tráfego de informações efetuados por seres humanos. Os dados sensíveis são disponibilizados diariamente em diversas plataformas, quando os autores propõem uma abordagem usando o teorema do resto Chinease para fornecer maior privacidade em dados, momento em que é realizado um processamento de transformação dos dados extraídos de servidores para clientes. Harikrishna, et al., (2022) propõem uma abordagem para a detecção e a classificação de objetos, melhorando a precisão para a identificação de objetos semelhantes. Algoritmos de SVM foram utilizados para realizar estas análises.

Os trabalhos mencionados nesta seção verificam que vários estudos, no contexto de segurança pública, têm como objetivo usar Big Data, incluindo dados estruturados e dados não estruturados, ML e User Generated Content (UGC) por meio de bancos de dados para análises preditivas, de classificação e agrupamento, bem como o desenvolvimento de modelos matemáticos para melhorar a segurança pública. A literatura existente, no entanto, tem focado principalmente na análise de dados que são exclusivamente estruturados ou exclusivamente não estruturados. Nenhuma metodologia até o momento combinou os dois. Propõe-se, neste trabalho, diferenciando-o da literatura existente, uma metodologia híbrida capaz de unir esses dados para alcançar resultados mais abrangentes e aumentar a possibilidade de melhorar o processo de tomada de decisão.

3 METODOLOGIA HÍBRIDA DE ANÁLISE DE DADOS E *DS.SECURITY*

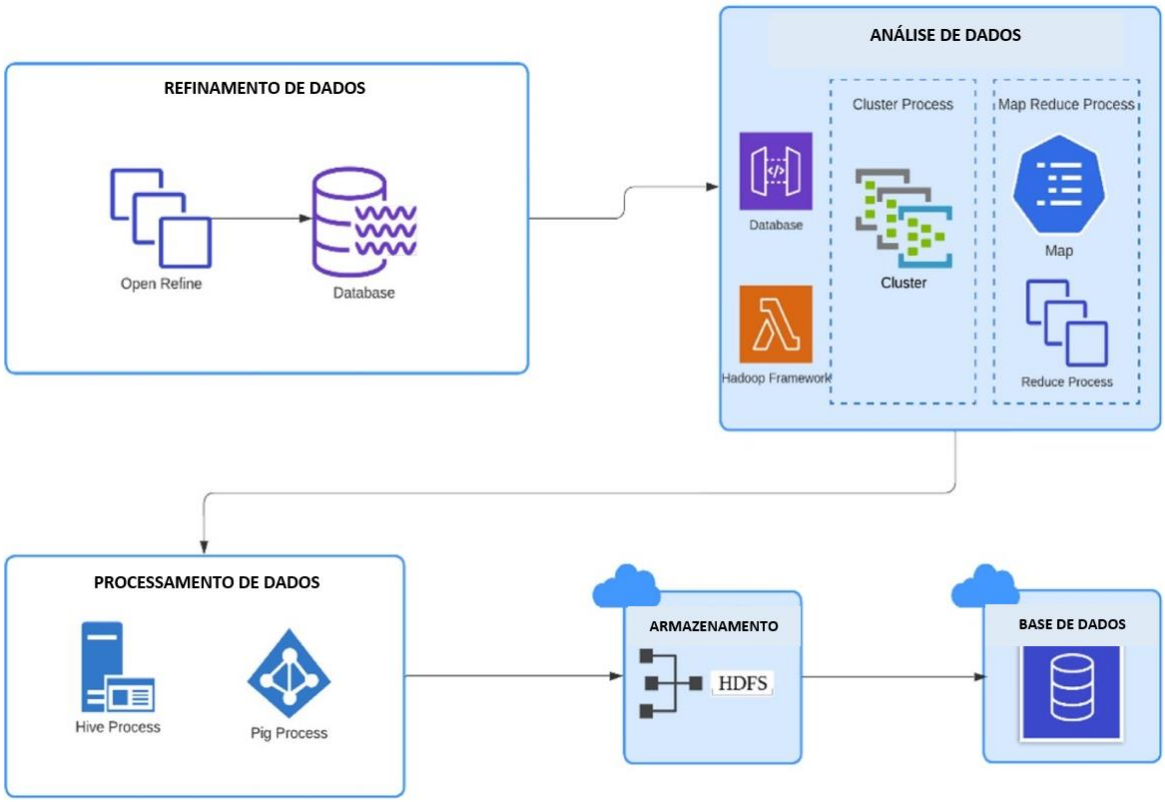
Neste capítulo será retratada a metodologia híbrida que permite a análise de dados estruturados e não estruturados, possibilitando a inclusão destes dados um único repositório, onde algoritmos de aprendizado de máquina podem atuar para realizar diversas atividades, entre elas a classificação de bairros, considerando a criminalidade, bem como a predição de crimes. Metodologia que está embarcada em um sistema de apoio à decisão chamado *DS.Security,* onde será retratado o funcionamento da arquitetura geral do sistema de apoio à decisão.

Em relação à metodologia híbrida, há dois frameworks principais: para os dados estruturados e para os dados não estruturados. Esta metodologia permite a concepção de um banco de dados que incorpore dados estruturados e dados não estruturados. O *DS.Security* traz três modelos para mostrar a aplicabilidade desta metodologia, contudo, a partir da base de dados estabelecida, pode-se criar outros tipos de modelos e aplicações de algoritmos, dependendo das necessidades no cenário vivenciado

3.1 *FRAMEWORK* PARA ANÁLISE DE DADOS ESTRUTURADOS

O processo de absorção e de refinamento de dados estruturados requer várias etapas, dado o grande volume de dados em cenários de policiamento público. É necessário, pois, estabelecer uma sequência lógica de ações para processar esses dados. No presente trabalho, estabelece-se uma estrutura que permite extrair, refinar, analisar, processar e armazenar esses dados, que são obtidos de órgãos governamentais e fornecem informações como local da ocorrência do crime, tipo de crime cometido e localização das viaturas policiais. A figura 8 ilustra o funcionamento do framework proposto:

Figura 8– Framework para dados estruturados



Fonte: Esta pesquisa (2022)

Onde:

1. Na etapa de refinamento, os dados são refinados para evitar possíveis erros que estão relacionados com a captura de dados. Nesta etapa é utilizado o algoritmo *Open Refine* para eliminar erros de espaço em branco, bem como erros ortográficos. O objetivo é garantir a qualidade dos dados. O *Open Refine* permite que um conjunto de dados seja analisado em sua totalidade, incluindo grandes volumes. É simples de usar e não requer grande poder computacional.

Seu passo a passo é simples. No primeiro momento há necessidade de carregamento da base de dados. Esse carregamento pode acontecer a partir de diversos tipos de formatos, como é o caso de *CSV, SQL* e *JSON*. O *software* permite também a inclusão direta com alguns tipos de sistemas de gerenciamento de banco de dados, como é o caso do *MySQL*. Após o carregamento, basta realizar o refinamento que desejar. Há diversos tipos de filtros, chamados de facetas, que permitem a eliminação de espaços em branco, erros em dados, dentre outros. Após

as modificações, os dados limpos são atualizados automaticamente pelo *Open Refine* no repositório geral em preparação para a segunda etapa.

2. Na etapa de análise dos dados, os dados estruturados foram mapeados para identificar semelhanças entre si. Este processo, denominado *mapear e reduzir*, possibilita eliminar linhas, garantindo assim a compacidade da base de dados para análises futuras. Processo que foi realizado utilizando o *Hadoop*, um *framework* que trabalha com grandes volumes de dados. No que diz respeito à análise de dados, a criação de *clusters* permite que vários computadores trabalhem simultaneamente com um grande volume de dados. Nesta etapa, é necessário estabelecer um *cluster* computacional para que vários computadores possam analisar os dados. Cada computador faz o mapeamento e a redução da base para que a terceira etapa possa começar. Um ponto importante nesta etapa: caso o usuário detenha de computadores com alto poder de processamento, um *cluster* não será necessário.

Para o estabelecimento de um *cluster* computacional, basta a instalação do *Hadoop Framework* nas máquinas que irão compor o *cluster*. O *Hadoop* estabelece que um computador será mestre e os demais serão computadores auxiliares. O ideal é que este processo aconteça com o sistema operacional Linux, que possui melhor desempenho e providencia maior robustez ao processo de clusterização.

Esta clusterização propicia que o *Hadoop* mapeie e reduza a base de dados refinada de acordo com o reconhecimento de padrões entre os dados. Esse processo é dividido entre as máquinas que fazem parte do *cluster*, permitindo que cada uma execute essas atividades. Este é um processo automático feito pelo *framework*.

3. Na etapa de processamento dos dados foram utilizadas as ferramentas de análises *Hive e Pig*. O *Hive* permite que *scripts* sejam usados para consultar quaisquer dados de acordo com as regras definidas pelo usuário. O *PIG*, usando uma linguagem de programação própria, processa esses dados segundo as necessidades do cenário do usuário. Ao precisar trabalhar com grandes volumes de dados, o *PIG* fornece para os usuários uma linguagem de *script* que possibilita transformar conjuntos de dados sem exigir código mais complexo.

Um ponto importante nessa fase é a necessidade de instalação tanto do *PIG* quanto do *HIVE*, uma vez que estes não vêm dentro do *Hadoop Framework*. Após a instalação, basta realizar as consultas desejadas, baseando-se em uma linguagem *script*, como acontece com a linguagem *script* mais utilizada no mundo em diversos sistemas de gerenciamento de banco de dados, o SQL. Esta parte permite selecionar

somente aqueles dados necessários para que o *Hadoop* possa efetuar a atualização e o armazenamento no banco de dados.

4. Na etapa de armazenamento de dados, utilizou-se o *Hadoop Distributed File System* (HDFS), um sistema de baixo a médio custo que distribui arquivos para armazenamento de dados estruturados e não estruturados. O *Hadoop* inicia esse processo com a formação do *cluster* e salva os dados consultados na etapa anterior em um banco de dados local ou em nuvem. Esta etapa é automática, sem a necessidade de executar qualquer *script* ou linha de código.

Para finalizar, após os dados serem armazenados e codificados no *Hadoop*, deve-se realizar o procedimento de normalização dos dados, que permite que aqueles que possuem alto valor não sejam classificados como ótimos, como é o caso da criminalidade, que quanto maior o número de crimes, pior será. Utiliza-se um algoritmo escrito em *Python* para realizar esse processo automaticamente. A normalização é feita usando a seguinte equação (equação 2).

Equação 2 – Equação Processo de Normalização

$$normalValue = \frac{item}{maxValue} \quad (2)$$

Ao usar um simples algoritmo, consegue-se realizar esse processo automaticamente.

```
Require: maxValue (inputList)
1: larger = max(inputList)
2: for item in inputList do
3: normalValue = (item / larger)
4: outputList.append (normalValue)
5: outputList
```

Este processo de normalização possibilita uma transformação na escala de avaliação dos dados estruturados, passando a utilizar um intervalo entre 0 e 1. No caso do algoritmo proposto, seu funcionamento é simples, onde, inicialmente, identifica-se o maior valor da lista e este valor será associado a uma variável chamada

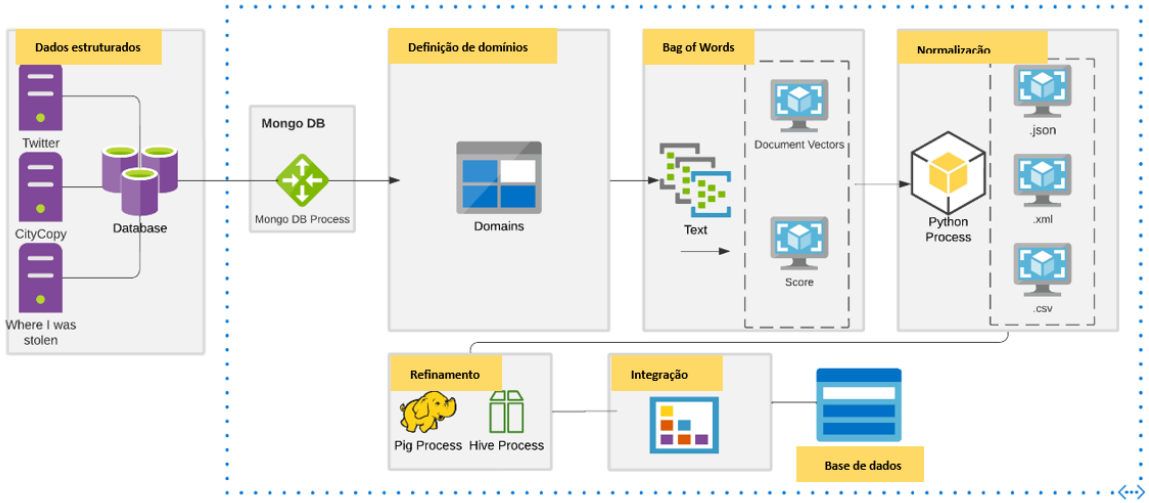
larger. Com um *for*, toda lista será percorrida e o valor do item será dividido pelo valor máximo, permitindo valores dentro de um intervalo numérico.

Observou-se que o processo de normalização se torna um ponto importante na padronização dos dados, pois ajuda a evitar que o algoritmo se torne enviesado para variáveis de maior ordem de grandeza.

3.2 *FRAMEWORK* PARA ANÁLISE DE DADOS NÃO ESTRUTURADOS

O próximo passo é analisar os dados não estruturados que podem vir de várias plataformas, como é o caso da plataforma *Onde Fui Roubado*, *CityCopy e Twitter*. Para tanto, tem-se um *framework* para análise deste tipo de dados (figura 9).

Figura 9– Framework para dados não estruturados



Fonte: Esta pesquisa (2022)

O processo de captura de dados não estruturados requer a definição de domínios, ou pontos estratégicos para a coleta de dados. Neste trabalho, cada comentário foi codificado em um único domínio: pessoas residentes na Região Metropolitana do Recife que foram vítimas de algum tipo de crime. Dados de outras

plataformas não exigiam definições de domínio, pois já eram dados do contexto de segurança pública.

O processo descrito acima acontece em três vertentes principais: o primeiro, relacionado ao *Twitter*. O *Twitter* disponibiliza sua API em versão para pesquisadores e estudantes que permite a identificação e a absorção de tweets, considerando critérios estabelecidos pelo usuário. Para essa extração é utilizada a linguagem de programação *Python* que possui uma biblioteca chamada *Tweepy*. Essa biblioteca permite acessar a API do *Twitter* e, consequentemente, acessar os dados do Twitter. Para isso o *Twitter* disponibiliza as formas de acesso, bem como as chaves correspondentes. A segunda vertente é relacionada ao *CityCopy* cuja a forma de absorção dos dados acontece a partir de arquivos *CSV*. A terceira vertente é o *Onde Fui Roubado* que também libera seus arquivos via *CSV*.

Para realizar a integração destas três fontes de dados em um único SGBD, que neste caso foi utilizado o Mongo DB por se tratar de um banco de dados não estruturados, a priori, utiliza-se um *software* de integração chamado *Pentaho Data Integration,* que funciona como uma ferramenta de Extração, Tranformação e Load – Carregamento (ETL) e permite a integração de várias bases de dados em um único repositório. Ele funciona a partir de transformações de bases e *jobs* (trabalhos), estes últimos possibilitam a integração final.

Após essa integração inicial entre as três instâncias de dados, o modelo *Bag of Words* (*BoW*) foi utilizado. *BoW* é um método de representar um documento como uma lista de palavras. Para realizar qualquer tipo de análise e recuperar informações de bases textuais em relação a bases estruturadas, deve-se reconhecer padrões encontrados no texto – mais especificamente, padrões nomeados, chamados entidades. Uma entidade é caracterizada por um padrão textual que está vinculado*, nomeado,* a uma determinada classe, como uma pessoa ou uma instituição. Para tanto, é usada uma base de entidades, ou seja, uma taxonomia ou ontologia do domínio de análise que representa as entidades de interesse que, possivelmente, estão presentes nos documentos coletados. Essa base precisa ser construída a partir de termos que tenham relevância potencial em relação ao domínio, o que pode exigir conhecimento prévio do domínio. A base pode ser construída manualmente, ou seja, os humanos inserem cada termo que consideram relevante, o que é um processo extremamente lento e caro; alternativamente, pode-se também construir essa base a partir da própria base de documentos, usando técnicas de reconhecimento de

entidades e/ou competição de termos. Com base nisso, 2.630 unigramas foram identificados após a remoção de palavras oclusivas do português, como o *the* e o *in*. Esta etapa pode ser executada via script, onde há definição direta dos domínios diretamente em qualquer sistema de gerenciamento de banco de dados ou a partir de algoritmos. Para esse trabalho houve a consulta direta via *Script SQL*.

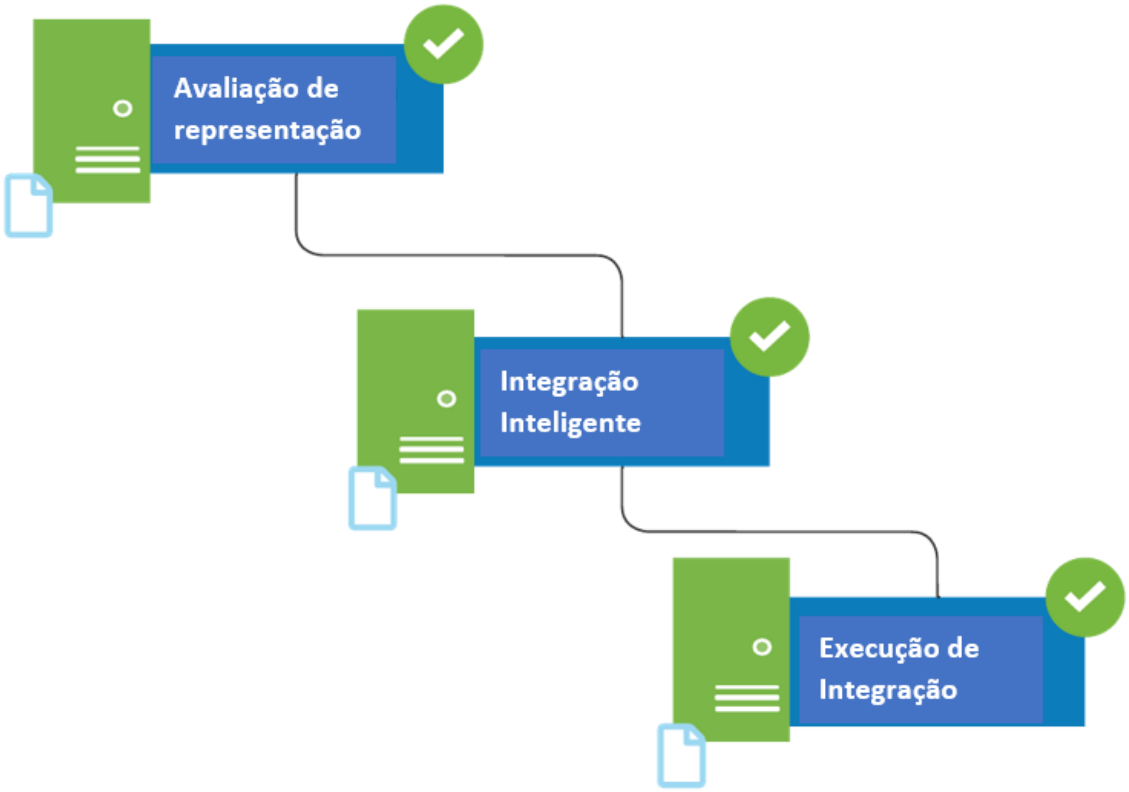
Como exemplo, para ilustrar esse contexto, a frase *Houve um assalto hoje perto de casa* pode ser minerada, pois contém a palavra *assalto,* que é um tipo de crime associado a esse contexto. O site captura geograficamente a localização do indivíduo, permitindo estabelecer a cidade, o bairro e a área de segurança integrada. Algumas informações não são capturadas, como nome e boletim de ocorrência, mas isso não afeta a análise final.

Por fim, realiza-se um procedimento de normalização, como com os dados estruturados. Esta etapa permite que arquivos estejam em formatos adequados para averiguação, ao mesmo tempo em que se refinam os dados para identificar erros. Este processo é igual ao refinamento de dados estruturados explicado anteriormente. Por fim, acontece a integração entre os dados estruturados e os dados não estruturados.

3.3 INTEGRAÇÃO DE DADOS

Este processo tem como finalidade permitir que os dados que são armazenados em diversos locais possam ser acessados, unificando a base de dados. Esta integração acontece através do *ETL*. O *ETL* é a sigla para *Extract, Transform e Load* e a mesma funciona para diversos tipos de arquiteturas, inclusive para a arquitetura de *Big Data.* Para este estudo foi desenvolvida uma estrutura de integração que é dividida em três partes principais (figura 10).

Figura 10 – Framework para integração dos dados



Fonte: Esta pesquisa (2022)

Essa estrutura de integração acontece a partir do *software Pentaho Data Integration*, já explicado anteriormente. Como há duas bases de dados, a ideia é unificá-las em uma única base de dados estruturada. Assim, as transformações realizadas pelo *software* necessitam passar por essas três fases principais.

Com relação a primeira etapa, há o processo chamado avaliador de representação. A finalidade desse avaliador é a identificação de dados, bem como *layouts* e padrões de arquivos que constam na base, considerando, também, os metadados de Sistemas de Gerenciamento de Banco de Dados (SGBDs). Esta etapa funciona da seguinte maneira: primeiramente há extração de metadados físicos de acordo com critérios previamente identificados; para esse trabalho dois critérios foram considerados: campos similares e/ou chaves de integração de dados e arquivos. Com isso, o modelo de representação mais adequado será direcionado para estes dados. Dentre estes modelos, a literatura apresenta o metadados de SGBDs relacionais; extratores de metadados por conteúdos estatísticos entre outros.

Para dados não estruturados tem-se extratores de metadados de bases federadas; extratores de metadados de SGBDs não relacionais; extratores inteligentes de texto e imagem. Estes extratores estão relacionados a um esquema de metadados chamado *Dublin Core* que tem como finalidade a descrição de objetos digitais, a exemplo de vídeos, áudios, texto e imagens.

Com relação a segunda etapa, tem-se um processo de otimização da fase anterior em que são identificadas as integrações que foram realizadas e o que é necessário para novos processos e melhorias de integração. Essa fase especifica cada chave de integração que combine os atributos da base de dados através do integrador de metadados. Alguns critérios são especificados, como é o cado da informação e sua disponibilidade, bem como o volume de dados do banco analisado. Por fim, a etapa de execução de integração faz a extração, transformação e carga de dados para a efetuação das junções dos dados em um único repositório de dados.

Após a integração final do banco de dados é preciso verificar a existência de ruídos. Os ruídos prejudicam o desempenho obtido por qualquer algoritmo de aprendizagem de máquina. Logo, removê-los melhora o esforço computacional. Neste sentido, o algoritmo *Decremental Reduction Optimization Procedures* é utilizado nesta metodologia, visando possibilitar a filtragem de ruídos de dados se estes forem classificados incorretamente por seus *k* vizinhos mais próximos e se sua remoção não prejudicar a classificação de outros dados. Essas regras corrigem o problema de remoção de agrupamentos inteiros e melhoram a generalização, com o custo de menor redução do conjunto de dados original, ideal para o cenário vivenciado neste trabalho. São levados em consideração valores fora do domínio; ausência de valores; inconsistência na forma de escrita e omissão de dados para minimizar o impacto negativo que estes ruídos possam representar nas informações extraídas, aumentando a relevância dos dados e melhorando a precisão dos algoritmos de aprendizagem de máquina.

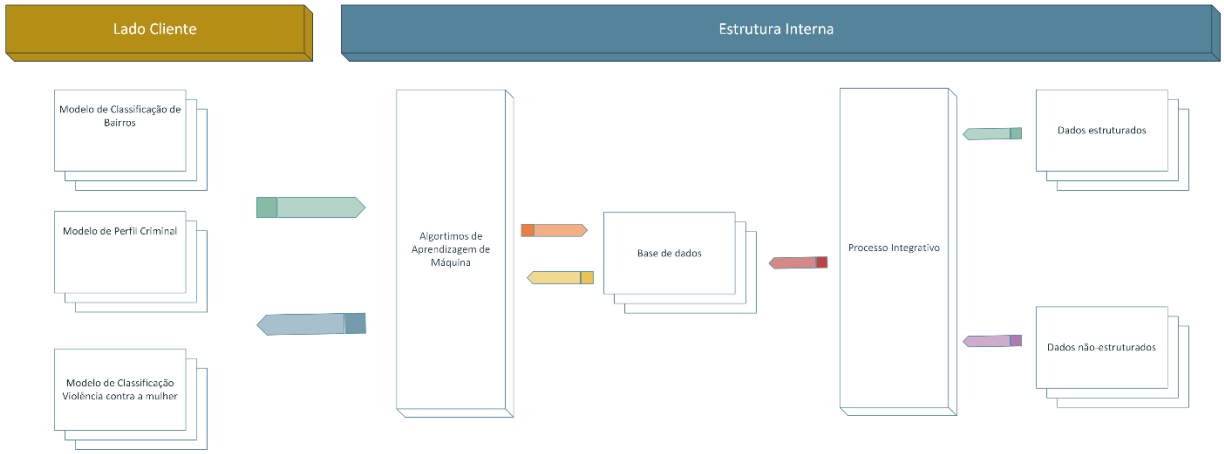
3.4 VISÃO GERAL DO *DS.SECURITY*

O *DS.Security* é um sistema de apoio à decisão para segurança pública que visa apoiar o processo decisório oferecendo ferramentas capazes de analisar grandes volumes de dados em tempo real, estando baseado na *web*, ou seja, disponível de forma simples para os usuários que precisam, bastando ter acesso à internet. O sistema, por sua vez, também permite a instalação local, desde que solicitado pelos departamentos

É um *software* registrado no Instituto Nacional da Propriedade Intelectual (INPI) sob o número **BR512020001467-4**).

A partir da figura 11, consegue-se compreender o funcionamento do sistema de apoio à decisão sob a perspectiva de seu ecossistema.

Figura 11– *DS.Security Ecossistema*



Fonte: Esta pesquisa (2022)

Onde:

Lado cliente: responsável pela interação sistema/homem. Neste momento o usuário poderá escolher qual modelo deseja entre os três disponíveis: modelo de análise de perfil; modelo de classificação de bairros; modelo de análise de violência contra a mulher.

Gerenciamento e análise do processamento de dados: o sistema de apoio à decisão concentra, aqui, todos os elementos favoráveis à análise de dados,

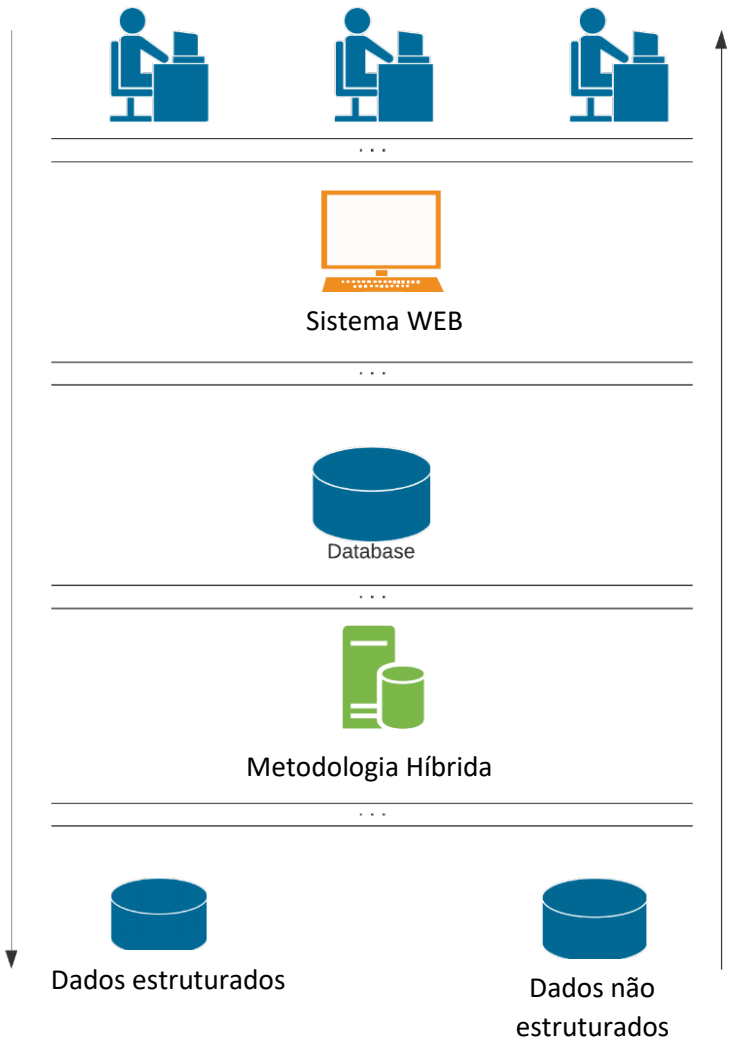
contemplando os algoritmos de aprendizagem de máquina, o repositório de dados e os *frameworks* para dados estruturados e dados não estruturados.

Base de dados: onde acontece a captura dos dados provenientes de várias plataformas e órgãos de segurança pública relacionados.

3.4.1 *DS.Security Back-End*

O *DS.Security* detém sistema embasado na metodologia híbrida mostrada no capítulo anterior, que funciona como intermediária entre o usuário final e a aquisição de dados (figura 12)

Figura 12– *Front View DS.Security*



Fonte: Esta pesquisa (2022)

O sistema atende a múltiplos usuários que inicialmente realizam a requisição do modelo que desejam operar. Após a escolha do modelo, imediatamente a metodologia híbrida de análise de dados entra em ação para analisar e incorporar dados estruturados e dados não estruturados que são adquiridos de múltiplas plataformas. Os modelos irão atuar neste repositório único de dados.

3.4.2 *DS.Security Front-End*

O *DS.Security* possui quatro telas principais: a primeira, que é composta pela tela de home; a segunda, terceira e quarta de acordo com os modelos de classificação de bairros, análise de perfil criminoso e violência contra a mulher, nesta ordem.

Figura 13 – Tela inicial *DS.Security*

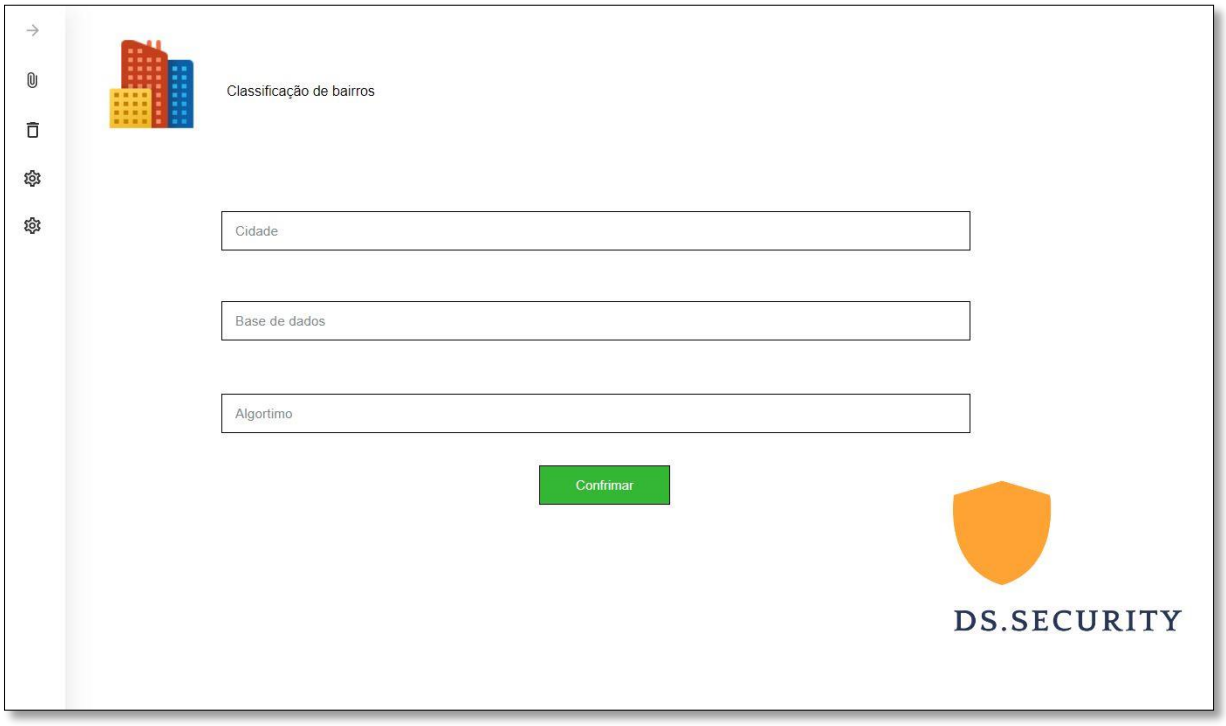
Fonte: Esta pesquisa (2022)

O usuário poderá escolher, de acordo com a tela inicial (figura 13), dentre os três modelos existentes, considerando o que se deseja. Os dados são inseridos apenas por administradores e o usuário comum não tem acesso.

Se o usuário optar pelo modelo de classificação de bairros, uma nova tela é exibida solicitando as seguintes informações: cidade em que se deseja realizar a classificação dos bairros, onde atualmente o *software* conta com a cidade do Recife. Além disso, o

usuário identificará a base de dados com a qual trabalhará entre as disponíveis. No fim, selecionará o algoritmo de aprendizagem de máquina desejado para a realização da análise (figura 14).

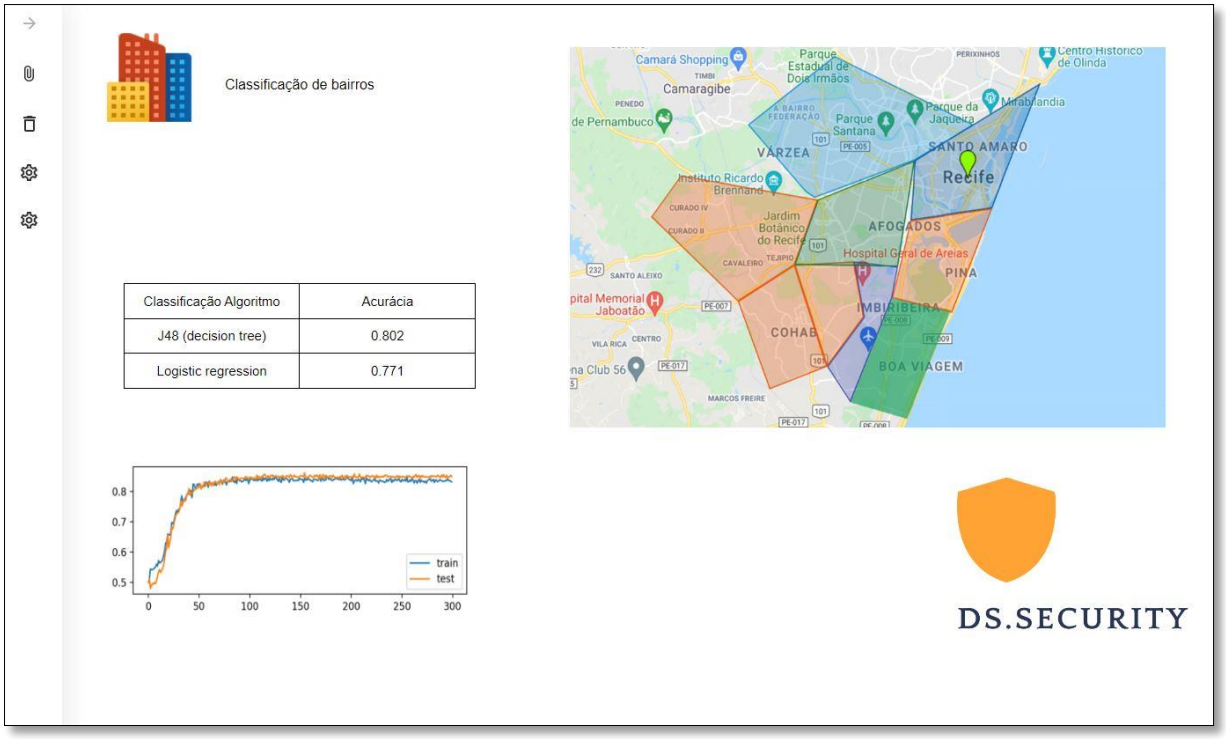
Figura 14 – Tela inicial do modelo de classificação de bairros



Fonte: Esta pesquisa (2022)

Após essa seleção inicial, o usuário é redirecionado para a segunda parte do sistema, onde haverá um *dashboard* informando a acurácia dos algoritmos selecionados, tanto a partir de uma tabela, quanto a partir de um gráfico de linha, bem como será mostrado um mapa com a classificação final dos bairros. Aqueles que estiverem em vermelho são considerados potenciais bairros para haver maior incidência de criminalidade; em azul vêm os bairros intermediários e, em verde, aqueles que obterão o menor número de criminalidades. A janela de tempo sempre é de uma semana, e estes resultados podem mudar diariamente (figura 15).

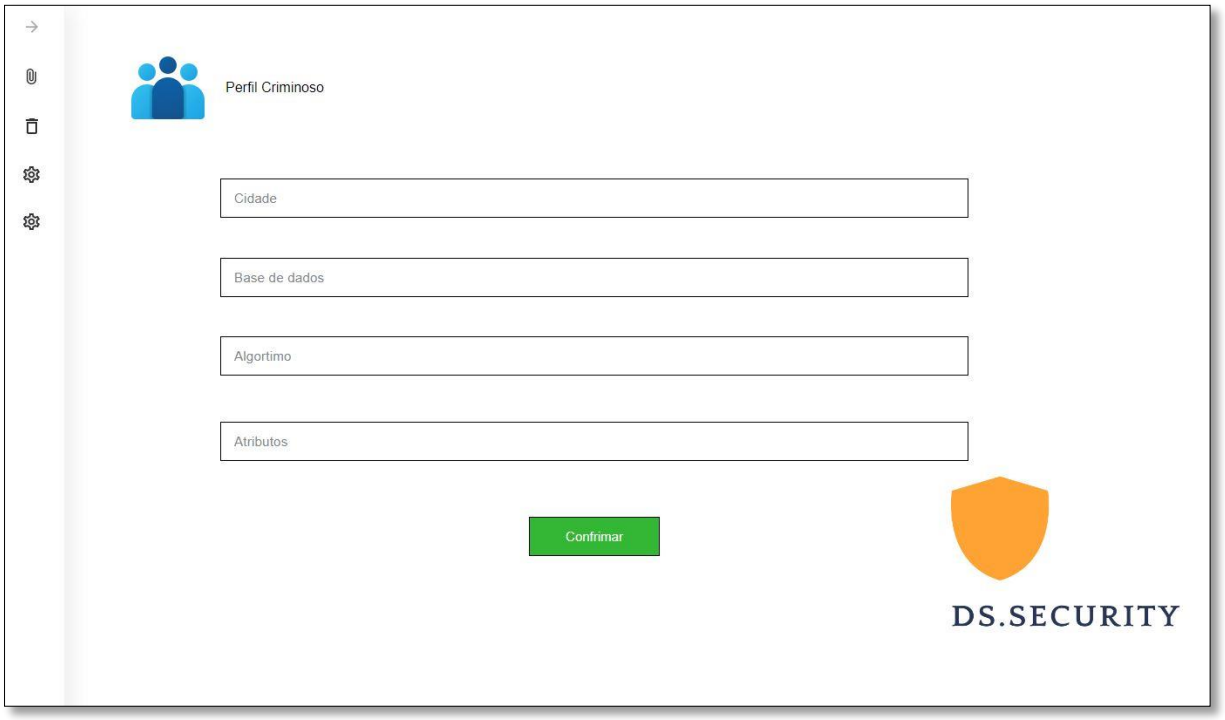
Figura 15 – Modelo de classificação de bairros



Fonte: Esta pesquisa (2022)

Se o usuário optar pelo modelo de perfil criminal, uma nova tela é exibida (figura 16) solicitando as seguintes informações: cidade em que deseja identificar o perfil criminoso, atualmente o software conta com a cidade de Recife. Além disso, o usuário poderá identificar com qual base de dados trabalhará entre as disponíveis e selecionar o algoritmo de aprendizagem de máquina desejado, bem como os atributos que serão levados em consideração pelo modelo.

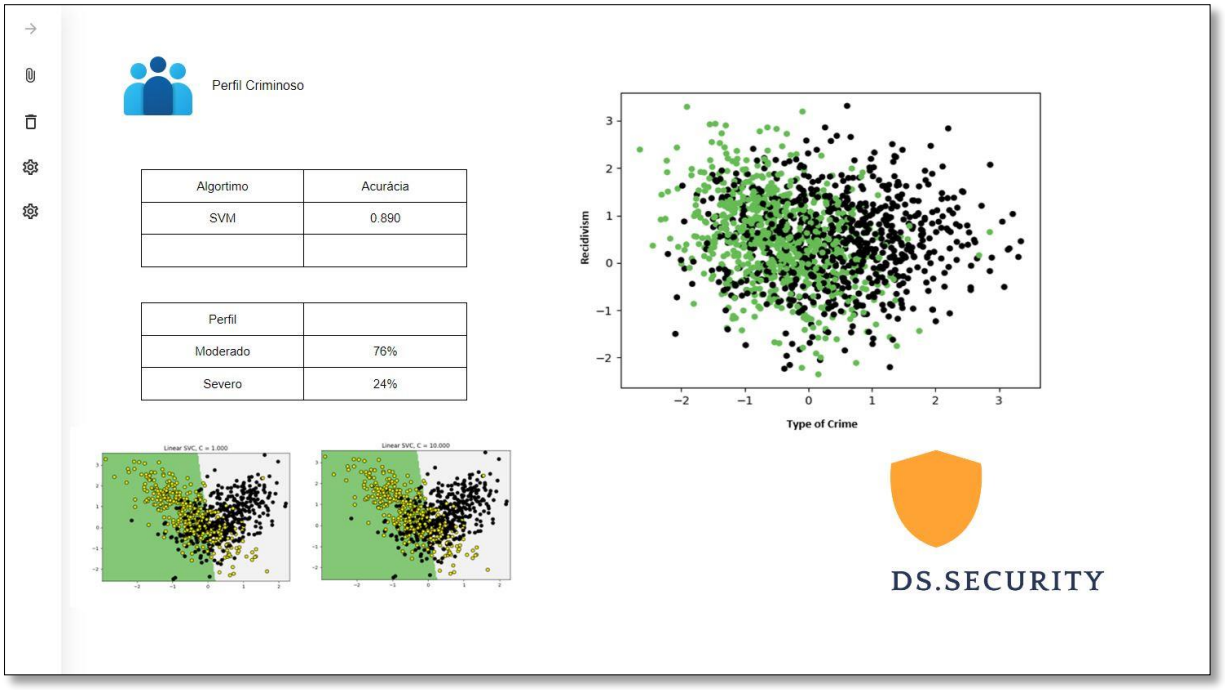
Figura 16 – Tela Inicial do modelo de perfil criminal



Fonte: Esta pesquisa (2022)

Após a seleção inicial, o usuário é redirecionado para a segunda parte do sistema, onde haverá um *dashboard* informando a acurácia dos algoritmos selecionados, em formato de tabela e gráfico, bem como será mostrada uma tabela informando a porcentagem de criminosos considerados moderados e severos. Gráficos são gerados para apresentar a dispersão dos dados (figura 17).

Figura 17 – Modelo de perfil criminal

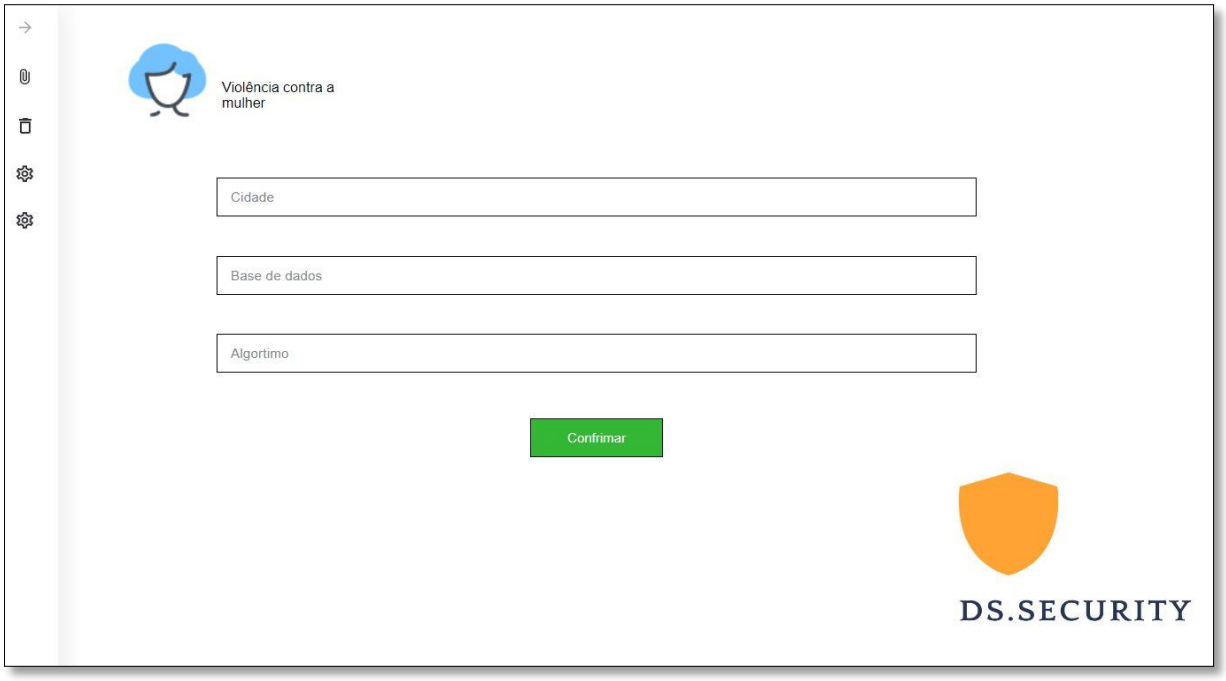


Fonte: Esta pesquisa (2022)

Por fim, se o usuário optar pelo modelo de classificação de bairros levando em consideração a violência contra a mulher (figura 18), uma nova tela é exibida para o usuário requerendo as seguintes informações: cidade em que se deseja realizar a classificação dos bairros, atualmente o software conta com a cidade de Recife. Além disso o usuário irá identificar com qual base de dados irá trabalhar dentre as disponíveis, bem como qual algoritmo de aprendizagem de máquina será utilizado para realizar essa classificação.

Além disso, o usuário poderá identificar com qual base de dados trabalhará entre as disponíveis e quais os atributos que serão considerados pelo modelo.

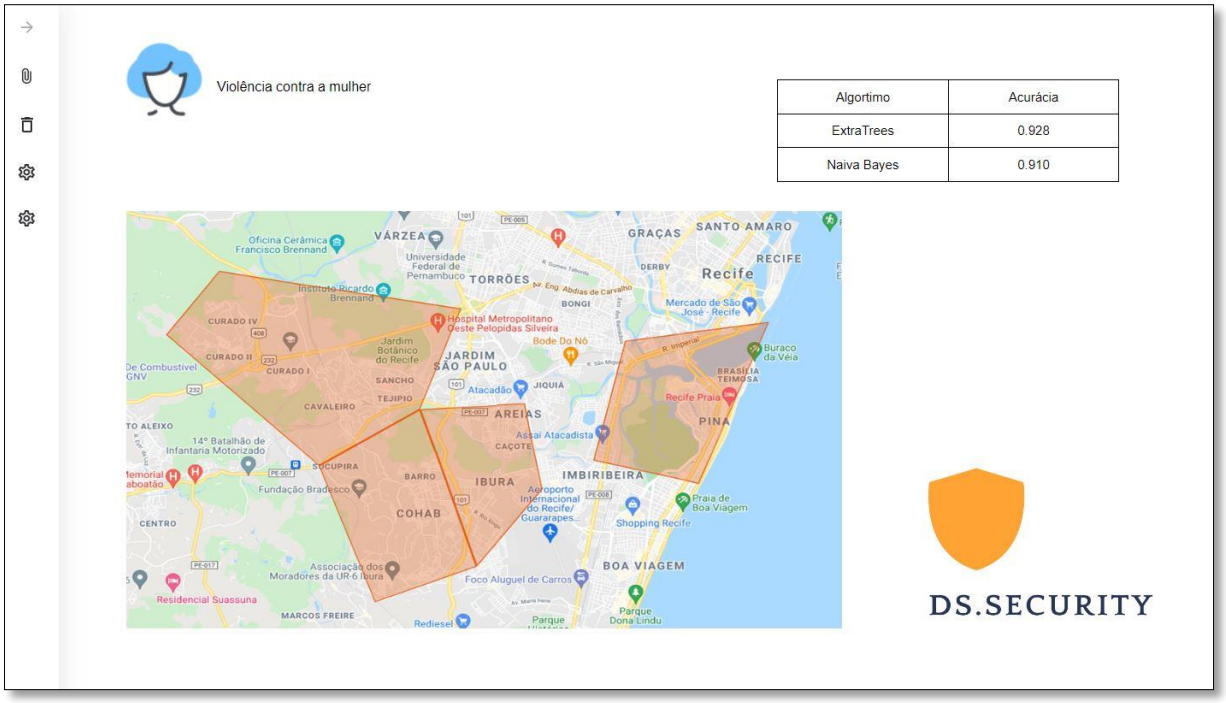
Figura 18 – Modelo de violência contra a mulher



Fonte: Esta pesquisa (2022)

Feita a seleção inicial, o usuário é redirecionado para a segunda parte do sistema, onde haverá um *dashboard* informando a acurácia dos algoritmos selecionados e como será retratado um mapa com a classificação final. Os bairros destacados em vermelho são considerados potenciais bairros para haver maior número de crimes contra a mulher, considerando uma janela de tempo de uma semana. Esses dados podem sofrer algum tipo de modificação dado que este é um sistema que trabalha em tempo real (figura 19).

Figura 19 – *Dashboard* modelo de violência contra a mulher



Fonte: Esta pesquisa (2022)

Cada modelo será retratado de forma mais aprofundada nos capítulos 4, 5 e 6 com estudos de casos para ilustrar a aplicabilidade de cada um. Estes capítulos ilustram aplicações da metodologia proposta em diferentes contextos em segurança pública. Estas aplicações são ilustrativas e tem como objetivo mostrar o aumento da acurácia de algoritmos de aprendizagem de máquina quando utiliza-se a combinação de dados estruturados e dados não estruturados (metodologia híbrida) na mesma análise em comparação com apenas dados estruturados (metodologia convencional).

Estas aplicações utilizaram-se de dados que estavam disponíveis no momento, mas em situações reais, precisariam ser estudados os dados que mais contribuem ou influenciam os objetos de cada aplicação. Neste sentido, outros dados poderiam ser adicionados ou até mesmo retirados dos modelos.

4 MODELO PARA CLASSIFICAÇÃO DE BAIRROS CONSIDERANDO A CRIMINALIDADE

Para este modelo, dois tipos de dados foram capturados: dados qualitativos não oficiais e dados quantitativos oficiais. Os dados não oficiais foram provenientes de redes sociais, nomeadamente o *Twitter* e duas outras plataformas *on-line*. A primeira dessas plataformas é *Onde Fui Roubado*, um sistema *on-line* colaborativo para mapeamento de ocorrências de furtos, roubos e outros tipos de crimes nas cidades brasileiras. Este serviço possibilita a visualização da incidência de crimes, em diversas localidades, com o objetivo de determinar as regiões mais perigosas. As denúncias são anônimas e podem ser consultadas por qualquer pessoa. O site segue os seguintes passos para a coleta de dados: 1) os usuários cadastram as informações de acordo com sua localidade por meio do aplicativo *Onde Fui Roubado*; 2) a informação passa a ser vinculada ao bairro em questão via georreferenciamento; e 3) a informação é divulgada para acesso público.

A segunda plataforma utilizada, o *CityCopy,* é de alerta social e foi criada com o objetivo de combater o crime. No *CityCopy*, os usuários podem avisar ativamente outras pessoas sobre crimes na comunidade e receber alertas em tempo real sobre o que está acontecendo em suas áreas de interesse. Os usuários geram o conteúdo no site, permitindo a análise em tempo real. Por fim, o *Twitter* foi utilizado para capturar dados da cidade que foi objeto de nosso estudo. Especificamente, foram coletados *tweets* de cidadãos que indicavam que um determinado incidente violento ou crime havia ocorrido na cidade e bairro do usuário.

Os dados oficiais utilizados neste trabalho foram provenientes da Secretaria de Defesa Social do estado de Pernambuco, Brasil. Esses dados foram captados com os órgãos envolvidos com a segurança pública nas áreas onde nossa metodologia seria implantada. Dados estruturados foram usados para contribuir para a classificação, previsão e análises de agrupamento. Os dados referentes aos números de homicídios, roubos e outros tipos de crimes são informados de acordo com o local onde ocorreram os crimes. Ambos os tipos de dados foram integrados em um único repositório usando uma metodologia híbrida identificada no capítulo 3.

Para ilustrar a aplicabilidade e o desempenho do modelo, foi estabelecido um cenário na cidade de Recife, Brasil. A Secretaria de Defesa Social tem dados

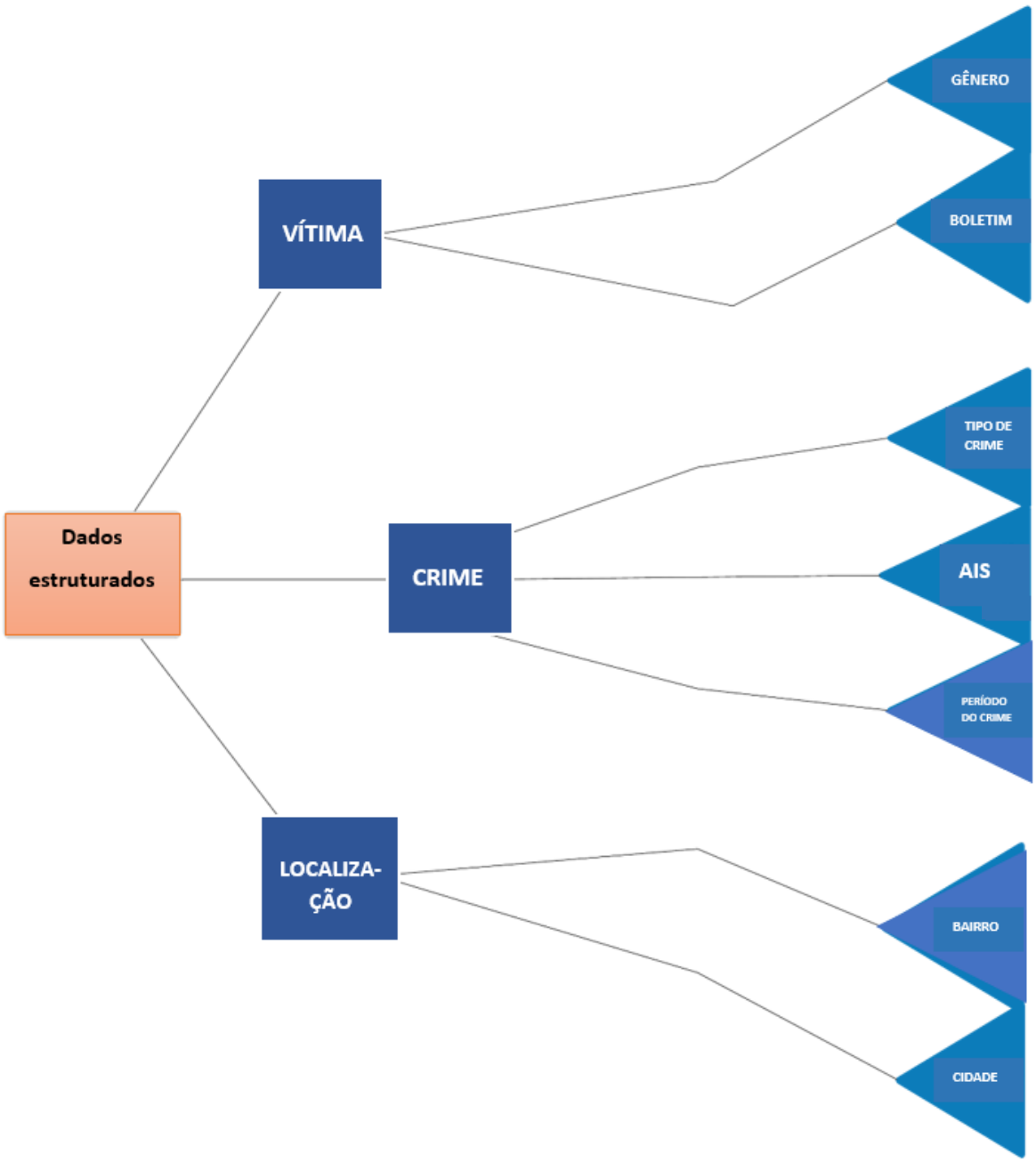
estruturados sobre a criminalidade na localidade. O departamento precisa conhecer em tempo real a situação dos bairros, considerando quais áreas são *hot spots,* ou seja, têm taxas de criminalidade mais altas, e quais são normais, ou seja, têm taxas de criminalidade mais baixas. Como essas informações estão em constante mudança, é preciso que as ações sejam coordenadas em tempo real para minimizar a ocorrência de crimes e aumentar a segurança na localidade. Portanto, a metodologia híbrida proposta neste trabalho é o primeiro ponto a contribuir para o processo decisório.

Os dados estruturados disponibilizados pela Secretaria de Defesa Social foram divididos em três pontos principais: vítima, crime e localização. A primeira categoria, dados relacionados à vítima, compreende dados que vão desde o sexo da vítima até o boletim de ocorrência do crime. O segundo, dados relacionados ao crime, inclui dados do tipo de crime, a área de segurança integrada e a hora em que ocorreu. O terceiro, dados relacionados à localização, contém informações como o bairro onde ocorreu o crime e as cidades envolvidas.

Em relação aos dados não estruturados, as informações mais relevantes para o estudo incluíram o tipo de crime cometido, o horário em que ocorreu, o bairro e cidade de execução. Estes atributos foram utilizados neste trabalho, uma vez que a inclusão desta informação assegura que a base de dados incorpore dados importantes para que os algoritmos *ML* possam realizar a classificação. *CityCopy*, *Onde Fui Roubado* e *Twitter* são utilizados para a captação destes dados.

Com isso, tem-se uma base final de dados com aproximadamente cinco milhões de dados, tendo como base os anos entre 2015 e 2021 (figura 20).

Figura 20 – Base de dados estruturados



Fonte: Esta pesquisa (2022)

4.1 CONFIGURAÇÃO DA BASE DE DADOS

A montagem da base de dados estruturados e dados não estruturados seguiram as etapas da metodologia híbrida proposta por esse trabalho. A configuração da base de dados segue o este mesmo padrão para os três estudos de casos, diferenciando, somente, no conteúdo dos dados.

Para a construção da base de dados estruturados houve, inicialmente, a necessidade do processo de refinamento da base. O refinamento permite que erros e espaços em branco sejam retirados da base, evitando que falhas no processo de análise e resultados. Entre estes erros, os mais comuns estavam associados com localidade. Por vezes, nomes de estados, como Pernambuco, eram escritos de várias maneiras, como Pernambuco, PE, Per, o que permite que cada erro desse seja associado com um estado diferente. Por outro lado, dados de crimes também não foram incluídos, associando o campo a um espaço em branco. Foi preciso, então, a remoção utilizando o algoritmo *Open Refine* previamente explicado no capítulo 3. Logo em seguida, torna-se necessário utilizar o *Hadoop Framework* devido à quantidade de dados que foi obtida. Para este estudo de caso foram coletados mais de três milhões de dados estruturados, o que necessita de um poder de processamento mais elevado. Desta forma, o *Haddop Framework* permite a inclusão de um *cluster* que possibilita a análise desse grande volume de dados. Para este *cluster* foram usados sete computadores, onde estes foram incorporados ao processo de divisão de tarefas para iniciar as análises dos dados. A partir desta clusterização pode-se realizar o Mapeamento e a Redução do Banco de Dados Estruturados.

O processo de mapeamento e redução possibilita identificar dados que poderiam ser associados, o que permitiria diminuir o banco de dados. No contexto deste trabalho, tipos iguais de crimes foram cometidos na mesma região, contudo, cada crime e sua localidade estavam em linhas diferentes do banco de dados. Este processo aumenta a necessidade de desempenho computacional, ao mesmo tempo em que dificulta a análise de dados. O quadro 1 representa essa estrutura.

Quadro 1– Exemplos de formação da base de dados

Tipo de crime	**ISA**	**Cidade**
Homicídio	5	Recife
Homicídio	5	Recife

Fonte: Esta pesquisa (2022)

No exemplo acima, verifica-se que há dois crimes (homicídios) que foram executados na mesma região (Recife). Logo, com o processo de mapeamento e redução da metodologia proposta neste trabalho, há simplificação da base de dados, como ilustra o quadro 2 abaixo.

Quadro 2 – Exemplos de formação da base de dados simplificada

Quantidade	**Tipo de crime**	**ISA**	**Cidade**
2	Homicídio	5	Recife

Fonte: Esta pesquisa (2022)

Após o processo, há inclusão de um novo atributo (quantidade) que identificará a quantidade de homicídios que houve em determinada região e reduzir a base de dados. O próprio mecanismo do *Haddop* realiza automaticamente esse processo e faz o armazenamento dos dados.

Quanto a montagem do banco de dados não estruturados para esse estudo de caso, foram absorvidos dados da plataforma *CityCopy*, *Onde Fui Roubado* e *Twitter*. Nestes dois primeiros, os dados já estão associados à criminalidade, logo basta selecionar a região desejada. No caso do *Twitter,* utiliza-se a API 2.0 da plataforma que permite limitar, buscar e recuperar *tweets,* considerando os domínios propostos pela metodologia híbrida. Para a conexão entre a API e as requisições de recuperação de *tweets*, um *scraper* foi desenvolvido em *Python*. Assim, *tweets* e textos são recuperados das três fontes, como por exemplo:

- Hoje, fui assaltado em frente de casa. Cuidado.
- No meu bairro teve dois homicídios na tarde de ontem.
- Não há como ter segurança na minha rua, ontem houve dois assaltos.

- Dupla de bandidos de moto assaltou um motoqueiro bem nesse cruzamento por volta de 6:30 de hoje.
- Pessoal de bicicleta. Rapaz jovem de boné praticando assaltos na região.
- Furto de celular!!!!! Agem em bando na redondeza.

Essas e tantas outras frases são divididas por palavras e associadas a um dicionário. Um dicionário possui palavras que permitem identificar, dentro de um texto/frase, os tipos de crimes cometidos. A título de exemplo:

Hoje, fui assaltado em frente de casa. Cuidado.

Nesta frase, verifica-se uma palavra que associa a um tipo de crime: *assaltado*. Esta é uma das palavras que consta no dicionário e permite a identificação pelo próprio, que passa a classificar este crime como um assalto. Tanto as frases, localidade, tipo de crime e horários são armazenados em um banco de dados não estruturado chamado Mongo DB.

4.2 ESTUDO DE CASO

Para avaliar o desempenho da metodologia proposta e classificar os bairros no cenário envolvido, primeiro foi necessário identificar a gravidade dos crimes. O Código Penal Brasileiro foi utilizado para realizar esta tarefa e uma vez que este código não se aplica a outros países, esta classificação deve, portanto, ser adaptada conforme necessário para cada nação.

O Código Penal Brasileiro faz distinção entre crimes instantâneos e crimes permanentes entre outros. Crimes instantâneos são aqueles cometidos em um determinado momento, ou seja, consumação imediata, sem qualquer prolongamento. O crime pode não necessariamente acontecer rapidamente; em vez disso, a consumação ocorre uma vez que seus elementos tenham sido reunidos. Um crime ser classificado como instantâneo não significa que o protagonista, ou seja, o

criminoso, não continue a se beneficiar dos efeitos do crime. Por exemplo, um roubo de carro é classificado como crime instantâneo, mesmo que o ladrão fique com o carro por vários dias, pois a consumação, ou seja, a subtração ou perda de valor, ocorreu no momento em que a violência, ameaça grave ou outros meios fez com que a vítima não resistisse. Outros exemplos de crimes instantâneos incluem assassinato, furto e roubo.

Crimes permanentes são aqueles em que a execução dura por um período de tempo determinado pelo sujeito ativo, ou seja, o criminoso. Um crime permanente ofende constantemente o bem jurídico e cessa conforme a vontade do agente. A extorsão por sequestro, por exemplo, é considerada um crime permanente. A relevância prática da determinação da permanência é estabelecer o início da contagem do prazo prescricional, que ocorre somente após a cessação da ofensa ao bem jurídico (Código Penal Brasileiro, 1988, artigo 111, inciso III).

Nosso estudo incorporou as partes relevantes do Código Penal Brasileiro e classificou os crimes em três intensidades, considerando se eram crimes instantâneos ou permanentes: baixo, médio e alto (tabela 1).

Tabela 1 – Crimes e suas categorias

Intensidade do crime	**Tipo de crime**
Baixo	Furto
Médio	Furto e roubo
Alto	Furto, roubo e homicídio

Fonte: Esta pesquisa (2022)

Roubo é o ato de retirar algo que pertence por lei a outra pessoa contra a vontade do legítimo proprietário, mas sem o uso de violência contra essa pessoa. O furto costuma ser praticado às escondidas para que o ladrão não seja notado (Código Penal Brasileiro, 1988). Os roubos consistem na subtração de algo que pertence a outra pessoa contra a vontade do legítimo proprietário em benefício do agente (criminoso) ou de terceiro com uso de violência ou grave ameaça ao legítimo proprietário do imóvel. Também são classificados como roubos os crimes em que o agente aplica violência ou atos de grave ameaça após a subtração dos bens para garantir a impunidade do crime (furto indevido). Homicídio é o ato de uma pessoa matar outra. Os homicídios podem ser divididos em várias subcategorias, como

infanticídio, eutanásia, pena de morte e legítima defesa, dependendo da circunstância em que ocorreu a morte.

Além desses dados, outros fatores devem ser levados em consideração para fazer as classificações, como a cidade em que o crime ocorreu, os bairros a ele associados, fatores sociodemográficos da vítima, por exemplo, sexo, escolaridade, renda e a área de segurança integrada da localidade específica. A composição desses dados mais os dados sobre os tipos de crimes e suas classificações são fundamentais para a identificação dos bairros.

4.2.1 Análise de desempenho dos algoritmos de aprendizagem de máquina neste cenário

Para realizar a classificação de bairros, foram utilizados algoritmos de *ML* supervisionados. Esses algoritmos tiveram aplicação no banco de dados obtido na utilização da metodologia híbrida. Esse banco de dados funcionou como um conjunto de treinamento para os algoritmos que geram classificações iniciais que podem ser modificadas (em tempo real) pela inclusão de outros dados.

Para avaliar o desempenho de cada algoritmo de *ML* após a integração dos dados, foi realizado um estudo comparativo de dois modelos principais. O primeiro modelo consiste em dados estruturados e o segundo era um composto de dados não estruturados e estruturados (ou seja, nossa metodologia proposta; tabela 2).

Tabela 2– Descrição dos modelos para identificação de taxa de acurácia

Modelos	**Descrição**
Modelo 1	Somente dados estruturados
Modelo 2	Dados estruturados e dados não estruturados

Fonte: Esta pesquisa (2022)

A F1-Score foi utilizada para medir o desempenho dos modelos utilizando validação cruzada de dez vezes. A pontuação F1 é uma métrica que combina precisão e *recall* para criar um número único que indica a qualidade geral de um modelo e funciona bem mesmo com conjuntos de dados que possuem classes desproporcionais. Neste trabalho, testamos três diferentes algoritmos de classificação

de *ML* para medir o desempenho do modelo: *Naive Bayes*; regressão logística; e J48, um classificador de Árvore de Decisão.

Os resultados de nossos testes de desempenho de classificação usando esses algoritmos de *ML* são mostrados na tabela 4.3. Dos dois modelos, o modelo 2, ou seja, que utiliza dados estruturados e não estruturados, obteve o melhor resultado no algoritmo J48 como classificador. No primeiro teste, o modelo 2 teve melhor desempenho com dois dos três algoritmos, enquanto o modelo 1 obteve melhor desempenho com o terceiro algoritmo. Uma comparação dos resultados do modelo 1 e do modelo 2, usando o classificador J48, mostra que o modelo 2 demonstrou melhor desempenho em mais de 10%. Essa diferença é estatisticamente significativa, com um valor de *p* menor que 0,001 usando o teste *t de Student.*

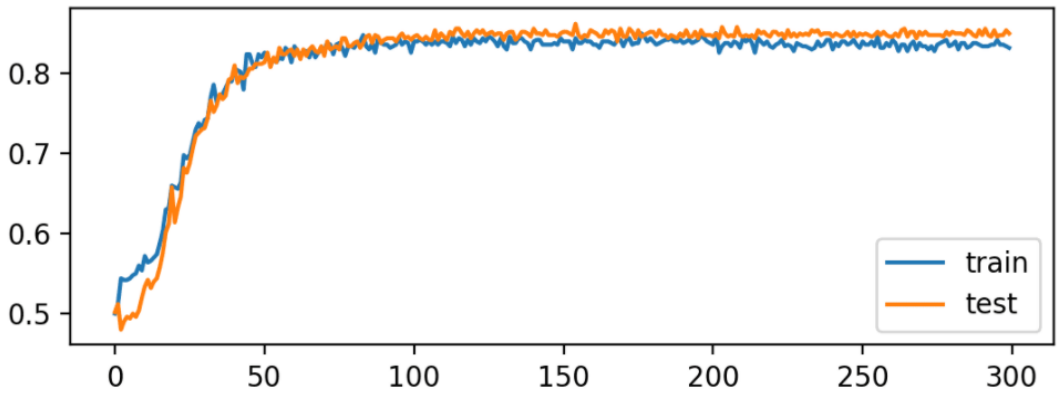
Tabela 3 – Taxa de acurácia dos modelos

Classificação do algoritmo	**Modelo 1**	**Modelo 2**
Naive Bayes	0.625	0.702
Logistic regression	0.732	0.771
J48 (decision tree)	0.721	**0.802**

Fonte: Esta pesquisa (2022)

Os resultados de acurácia para o algoritmo de classificação da Árvore de Decisão J.48, considerando o modelo 2, mostraram um resultado semelhante (figura 21).

Figura 21 - Desempenho do modelo



Fonte: Esta pesquisa (2022)

Para este resultado foi utilizada a seguinte equação (equação 3):

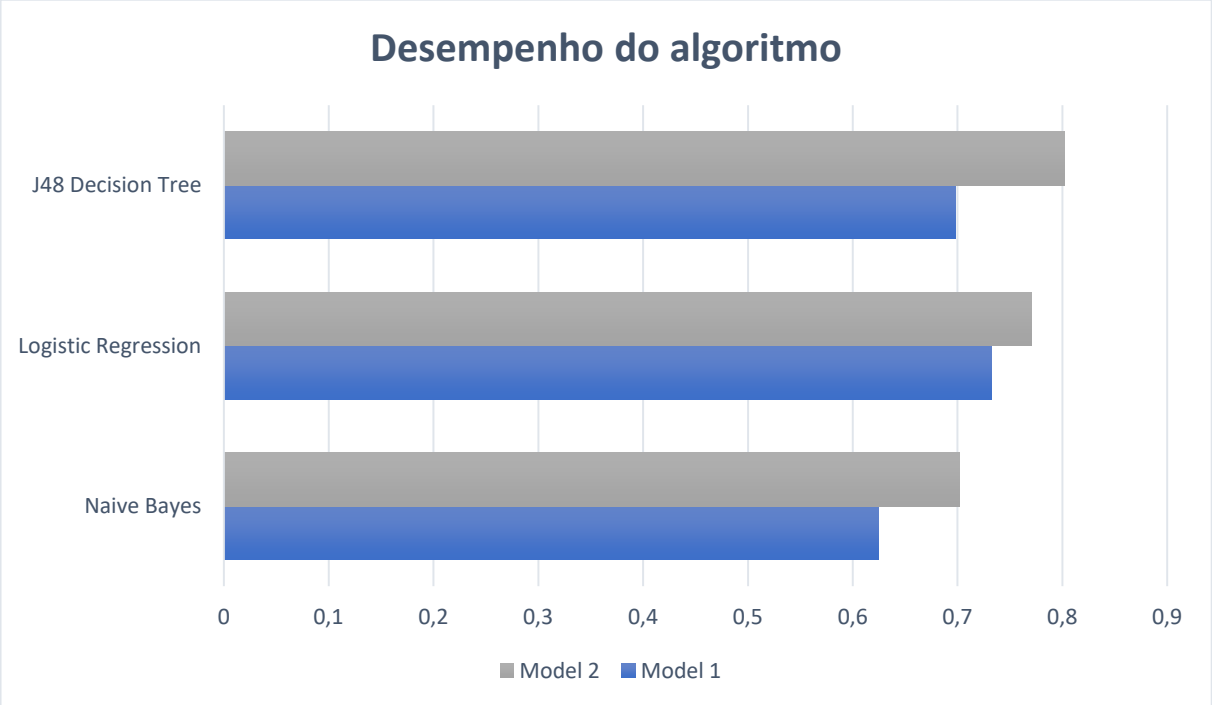
Equação 3 – Equação Acurácia Balanceada

$$Balanced\ Accuracy = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (3)$$

Essa fórmula realiza a divisão dos acertos pelos acertos e erros, onde *TP* é *true positive*; *TN* é *true negative*; *FN* é *false negative* e *FP false positive*. Consideramos a acurácia balanceada, uma vez que nesta não há influência do desbalanceamento de classes. Bibliotecas em *Python*, como é o caso do *SKlearn* podem realizar este cálculo de forma automática.

Com base nesses resultados, fica claro que combinar os dois tipos de dados em um único repositório é mais eficiente. Esta combinação permite aumentar o leque de possibilidades de aprendizagem que o algoritmo obterá (figura 22).

Figura 22 – Comparação entre algoritmos e modelos



Fonte: Esta pesquisa (2022)

Por outro lado, o teste F1 Score no algoritmo J48 obteve 0,810, resultado semelhante ao identificado pela acurácia balanceada, comprovando o bom desempenho do algoritmo dentro da base de dados concebida a partir da metodologia híbrida de análise de dados.

O modelo proposto por esse trabalho pode ser usado com outras técnicas de aprendizagem de máquina, como é o caso das redes neurais artificiais e florestas aleatórias, além de outros algoritmos de classificação, como é o caso do *ExtraTrees*. Ao considerar este cenário, o teste de acurácia revelou melhores resultados ao se utilizar o modelo 2 (tabela 4).

Tabela 4 – Taxa de acurácia dos modelos

Algoritmos	**Modelo 1**	**Modelo 2**
Redes Neurais	0.681	**0.720**
Florestas Aleatórias	0.715	**0.823**
ExtraTrees	0.705	**0.801**

Fonte: Esta pesquisa (2022)

Contudo, para esse estudo de caso foram considerados os algoritmos de aprendizagem de máquina *J48, Logistic Regression e Naive Bayes*, uma vez que o primeiro permite examinar dados de forma categórica e contínua; o segundo produz uma predição levando em consideração uma variável categórica; e o terceiro é um classificador probabilístico que assume que as *features* são independentes entre si. Desta forma, estes algoritmos possuem características que estão relacionadas ao problema enfrentado pelo departamento de segurança pública local.

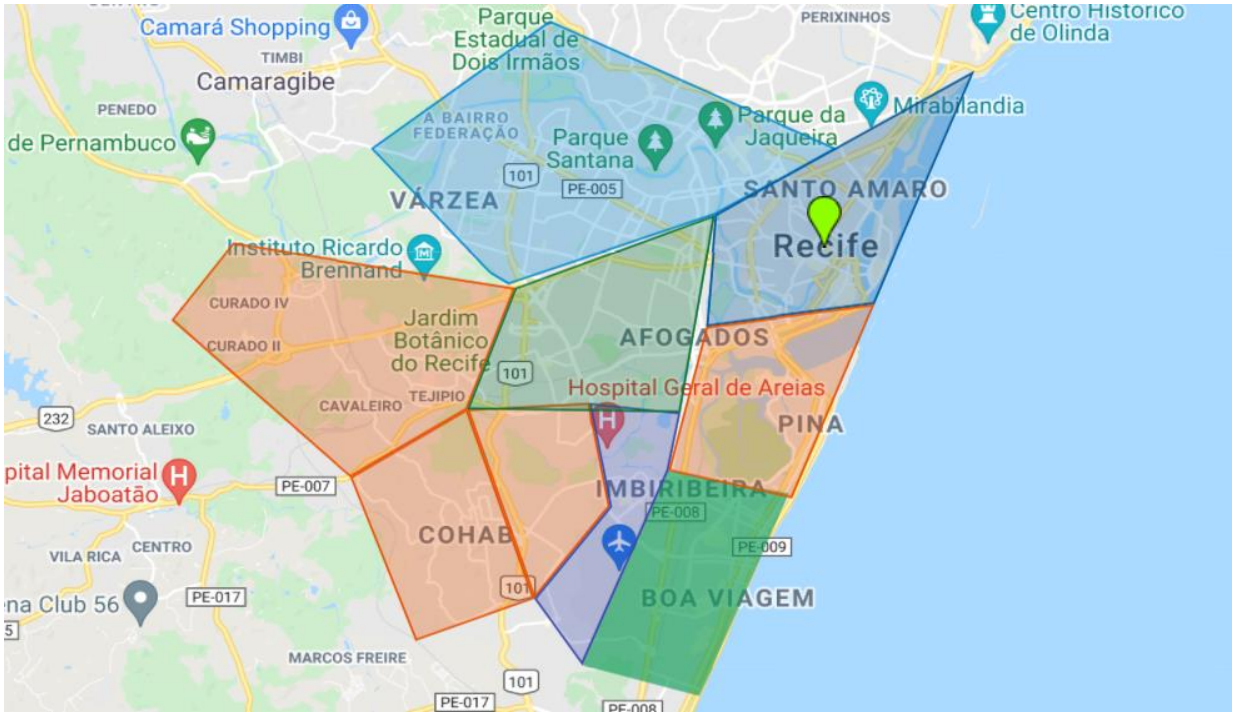
O teste F1 Score no algoritmo *Random Florest* obteve 0,815, resultado semelhante ao identificado pela acurácia balanceada, comprovando, também, o bom desempenho deste algoritmo na base de dados correspondente.

4.3 RESULTADOS E DISCUSSÕES

Ao considerar o cenário de classificação de bairros na cidade do Recife, o sistema providencia o mapa abaixo (figura 23). As áreas em vermelho representam os bairros com maior propensão de haver crimes de categoria elevada, as áreas em azul são os bairros com tendência de haver crimes de categoria intermediária e as

áreas em verde são os bairros com propensão de haver crimes com baixos níveis de gravidade.

Figura 23– Mapa com resultado final



Fonte: Esta pesquisa (2022)

Esse resultado pode mudar constantemente à medida que o banco de dados recebe novos dados, pois é um sistema que trabalha em tempo real. Os bairros classificados como de alta gravidade têm histórico de crimes de diversos tipos, incluindo roubos, furtos e homicídios.

O departamento de segurança pública pode usar esses resultados para tomar as melhores decisões. Iniciativas como aumentar o número de policiais, de viaturas e programar o uso de viaturas em regiões de alta gravidade estão entre as que são visíveis e devem contribuir para melhorias na segurança pública nestas regiões. Esta metodologia contribui, portanto, diariamente, como suporte ao processo de tomada de decisão.

5 CLASSIFICAÇÃO DO NÍVEL CRIMINAL EM SEGURANÇA PÚBLICA

Para este modelo, o algoritmo *Suport Vector Machine (SVM)* passa a ser utilizado. Como dito no capítulo 2, o *SVM* é um método de aprendizado de máquina que tenta obter dados de entrada e classificá-los em uma das duas categorias. Um *SVM* requer o uso de um conjunto de dados de treinamento de entrada e saída para construir o modelo *SVM* que pode ser utilizado para classificar novos dados. A classificação ocorrerá em duas esferas: presos com perfil moderado e presos com perfil severo.

5.1 CONFIGURAÇÃO DA BASE DE DADOS

No que se refere, especificamente, à montagem da base de dados para esse estudo, a princípio os dados estruturados foram coletados e refinados, baseando-se na metodologia híbrida de análise de dados, que é processo idêntico ao estudo de caso apresentado no capítulo 4. A diferença, neste estudo, são os conteúdos dos dados. Os dados prisionais têm em conta a localidade em que ocorreu o nascimento, tipo de crime que cometeu, fatores educacionais, reincidência e fatores sociais. Estes dados são sensíveis e protegidos pelos órgãos envolvidos em segurança pública. O processo de refinamento, mapeamento/redução e armazenamento segue as etapas da metodologia híbrida de análise de dados proposta por esse trabalho.

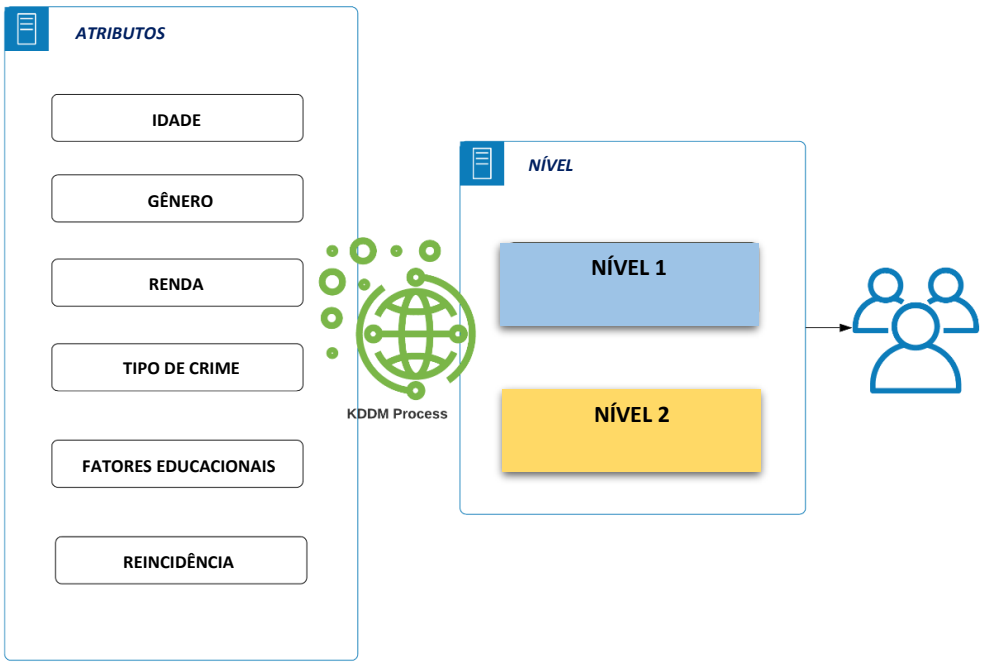
Para os dados não estruturados, também foram selecionados dados da plataforma *Onde Fui Roubado, CityCopy* e *Twitter*. Os dados são focados nos tipos de crimes e os fatores sociais que estão envolvidos nas localidades, permitindo que algoritmos de aprendizagem de máquina treinem e possam identificar os resultados a partir de uma base de testes. O processo de configuração desta base, bem como a integração entre as bases, acontece da mesma forma que à do estudo de caso apresentado no capítulo anterior.

5.2 ESTUDO DE CASO

Ao considerar as características do algoritmo, torna-se necessário identificar os atributos para que o algoritmo de *SVM* possa realizar a classificação. Essa identificação acontece a partir do nosso repositório geral de dados. Para este modelo é utilizado um repositório de população carcerária e crimes de CVLI, levando em consideração sua localidade, bem como crimes que acometem a população diariamente (dados não estruturados). Estes dados passam pelo processo de integração, como discutido no capítulo 3, e, por fim, tem-se um único repositório de dados, com diversos atributos.

Torna-se importante medir e identificar os atributos mais relevantes para o problema em questão. Sete atributos e seus relacionamentos se mostraram mais relevantes para a classificação desses indivíduos, pois a combinação de dados relacionados a esses atributos contribui para o aprendizado dos algoritmos *SVM* (figura 24).

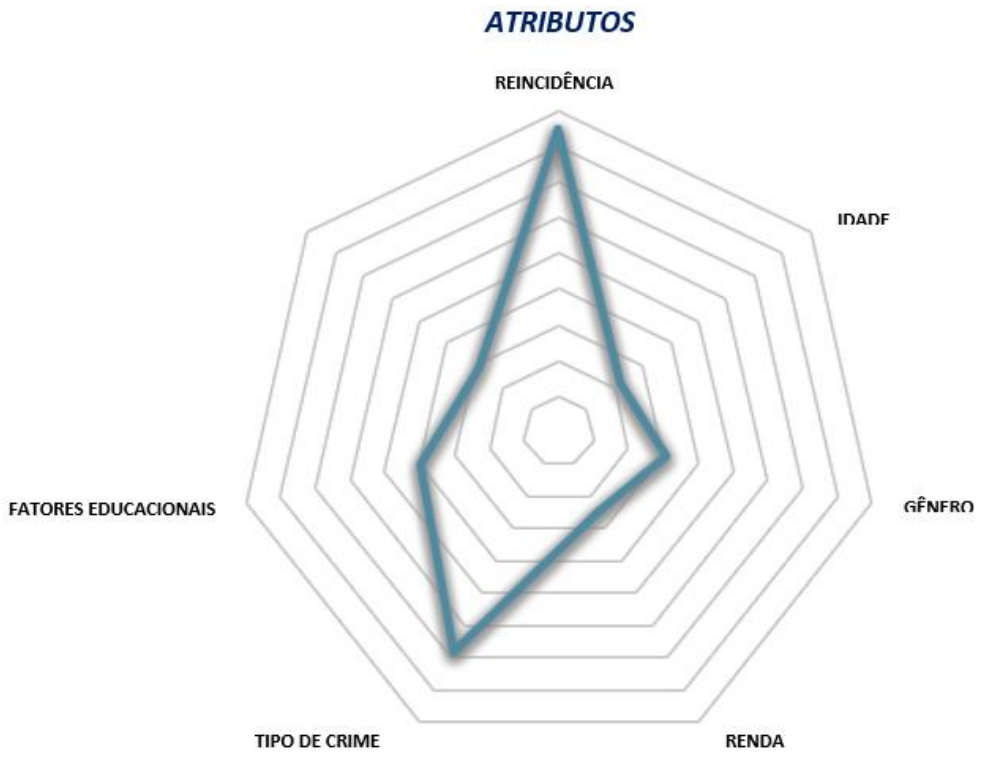
Figura 24 – Atributos para análise de dados



Fonte: Esta pesquisa (2022)

Baseando-se nesse processo, dois atributos foram selecionados e determinados como ideais para o processo de classificação (figura 25).

Figura 25 – Desempenho dos atributos



Fonte: Esta pesquisa (2022)

Assim, a reincidência e a criminalidade obtiveram os melhores índices considerando os dados disponíveis. Os dados têm tendências e comportamentos que, quando alinhados, permitem conhecê-los melhor.

5.2.1 Nível de crimes

Dois tipos de níveis foram estabelecidos para este estudo de caso: nível 1 e nível 2. O perfil do crime é definido como uma técnica de investigação da cena do crime utilizada para analisar padrões de comportamento que melhor definem um crime violento ou uma série de crimes que podem estar associados com o propósito de identificar as características do suposto infrator. Essa técnica integra processos de coleta e análise de uma cena de crime para prever o comportamento, características

de personalidade e indicadores sociodemográficos de infratores que cometeram crimes semelhantes, estreitando o campo de suspeitos e auxiliando em suas prisões.

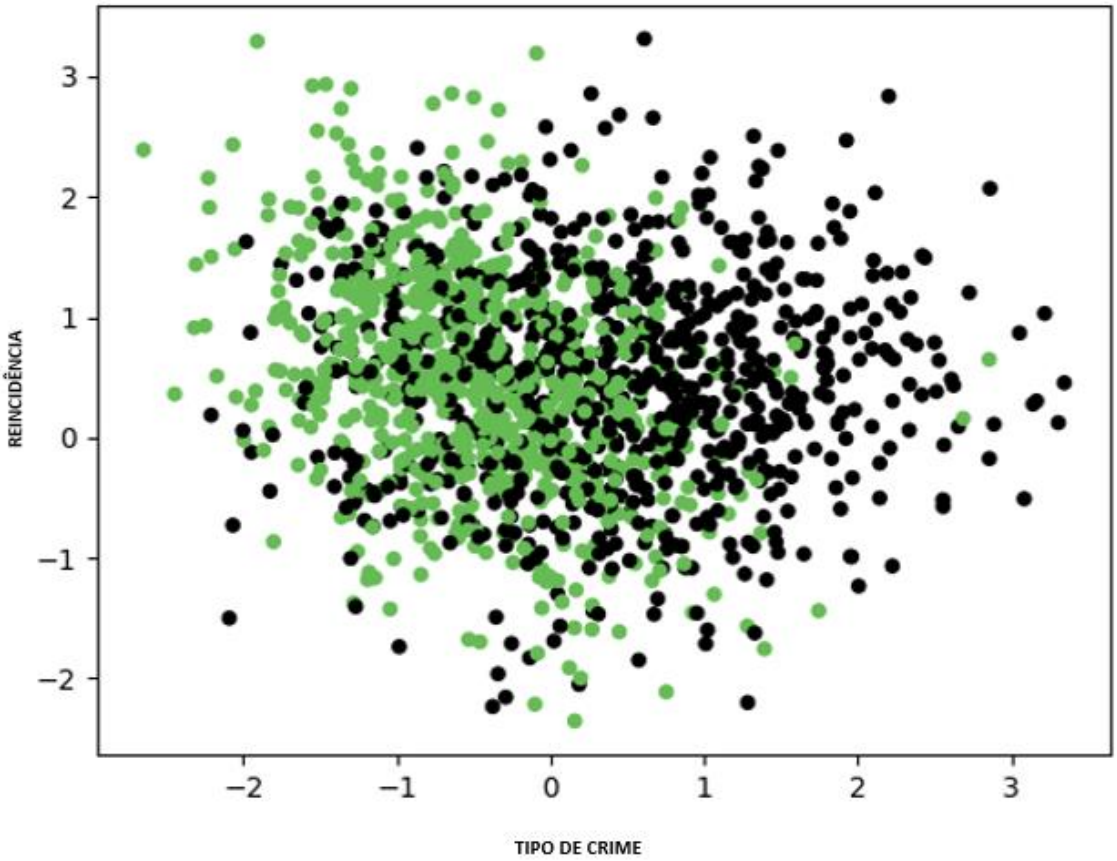
Os níveis criminais foram divididos em categorias nível 1 e nível 2:

- No **nível 1**, o indivíduo cometeu crimes menores, como roubo e furto. Além disso, fatores demográficos impulsionam esse nível, incluindo ensino fundamental ou médio, nascimento em um bairro não violento e ausência de reincidência.
- As pessoas que se enquadram no **nível 2** cometeram crimes considerados mais graves, como roubos e homicídios. Os fatores demográficos que impulsionam esse nível incluem baixa escolaridade, nascimento em bairros violentos e familiares, que cometeram crimes.

5.2.2 Aplicação do algoritmo de *SVM*

O algoritmo *SVM* pode ser executado após o processo de classificação ser estabelecido. Nesse caso, foram considerados dois atributos, reincidência e tipo de crime, com duas classificações possíveis: nível 1 e nível 2. A primeira etapa envolve o treinamento *SVM*, neste banco de dados, para identificar padrões nos dados. A ideia é verificar as características em comum entre os atributos do crime e da reincidência nos indivíduos (figura 26).

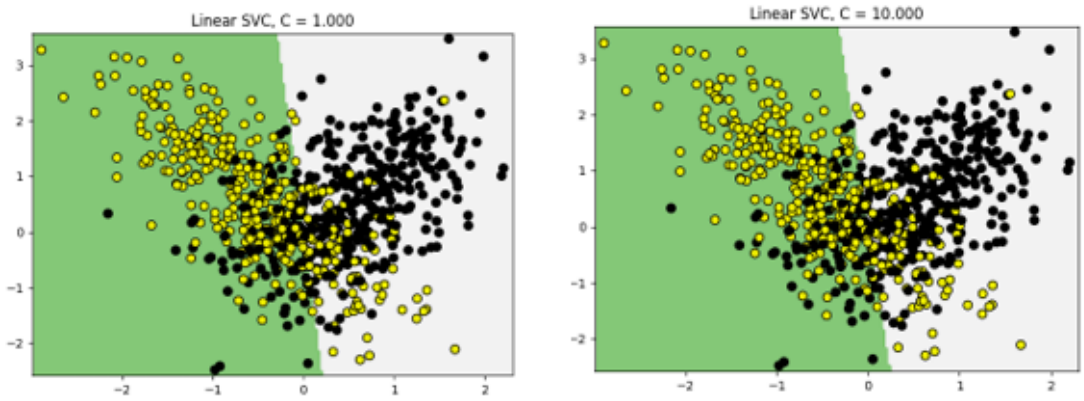
Figura 26 – Padrões de dados



Fonte: Esta pesquisa (2022)

É possível, assim, realizar a classificação geral dos dados a partir do parâmetro que controla a complexidade do modelo. Pode-se observar que à medida que o parâmetro C se torna maior, há uma fronteira que tenta não permitir erros de classificação. Os pontos pretos na figura 9 indicam pontos de dados que pertencem à categoria grave e os círculos amarelos representam pontos de dados que pertencem à categoria moderada. Cada um dos pontos de dados individuais tem um valor de entrada exclusivo 1, representado por sua posição no eixo x, e um valor de entrada exclusivo 2, representado por sua posição no eixo y. Todos esses pontos foram mapeados no espaço bidimensional (figura 27).

Figura 27 – Processo de classificação (parâmetro C)



Fonte: Esta pesquisa (2022)

Um *SVM* pode classificar dados criando um modelo desses pontos no espaço bidimensional. O *SVM* observa dados no espaço bidimensional e usa um algoritmo de regressão para encontrar um hiperplano unidimensional, também conhecido como linha, que separa os dados com mais precisão entre suas duas categorias. O *SVM*, então, usa essa linha divisória para classificar novos pontos de dados na categoria 1 ou na categoria 2.

Quanto ao desempenho do algoritmo considerando duas métricas, F1 Score e Acurácia Balanceada, verificar mais detalhes no capítulo 4, observa-se bom desempenho. Para efeito comparativo há dois modelos: no primeiro há aplicação do algoritmo *SVM* somente em dados estruturados e no segundo modelo aplica-se o algoritmo de *SVM* no banco de dados oriundo da metodologia híbrida de análise de dados. Este é o resultado a partir do F1 Score (tabela 5):

Tabela 5 – Taxa de F1 Score dos modelos para perfil criminal

Algoritmos	**Modelo 1**	**Modelo 2**
Support Vector Machine	0.665	**0.810**

Fonte: Esta pesquisa (2022)

Com relação ao teste de acurácia balanceada, obteve-se (tabela 6):

Tabela 6 - Taxa de acurácia balanceada dos modelos para perfil criminal

Algoritmos	Modelo 1	Modelo 2
Support Vector Machine	0.609	**0.876**

Fonte: Esta pesquisa (2022)

5.3 RESULTADOS E DISCUSSÕES

Para este estudo de caso, cerca de 76% dos indivíduos criminosos foram classificados como nível 2 e 24% foram classificados como nível 1. Esses resultados consideram os atributos identificados a partir do banco de dados que, segundo o modelo, são atributos relevantes para identificar e mensurar a classificação desses indivíduos.

O algoritmo *SVM* alocou 76% dos indivíduos em nível 2. Observa-se que alguns fatores impactam esse resultado mais do que outros. Os algoritmos *SVM* sempre consideram a relação entre os atributos e seus respectivos pesos. A definição dos atributos, realizada inicialmente pela extração de conhecimento da base de dados, é um fator decisivo para o nível de precisão do algoritmo.

Para este resultado, dois atributos e seus relacionamentos direcionaram o resultado do algoritmo. O primeiro tem relação ao tipo de crime que o indivíduo cometeu. Esse tipo de atributo permite entender a gravidade do crime e como isso está relacionado a fatores demográficos, como o bairro onde o crime foi executado. O segundo critério verifica se o indivíduo é reincidente, identificando se já houve o retorno ao processo carcerário.

Essas informações destinam-se a ajudar os investigadores a conhecer as características de sujeitos criminais desconhecidos, criar uma lista de suspeitos ou restringir um grupo de suspeitos, desenvolver perfis e realizar análises motivacionais. Além disso, essa informação busca identificar o infrator e serve para consulta, auxiliando detetives e demais investigadores na solução de casos e atores judiciais em processos decisórios.

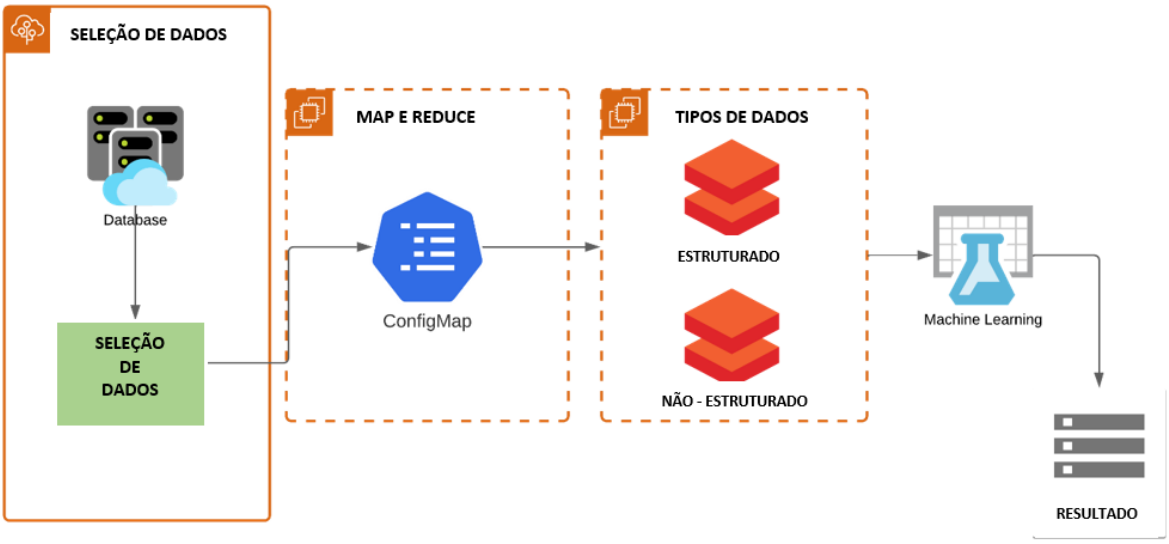
Por fim, esse modelo pode ser utilizado pelos componentes policial e de segurança pública, em conjunto com dados de presos e criminosos, para identificar os suspeitos e a população carcerária de acordo com seu nível. As políticas de segurança pública tendem a ser mais eficientes porque as ações serão direcionadas

a esses indivíduos, bairros e populações carcerárias para evitar possíveis reincidências.

6 CLASSIFICAÇÃO DE BAIRROS CONSIDERANDO VIOLÊNCIA CONTRA A MULHER

Este é o terceiro modelo incorporado ao *DS.Security*. Trata-se de um modelo que tem como finalidade classificar bairros considerando a violência contra a mulher. Apesar de classificar estes bairros, este modelo apresenta uma característica preditiva, pois identificará quais bairros são potenciais candidatos a ter violência contra a mulher. Para isso, o modelo utiliza a metodologia híbrida de análise de dados, incorporando dados de criminalidade em âmbito geral e dados de criminalidade contra a mulher. Abaixo, ilustração do funcionamento desse modelo (figura 28).

Figura 28 - Estrutura do modelo de violência contra a mulher



Fonte: Esta pesquisa (2022)

1 - Seleção de Dados: verifica-se a base de dados associada, bem como os tipos de dados existentes. É neste momento que há uma pré-análise dos dados para identificar dados estruturados e não estruturados.

O processo para seleção e configuração de dados baseia-se na metodologia híbrida de análise de dados proposta por esse trabalho. No primeiro momento, há captura de dados estruturados e no segundo, de dados não estruturados. As configurações seguem a mesmas instruções observadas no estudo de caso do

capítulo 4, na sessão 4.1, mas com o foco em dados sobre violência contra a mulher.
2 - Processo *Map Reduce*: a função do *Map Reduce* é buscar padrões e reduzir o número de linhas para otimizar o banco de dados.
3 - Dados estruturados e não estruturados: para cada bloco de dados existe uma forma de analisá-los. No caso de dados estruturados, pode-se usar ferramentas convencionais de extração e análise, como *scripts* de medição de dados padrão. No caso de dados não estruturados, é estabelecer uma sequência lógica de análise, desde a identificação dos padrões, até a sua transformação em dados estruturados, utilizando-se a metodologia híbrida do *DS.Security* para analisá-los e obter um único repositório de dados.
4 – Aprendizado de Máquina: a partir da identificação e do armazenamento dos dados, pode-se associar algoritmos de aprendizado de máquina que terão como objetivo realizar a classificação de bairros tendo como base os crimes contra a mulher.

Com o objetivo de extrair conhecimento dos bancos de dados envolvidos na segurança pública, considerando a violência contra a mulher, foi realizado um estudo de caso no estado de Pernambuco.

Os dados incluídos na análise foram (considerando dados estruturados): tipo de crime na localidade; quantidade de crimes envolvidos; ações realizadas no bairro; e fatores educacionais em relação aos dados não estruturados, foram também utilizados dados de redes sociais, como o *Twitter* e de duas plataformas *on-line*. A primeira, *Onde Fui Roubado* e a segunda, *CityCopy* (para saber mais detalhes, ler o capítulo 5). Esta análise tem como objetivo saber como a população reage, tanto ao crime quanto às políticas estabelecidas pelo governo para minimizar o impacto do crime na cidade.

Com relação a classificação, basicamente o sistema identifica o bairro com maior incidência de violência contra a mulher. Assim, automaticamente o bairro que obtiver maior incidência será destacado no mapa com a cor vermelha.

Para esse estudo foram usados algoritmos de aprendizagem de máquina supervisionados, como é o caso das redes neurais artificiais, florestas aleatórias e *ExtraTrees.* Ao realizar um teste de acurácia, foi observada a seguinte classificação (tabela 7):

Tabela 7 – Teste de acurácia

Algoritmos	Acurácia
Redes Neurais	0.791
Florestas Aleatórias	0.805
ExtraTrees	0.885

Fonte: Esta pesquisa (2022)

Observa-se que o algoritmo *ExtraTrees* obteve bom desempenho no teste de acurácia e este algoritmo será o selecionado para a classificação final. A classificação final ocorre levando em consideração os quatros bairros com maior chance de haver violência contra a mulher em uma semana.

Para efeito de comparação, fizemos uma análise envolvendo apenas dados estruturados, não utilizando a metodologia híbrida. Os testes mostram que a metodologia híbrida traz melhores valores. Como resultado, tem-se (tabela 8):

Tabela 8 – Teste de acurácia não utilizando metodologia híbrida

Algoritmos	Acurácia	F1 - Score
Redes Neurais	0.691	0.702
Florestas Aleatórias	0.735	0.710
ExtraTrees	0.745	0.723

Fonte: Esta pesquisa (2022)

Este cenário estabelece a necessidade de compreensão que a violência contra a mulher não precisamente significa alguma violência física, também estão incluídos qual ato ou violência de gênero que tem como resultado principal dano ou sofrimento físico, sexual ou psicológico para a mulher e isto inclui tanto ameaças, coerção ou privação arbitrária de liberdade.

A violência física pode incluir tapas, empurrões, socos, chutar, torcer os braços, dentre outros crimes. O abuso psicológico e emocional pode incluir uma série de comportamentos controladores, como controle das finanças, isolamento da família e amigos, humilhação contínua, ameaças contra crianças ou ameaças de ferimentos ou morte. O abuso financeiro ou econômico inclui o controle forçado do dinheiro ou outros ativos de outra pessoa. Também pode envolver roubar dinheiro, não permitir que a vítima participe de quaisquer decisões financeiras ou impedi-la de ter um emprego.

De modo geral, os crimes contra a mulher podem ser (quadro 3):

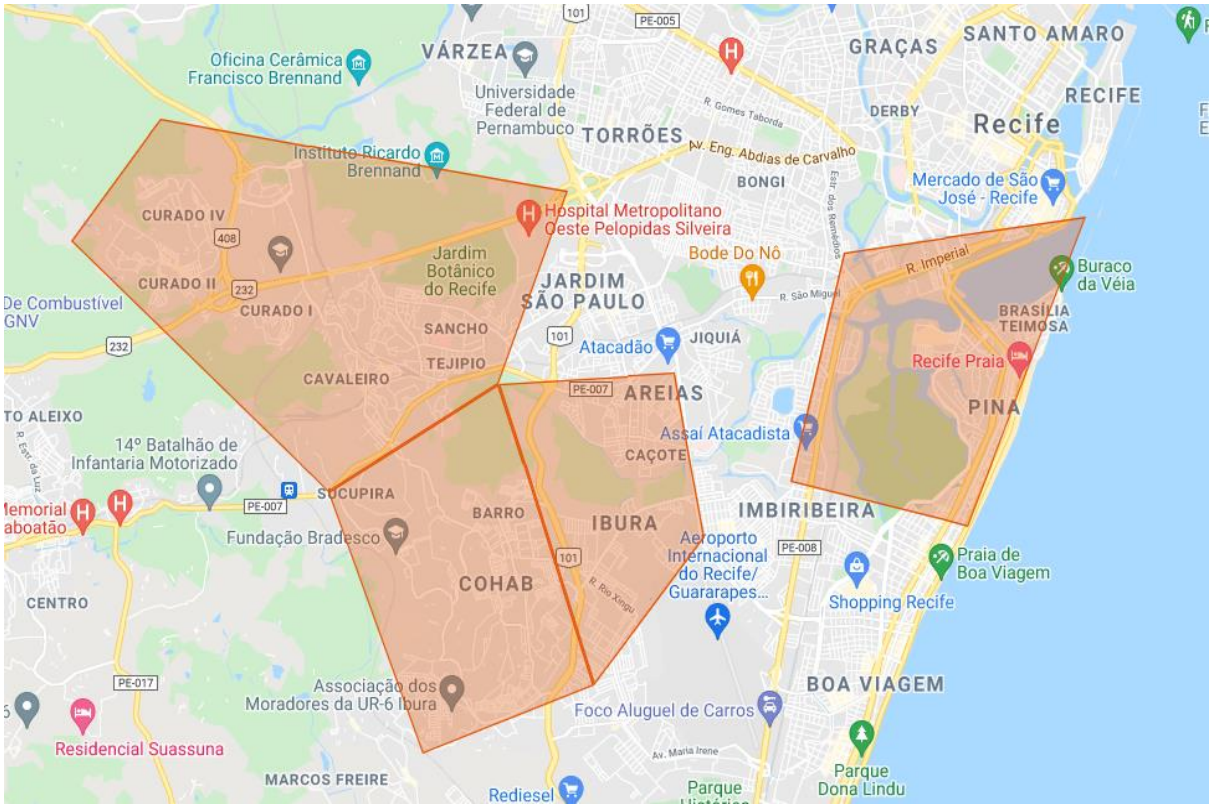
Quadro 3– Tipos de violência contra a mulher

Tipos de crimes	**Significado**
Lesão corporal	Conduta de ofensa a integridade física ou a saúde de uma pessoa.
Ameaça por violência doméstica	Ameaçar uma determinada pessoa com qualquer tipo de crime.
Estupro de vulnerável	Consiste na imposição da prática sexual por ameaça ou violência.
Maus-tratos por violência doméstica	Maus-tratos físicos e/ou psicológicos.
Dano por violência doméstica	Dano ocasionado por violência de qualquer natureza.
Injúria por violência doméstica	Ofensa que venha atingir a pessoa, em desrespeito a seu decoro, a sua honra, a seus bens ou a sua vida.
Difamação por violência doméstica	Imputação ofensiva atribuída contra a honra da mulher com a intenção de desacreditá-la na sociedade em que vive.
Vias de fatos por violência doméstica	Qualquer um dos crimes anteriores seguidos de morte.

Fonte: Esta pesquisa (2022)

Neste sentido, o algoritmo de aprendizagem de máquina considera tantos os aspectos sociais dos bairros, quanto no quantitativo da criminalidade física contra a mulher. Isto permite direcionar o comportamento daquele bairro em detrimento deste tipo de violência.

Figura 29 – Resultado do modelo de violência contra a mulher



Fonte: Esta pesquisa (2022)

Ao considerar o cenário de classificação de bairros na cidade do Recife relacionado à violência contra a mulher, o sistema providencia o mapa acima (figura 29). As áreas em vermelho representam os bairros com maior propensão de haver crimes contra a mulher. São quatro bairros ao todo, a saber: Curado II/Curado IV, Cohab, Ibura e Pina. A relação dos dados estruturados e dados não estruturados *linkados* a um repositório único de dados, permite que o algoritmo *ExtraTrees* possa analisar e realizar esta classificação. Por outro lado, é importante salientar que este resultado poderá mudar com frequência, à medida que novos dados forem sendo inseridos no sistema e novas prospecções apareçam, pois é um sistema que trabalha em tempo real.

Houve também a aplicação deste modelo tendo como referência cidades pernambucanas. Para esse estudo foi utilizado o algoritmo *ExtraTrees* que obteve os seguintes resultados (tabela 9):

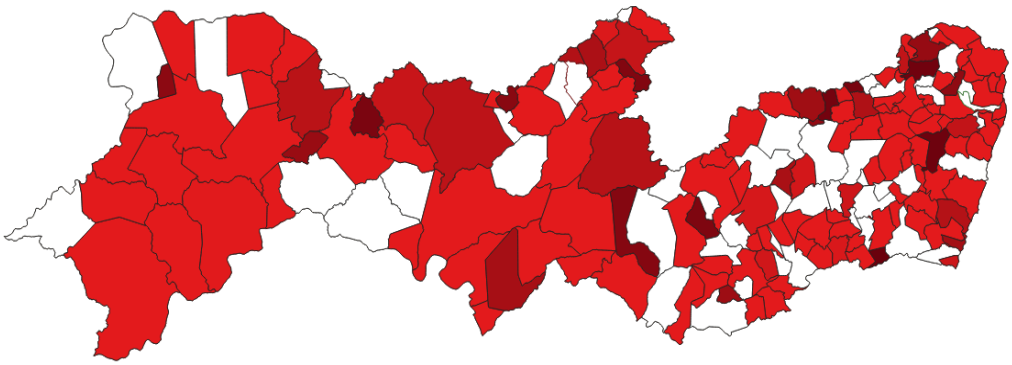
Tabela 9 – Desempenho do algoritmo para classificação de regiões

Algoritmos	**Acurácia**	**F1 - Score**
ExtraTrees	0.9	0.91

Fonte: Esta pesquisa, 2022

O algoritmo obteve bom resultado nos dois tipos de testes e evidenciou as regiões mais propensas a ter violência contra a mulher, considerando uma determinada janela de tempo (para este trabalho foi de 15 dias). Essa janela de tempo pode ser modificada pelo usuário a depender da necessidade (figura 30).

Figura 30 – Resultado do modelo de violência contra a mulher por cidade



Fonte: Esta pesquisa, 2022

As regiões estão divididas em três categorias principais: alta incidência de crimes contra a mulher (representada pela cor vermelha escura); média incidência de crimes contra a mulher (indicada pela cor vermelha clara); baixa incidência contra a mulher (representada pela cor branca) apresentadas no (quadro 6.2):

Quadro 4– Classificação de regiões considerando a violência contra a mulher

Classificação	**O que é levado em consideração?**
Alta incidência	Para os três casos são considerados:
Média incidência	Histórico da localização com relação a criminalidade; tipos de crimes contra a
Baixa incidência	mulher; fatores sociais da região (IDH, renda populacional); outros tipos de crimes executados na região.

Fonte: Esta pesquisa (2022)

Conforme maior a tonalidade da cor vermelha, maior a incidência de crime contra a mulher em suas mais diversas esferas. Percebe-se que algumas cidades estão com um vermelho mais acentuado. Estas cidades, historicamente, tendem a ter violência contra a mulher em um curto intervalo de tempo. Neste caso, não necessariamente haverá homicídio, mas poderá concentrar outros tipos de crimes aqui detalhados. O ponto importante neste modelo é conceber uma visão geral para os departamentos envolvidos, com a finalidade de estabelecer ações de controle a este tipo de crime, bem como ações de prevenção, como estabelecer delegacias de combate à violência contra a mulher, conscientização, patrulhamento de viaturas e definições de rotas para tráfego policial em localidades complexas. O modelo se complementa nas duas vertentes, identificação de bairros e cidades, proporcionando uma abrangência nas análises relacionadas.

7 CONSIDERAÇÕES FINAIS

Em segurança pública, órgãos governamentais necessitam discutir políticas que garantem ações que providenciem proteção para a sociedade. Essas ações impactam diretamente a vida das pessoas e podem produzir efeitos não intencionais. Nesse contexto, torna-se importante a análise de dados de forma contínua, pois as medidas são implementadas em tempo real levando em consideração estas análises. Além disso, é importante coletar dados não estruturados de várias plataformas. Esses dados podem relatar o cotidiano da população com maior precisão.

7.1 PRINCIPAIS CONTRIBUIÇÕES E POTENCIAIS DE USO DESTE ESTUDO

À luz dessas considerações, este trabalho propôs uma metodologia híbrida de análise de dados incorporado a um sistema de apoio a decisão chamado *DS.Security* capaz de considerar tanto dados estruturados quanto dados não estruturados. Esta metodologia permite que dados sejam processados em tempo real, permitindo que os órgãos envolvidos na segurança pública possam executar ações imediatas de combate à criminalidade, que vão desde o escalonamento policial até ações de conscientização para a população. Este estudo, como informado na sessão metodológica, obteve reconhecimento nacional e internacional, garantindo premiações, registro de software pelo INPI e trabalhos publicados em congressos nacionais, congressos internacionais e periódicos, evidenciando seu impacto social, econômico e tecnológico.

Esta metodologia permite que ações em segurança pública possam ser mais eficientes e voltadas diretamente para a sociedade, permitindo maior economia dos órgãos governamentais, uma vez que ações poderão ser mais efetivas, sem a necessidade de retrabalho. Por outro lado, novas metodologias em tratamento de dados permitem melhorias em seu processo de integração, bem como, na inclusão de novas tecnologias presentes na literatura.

Nossa metodologia híbrida apresentou excelentes resultados de acurácia e F1 score em três modelos ilustrativos, pois a composição de dados estruturados e não estruturados melhorou a precisão dos algoritmos de classificação. Evidencia-se que esta metodologia pode ser empregada em qualquer modelo que for de necessidade do órgão ou departamento.

Este estudo sugere que a aplicação de algoritmos de ML e outras técnicas de IA em uma combinação de dados estruturados e não estruturados produz maior precisão, como demonstrado na aplicação do presente trabalho (classificação de bairros de acordo com a gravidade do crime). No entanto, outras análises podem ser feitas para subsidiar outras decisões no contexto da segurança pública.

A contribuição mais significativa deste trabalho para a literatura é a perspectiva única que oferece sobre segurança pública a partir de uma abordagem de ML. A análise da UGC vinculada à segurança pública fornece resultados mais precisos que permitem análises mais robustas das ações necessárias para trazer melhorias às comunidades e bairros em questão. Este estudo também constatou que os dados oficiais fornecidos pelas secretarias de segurança pública são fundamentais para ampliar a análise e, consequentemente, alcançar resultados mais precisos. Este trabalho amplia a discussão sobre políticas de segurança pública ao mesmo tempo em que avança *DSSs* que incorporam ML, IA e UGC.

No nível local, esse tipo de análise pode melhorar as comunidades com os maiores índices de criminalidade, seja ela de baixa, média ou alta intensidade. As ações de segurança pública podem ser direcionadas para minimizar a criminalidade e aumentar a sensação geral de segurança. Em um contexto em que há recursos limitados, esta análise pode possibilitar a melhor distribuição desses recursos além de apoiar os processos de tomada de decisão em tais cenários. Com isso, este trabalho contribui para aspectos sociais, econômicos e tecnológicos.

7.2 LIMITAÇÕES E TRABALHOS FUTUROS

Quanto as limitações deste trabalho podem-se identificar processamento computacional e dependência constante de dados. Por ser um modelo que detém da necessidade constante de dados, necessita de um poder computacional mais elevado. Esta parte foi minimizada com a inclusão do *Hadoop* e formação de *cluster,* contudo ainda não é uma solução efetiva para todos os cenários. Quanto a aquisição de dados, sempre será um desafio, uma vez que os dados precisam ter qualidade. Com a inclusão de etapas de refinamento, este problema é minimizado, mas, a depender do cenário, outras etapas precisam ser consideradas.

Além disso, cada modelo retratado nos capítulos 4, 5 e 6 evidenciaram aplicações da metodologia proposta em diferentes contextos em segurança pública.

Estas aplicações foram ilustrativas e teve como objetivo mostrar o aumento da acurácia de algoritmos de aprendizagem de máquina com a utilização da metodologia híbrida de análise de dados. Desta forma, os dados utilizados nesse processo foram disponibilizados no momento das aplicações. Em um cenário real torna-se importante estudar os dados disponíveis, evidenciando aqueles que mais contribuem para o processo.

Para trabalhos futuros torna-se interessante permitir que outros órgãos governamentais, como é o caso de setores educacionais, estejam integrados ao sistema de apoio a decisão, bem como a metodologia híbrida de análise de dados. A integração de dados entre os órgãos governamentais conduz uma compreensão maior da origem da criminalidade, bem como da trajetória do crime em uma dada localidade, permitindo que ações de prevenção sejam mais eficientes em segurança pública. Por outro lado, com maior número de dados e em diversos contextos governamentais, algoritmos de aprendizagem de máquina ampliam o reconhecimento de padrões e comportamento destes dados, permitindo maior índice de acurácia destes algoritmos.

REFERÊNCIAS

ANUSHA, C. RUPA, G. Samhitha, Region based detection of ships from remote sensing satellite imagery using deep learning, in: **2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM),** p. 118–122, 2022.

ANDREWS, D. A., & BONTA, J. **The psychology of criminal conduct**. New York: Routledge, 2016.

BRASIL. **Constituição da República Federativa do Brasil de 1988**. Disponível in: <https://www.senado.leg.br/atividade/const/con1988/con1988_26.06.2019/art_144_.asp>. Access in: Feb. 23, 2021.

BALAS, S. KAMOLPHIWONG, K.L. Du. Sentimental Analysis and Deep Learning. **Advances in Intelligent Systems and Computing**, Vol. 1408, Springer, Singapore, 2022.

CAMACHO-COLLADOS, M., & LIBERATORE, F. A. Decision Support System for predictive police patrolling. **Decision Support Systems**, v.75, p. 27-37, 2015.

CHIRAMDASU, G. SRIVASTAVA, S. BHATTACHARYA, P.K. REDDY, T. REDDY, G. Malicious URL detection using logistic regression. *In*: 2021 IEEE INTERNATIONAL CONFERENCE ON OMNI-LAYER INTELLIGENT SYSTEMS (COINS). p. 1–6, 2021.

COHEN, J. Development of Crime Forecasting and Mapping Systems for Use By Police, Technical report, U.S. **Department of Justice**, 2020.

DE LIMA, F., MARINHO, E. Public security in Brazil: Efficiency and technological gaps. **EconomiA**, v.18, p. 129-145, 2017.

ELGENDY, N. & ELRAGAL, A, Big Data Analytics in Support of the Decision-Making Process. **Procedia Computer Science**. n.100, p. 1071-1084, 2016.

GRILLENBERGER, A., FAU, F.E.: Big data and data management: a topic for secondary computing education. *In:* PROCEEDINGS OF THE 10TH ANNUAL INTERNATIONAL CONFERENCE ON INTERNATIONAL COMPUTING EDUCATION RESEARCH. p. 147–148, 2014

GONG, J. SUN, Zhang. The discussion of the possibility and development pattern for constructing the public security management system based on big data, *In*: INTERNATIONAL CONFERENCE ON MATERIALS ENGINEERING AND INFORMATION TECHNOLOGY APPLICATIONS, MEITA. p. 365–369, 2020.

GERBER, M. Predicting crime using Twitter and Kernel density estimation. **Decision Support System**. v.61, p.115–125, 2014.

HASSIO, A., HARVIAINEN, T., SAVOLAINEN, R. Information needs of drug users on a local dark Web marketplace. **Information Processing and Management**. Article In

Press, 2020

HARIKRISHNA, J. RUPA, C., GIREESH, R. Deep learning-based real-time object classification and recognition using supervised learning approach. **Sentimental Analysis and Deep Learning.** p. 129-139, 2021

INGO, S., ANDREAS, C. Support Vector Machines, Springer-Verlag New York, 2008.

JAVED, P. BURNAP, O. Rana, Prediction of drive-by download attacks on Twitter. **Inf. Process. Manage**. v.56, p.1133–1145, 2019.

LIMA, T. VIEIRA, E. Costa. Evaluating deep models for absenteeism prediction of public security agents. **Appl. Soft Comput**. v.91, 2020.

LUONG, W. Applying risk/need assessment to probation practice and its impact on the recidivism of young offenders. **Crim. Justice Behav**. v.3, p.1177–1199, 2011.

LAMARI, Y. FRESKURA, B., ABDESSAMAD, A. EICHBERG, S. DE BONVILLER, S. Predicting spatial crime occurrences through an efficient ensemble-learning model. **Int. J. Geo-Inf.** v.9, p. 1–11, 2020.

LI, S.C. KUO, T. An intelligent decision-support model using FSOM and rule extraction for crime prevention. **Expert Syst. Appl**. p.7108–7119, 2010.

MINISTÉRIO PÚBLICO. Tipos de violência. Disponível em: < https://www.cnmp.mp.br/portal/index.php>. Acesso em: 20 de jul. 2022.

MURRAY, J., CERQUEIRA, D.R.C., KAHN. T, Crime and violence in Brazil: Systematic review of time trends, prevalence rates and risk factors. **Aggression and Violent Behavior**. v.18, p. 471-483, 2019.

MORONEY, L. AI and Machine Learning for Coders: A Programmer's Guide to Artificial Intelligence. O´REILLY, 2020

MCFADZIEN, K., SHERMAN, L.W. Hold them or fold them: evidence-based decisions to discontinue investigations of non-domestic minor violence. **Policing.: Int. J.** v.44, p. 643–654, 2021.

OLVER, M.E., STOCKDALE, K.C. WORMITH, S. Risk assessment with young offenders: A meta-analysis of three. **Crim. Justice Behav**. v.36, p.329–353, 2009.

POLETO, T; De Carvalho, V.D.H. Costa, A.P.C.S. The full knowledge of big data in the integration of interorganizational information: An approach focused on decision making, Int. J. **Decis. Support Syst. Technol**. v,9, p.16–31, 2017.

PETER, F. Machine Learning: The Art and Science of Algorithms That Make Sense of Data, Cambridge University Pres, 2018.

QUINLAN, J.R. Introduction of decision trees. **Machine Learning**, v.1, p. 81–106, 1986.

RENU, R.S., MOCKO, G., & KONERU, A, Use of Big Data and Knowledge Discovery to create Data Backbones for Decision Support Systems. **Procedia Comput. Sci**. v.20, 2013.

RUPA, T.R., GADEKALLU, M.H., ABIDI, A. Computational system to classify cyber crime offenses using machine learning. **Sustainability**, v.12, 2020.

RUPA, G., SRIVASTAVA, S., BHATTACHARYA, P., REDDY, T.R. A machine learning driven threat intelligence system for malicious URL detection, *IN*: THE 16TH INTERNATIONAL CONFERENCE ON AVAILABILITY, RELIABILITY AND SECURITY (ARES 2021), ASSOCIATION FOR COMPUTING MACHINERY. New York, p. 1–7, 2021

RUMMENS, H. The effect of spatiotemporal resolution on predictive policing model performance. **Int. J. Forecast**. v.37, p.125–133, 2020.

RAJESH, P., NARASIMHA, G., RUPA, C. Privacy preserving technique in data mining by using Chinease remainder theorem. *In*: ICECCS 2012: ECO-FRIENDLY COMPUTING AND COMMUNICATION SYSTEMS PP 434–442, 2012.

SOARES, F., SIVEIRA, T., FREITAS, H. Hybrid approach base don SARIMA and artificial neural networks for knowledge discovery applied to crime rates prediction, *IN*: INTERNATIONAL CONFERENCE ON ENTERPRISE INFORMATION SYSTEMS. v. 1, p. 407–415, 2020

SILVA, S. GONZÁLES, C.F.P. ALMEIDA, S.D.J. BARBOSA, H. LOPES, C. An interactive visualization system for analyzing crime data in the state of Rio de Janeiro, *In:* 19TH INTERNATIONAL CONFERENCE ON ENTERPRISE INFORMATION SYSTEMS. v. 1, Portugal, 2017.

SUMMERS, L., ROSSMO, D.K. Offender interviews: implications for intelligence-led policing. **Policing.: Int. J.** v.42, p. 31–42, 2021.

STEEVES, G. PETTERINI, F.C., MOURA, G. The interiorization of Brazilian violence, policing, and economic growth. **Economia**. v. 16, p.359-375, 2015.

TURET, J.G., COSTA, A.P.C.S. Big Data Analytics to Improve the Decision-Making Process in Public Safety: A Case Study in Northeast Brazil. *In*: DARGAM, F., DELIAS, P., LINDEN, I., MARESCHAL, B. (EDS) DECISION SUPPORT SYSTEMS VIII: SUSTAINABLE DATA-DRIVEN AND EVIDENCE-BASED DECISION SUPPORT. ICDSST 2018. LECTURE NOTES IN BUSINESS INFORMATION PROCESSING. v. 313. Springer, Cham. 2018.

TURET, J., COSTA, A.P.C.S., ZARATÉ, P. Selection of an Integrated Security Area for locating a State Military Organization (SMO) based on group decision system: a multicriteria approach. *In*: INTERNATIONAL CONFERENCE ON DECISION SUPPORT SYSTEMS TECHNOLOGY, 2020.

TURET, J., COSTA, A.P.C.S., Hybrid methodology for analysis of structured and unstructured data to support decision-making in public security. **Data & Knowledge Engineering**. v. 141, 2022.

TURET, J., COSTA, A.P.C.S., Ds.security v.2 (sistema de apoio à decisão para gerenciamento da segurança pública). *In:* ENCONTRO NACIONAL DOS PROGRAMAS DE PÓS GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO – EPPGEP, 2020

TURET, J., COSTA, A.P.C.S. Knowledge Discovery Database Model to classify neighborhoods according to violence against women. *In:* International Conference on Decision Support Systems Technology, 2022.

WANG, J., CHEN, Q., GONH, H. STMAG: A spatial-temporal mixed attention graph-based convolution model for multi-data flow safety prediction**. Information Sciences**. v.525, p. 16–36, 2020.

WANG, X., ZHAO, K., CHA, S., AMATO, M., COHN, A., PEARSON, J., PAPANDONATOS, G., GRAHAM, A. Mining user-generated content in an online smoking cessation community to identify smoking status: A machine learning approach. **Decision Support Systems. v.**116, p. 26–34, 2019.

WHITE, T. Hadoop: The Definitive Guide, 4th Edition, O'Reilly Media, 2015.

WILLIAMS, M.L., BURNAP, P., SLOAN, L. Crime sensing with big data: the affordances and limitations of using open-source communications to estimate crime patterns. **Br. J. Criminology**, v.57, p.320–340, 2017.

WAWRZYNIAK, Z., PYTLAK, R., CICHOSZ, P., JANKOWSKI, S., BOROWIK, G., OLSZEWSKI, W., SZCZECHLA, E., MICHALAK. P. The data-based methodology for crime forecasting, *In*: PHOTONICS APPLICATIONS IN ASTRONOMY, COMMUNICATIONS, INDUSTRY, AND HIGH-ENERGY PHYSICS EXPERIMENTS, 2020.

WIN, K. LI, K. CHEN, J., .VIGER, P., LI. K. Fingerprint classification and identification algorithms for criminal investigation: A survey. **Future Gener. Comput. Syst.** v.110, p.758–771, 2020.

XUE, Y. D.E. BROWN, D. Spatial analysis with preference specification of latent decision makers for criminal event prediction. **Decis. Support Syst.** v.41, p.560–573, 2006

YANG, B., LIU, L., LAN, M., WANG, Z., ZHOU, H., YU, H. A spatio-temporal method for crime prediction using historical crime data and transitional zones identified from nightlight imagery. **International Journal of Geographical Information Science**. v.34, p. 1740-1764, 2020.

ZHANG, X., LIU, L., XIAO, L., JI, J, Comparison of Machine Learning Algorithms for Predicting Crime Hotspots. **IEEE Access**. v.8, p. 18302-18310, 2022.