UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
FACULDADE DE COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

HELDER MATEUS DOS REIS MATOS

UM CLASSIFICADOR SUPERVISIONADO PARA RELATOS POLICIAIS NO ESTADO DO PARÁ

Belém
2022

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
FACULDADE DE COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

HELDER MATEUS DOS REIS MATOS

UM CLASSIFICADOR SUPERVISIONADO PARA RELATOS POLICIAIS NO ESTADO DO PARÁ

Trabalho de Conclusão de Curso apresentado para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Reginaldo Cordeiro dos Santos Filho

Belém
2022

Matos, Helder M. dos R.
UM CLASSIFICADOR SUPERVISIONADO PARA RELATOS POLICIAIS NO ESTADO DO PARÁ/ HELDER MATEUS DOS REIS MATOS. – Belém, 2022.
68 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Reginaldo Cordeiro dos Santos Filho

Monografia – UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO, 2022.

1. Mineração de Dados. 2. Aprendizado de Máquina. 3. Segurança Pública. I. Título.

HELDER MATEUS DOS REIS MATOS

UM CLASSIFICADOR SUPERVISIONADO PARA RELATOS POLICIAIS NO ESTADO DO PARÁ

Trabalho de Conclusão de Curso apresentado para obtenção do grau de Bacharel em Ciência da Computação.

Data da Defesa: 16 de dezembro de 2022
Conceito: Excelente

Banca Examinadora

Prof. Dr. Reginaldo Cordeiro dos Santos Filho
Faculdade de Computação - UFPA
Orientador

Prof. Dr. Claudomiro de Souza de Sales Júnior
Faculdade de Computação - UFPA
Membro da Banca

Prof. Dr. Filipe de Oliveira Saraiva
Faculdade de Computação - UFPA
Membro da Banca

Belém
2022

Este trabalho é dedicado a todos os professores,
os gigantes em que cujos ombros nos apoiamos,
e que nos permitem ver cada vez mais longe.

AGRADECIMENTOS

A todos os professores que fizeram parte de minha jornada, por contribuírem tanto para minha formação educacional e profissional, mas também pela formação como ser humano.

Ao professor orientador Reginaldo Santos, pelo apoio, direcionamentos e paciência concedidos durante todo o período em que estive envolvido em atividades de pesquisa, e sem os quais não poderia ter concluído este trabalho e várias outras etapas da minha formação.

Aos membros da banca, pela pronta disponibilidade em avaliarem este trabalho e estarem presentes na defesa, assim como por me ajudarem a acender a chama do pensamento científico e motivarem a seguir as demais etapas da formação acadêmica.

À esta universidade, por proporcionar a oportunidade e os recursos para realização de minha formação, assim como à esta faculdade e corpo docente, pela solicitude e esforços oferecidos durante os últimos quatro anos.

À minha família, em especial a ambas a minha mãe e avó, Vaneide Matos e Dulcemira Matos, pelo suporte, paciência e conselhos concedidos durante o período de minha formação.

Aos colegas de curso com os quais compartilhei todo o tipo de experiências, felizes e tristes, empolgantes e entediantes, eruditas e lúdicas, especialmente aos meus irmãos e irmãs da turma de 2019, Carolina Lins, Gabriel Aragão, Hádria Farias, João Lima, Lívia Carrera, Lucas Nobre e Samuel Aguiar.

Aos colegas de pesquisa que fizeram parte da construção deste trabalho e de demais contribuições, especialmente a Renan Cunha e Samara Souza.

À SIAC-PA, na figura de Cleyton Costa, Luis Gonçalves, e Marilene Tavares, por ter tornado possível a execução do trabalho e aberto as portas para troca de conhecimento entre a academia e o serviço público estadual.

A todos que direta ou indiretamente contribuíram para a realização deste trabalho e da minha integralização neste curso, muito obrigado.

"Nós somos moldados pelas ferramentas que usamos, em particular:
os formalismos que usamos moldam nossos hábitos de pensamento,
para melhor ou para pior, e isso significa que devemos ter muito cuidado na
escolha do que aprendemos e ensinamos, já que desaprender não é possível."
(Edsger W. Dijkstra)

RESUMO

Os diversos setores públicos estão se voltando para as tendências de aplicações de ciência e mineração de dados, muito em razão do aumento exponencial do volume de seus dados ao longo dos últimos anos, da consequente demanda cada vez mais frequente por informações escondidas na massa de dados gerados a cada dia, e das soluções oferecidas por estas áreas do conhecimento na automação e melhoria de processos internos. A segurança pública tem um grande potencial de adquirir benefícios gerados por ferramentas de automação de extração de conhecimento em bases de dados, incluindo a classificação de textos inclusos em relatos policiais. Este trabalho descreve o desenvolvimento de um classificador supervisionado de relatos policiais, construído a partir do conhecimento extraído de bases de dados públicas de relatos policiais, para os anos entre 2019 e 2021, no estado do Pará, Brasil. Dentre as técnicas utilizadas, destacam-se o uso de da metodologia de mineração de dados CRISP-DM, Processamento de Linguagem Natural, vetorização de sequências de texto através de *word embeddings* e um modelo de aprendizado profundo baseado em Redes Neurais Convolucionais. Este modelo alcançou uma acurácia geral de aproximadamente 78% para a predição de 463 classes únicas relacionadas com segurança pública. Tais classes incluem categorias derivadas da legislação brasileira, como homicídio, furto, roubo, estupro e ameaça, com a inclusão de classes específicas ao ambiente policial, como a comunicação de óbito, a morte por intervenção de agente do estado e o tráfico de drogas. O modelo resultante também foi usado para melhoria de processos estatísticos de analistas criminais, tanto em termos quantitativos, quanto qualitativos, através da implantação de uma ferramenta de classificação de relatos policiais diários do estado do Pará, reduzindo os esforços diários de processamento e consolidação dos dados em até 5 horas.

Palavras-chave: Mineração de Dados. Aprendizado de Máquina. Segurança Pública.

ABSTRACT

Public sectors are becoming more aware of the tendencies of data science and data mining applications, due to the exponential increase of its data volume over the recent years, the consequent and more frequent demand for hidden information in the massive amount of data generated daily, and the solutions offered by these fields of study over the automation and improvement of internal processes. Public security holds a huge potential of acquiring benefits from tools of automatic knowledge extraction on databases, including classification of text included on police records. This paper describes the development of a supervised classifier for police records, constructed upon knowledge extracted from police report public databases, in the years between 2019 and 2021, in the state of Pará, Brazil. Among the utilized techniques, it can be highlighted a data mining methodology based on CRISP-DM, Natural Language Processing, text sequence vectorization through word embeddings, and a deep learning model based on Convolutional Neural Networks. The model achieved an overall accuracy of approximately 78% for the prediction of 463 unique labels related to public safety. These labels include categories derived from the Brazilian legislation, such as murder, theft, robbery, rape, and threat, adding to labels specific to the policial environment, for instance death notice, death due to state officer intervention, and drug trafficking. The resulting model was used to improve the statistical processes of criminal analysts, both in quantitative and qualitative terms, through the deployment of a police record classification tool in the state of Pará, reducing the daily efforts of data processing and consolidation to at most 5 hours.

Keywords: Data Mining. Machine Learning. Public Security.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
AISP	Área Integrada de Segurança Pública
ANN	Artificial Neural Networks
BOP	Boletim de Ocorrência Policial
CNN	Convolutional Neural Network
CRISP-DM	CRoss-Industry Standard Process for Data Mining
ETL	Extração, Transformação e Carga
GAN	Generative Adversarial Network
HAN	Hierarchical Attention Network
KDD	Knowledge Discovery in Databases
KNN	K-Nearest Neighbors
LSTM	Long-Short Term Memory
NLP	Natural Language Processing
PRODEPA	Empresa de Processamento de Dados do Estado do Pará
RISP	Região Integrada de Segurança Pública
RNN	Recurrent Neural Networks
SEGUP-PA	Secretaria de Segurança Pública e Defesa Social do Estado do Pará
SIAC-PA	Secretaria Adjunta de Inteligência e Análise Criminal do Estado do Pará
SISP	Sistema Integrado de Segurança Pública
SOM	Self-Organizing Map
SUSP	Sistema Único de Segurança Pública
SVM	Support Vector Machines
TIC	Tecnologia da Informação e Comunicação
UF	Unidade Federativa
UFPA	Universidade Federal do Pará

SUMÁRIO

1 INTRODUÇÃO

Segurança pública é um dos setores de maior interesse para a administração pública das sociedades modernas. O esforço de manter a ordem pública pode ser observado desde o surgimento da filosofia do direito, descrito como o estudo orgânico e sistemático de um conjunto de regras que incluem o direito público e criminal (SOUZA, 1972).

No Brasil, o Sistema Nacional de Informações de Segurança Pública, Prisionais, de Rastreabilidade de Armas e Munições, de Material Genético, de Digitais e de Drogas (SINESP) é a principal infraestrutura de organização de dados e informações criminais (BRASIL, 2019). Criado em 2012, o SINESP é um sistema informacional seguro e padronizado, que facilita a comunicação entre os membros do Sistema Único de Segurança Pública (SUSP). Cada estado da federação é responsável pela coleta, normalização e entrega dos dados extraídos das delegacias policiais ao longo do país. Essa integração requer uma grande quantidade de esforço, especialmente nos interiores, onde tais sistemas tendem a falhar devido a falta de manutenção, escassez de delegacias computadorizadas ou profissionais capacitados a registrar e rotular um grande volume de registros policiais.

O aumento da quantidade de dados gerados em nossa sociedade se tornou o catalisador para a digitalização de processos ao longo de diferentes setores públicos. Mais recentemente, a aplicação de ferramentas de ciência e mineração de dados se tornou uma tendência a ser observada por esses setores.

Sob essas circunstâncias, o presente trabalho objetiva descrever o desenvolvimento de um classificador supervisionado de relatos policiais, usando dados colhidos junto a Secretaria Adjunta de Inteligência e Análise Criminal (SIAC), uma instituição ligada a Secretaria de Segurança Pública e Defesa Social (SEGUP) do estado do Pará, Brasil. Uma investigação das ferramentas usadas em mineração de dados aplicados a dados textuais foi conduzida, o que auxiliou na construção de uma ferramenta de processamento de relatos policiais para predição de um tipo de evento cuja delimitação conceitual advém da legislação brasileira em conjunto com termos do ambiente policial, com a premissa de se tornar generalizável para diferentes cenários e jurisprudências. O modelo de classificação resultante tem o potencial de se tornar um mecanismo de aceleração dos processos estatísticos da secretaria, junto com a análise qualitativa automatizada de uma grande quantidade de dados criminais.

1.1 Contexto

A SIAC tem como principais objetivos o tratamento, a análise de inteligência, a consolidação, o levantamento de relatórios estatísticos, e o armazenamento dos dados de segurança pública do estado do Pará. Os dados são fornecidos pelo Sistema de Informação da Segurança Pública (SISP), que realiza o gerenciamento de dados de boletins de ocorrência, procedimentos,

laudos e demais documentos da Polícia Civil, e chegam à secretaria por intermédio da Empresa de Processamento de Dados do Estado do Pará (PRODEPA), responsável por prover soluções de tecnologia da informação e comunicação ao estado.

A SIAC realiza o processamento da base de dados diária de Boletins de Ocorrência Policiais (BOPs), registrados nos sistemas da polícia civil do estado e que também são incorporados a diversos sistemas de informação. Os boletins de ocorrência possuem diversos campos a serem preenchidos pelo relator (escrivão da polícia civil nas delegacias ou a própria vítima através da delegacia virtual), que posteriormente são extraídos dos sistemas de registro e armazenados em sistemas de bancos de dados, permitindo um acesso facilitado aos diversos interessados nos dados de segurança pública do Pará.

Dentre os campos preenchidos pelo relator, destaca-se o *relato*, uma descrição textual da ocorrência que narra o evento ocorrido, lista os participantes e seus papeis (relator, vítimas e autores), e que geralmente acompanha identificações de localidade (local do fato e do registro), tempo (data/hora do fato e do registro) e *modus operandis*, como o meio empregado usado no delito. Muitas dessas informações são extraídas do relato e discretizadas em outros atributos, de forma que o relato é o componente central para extração de conhecimento para cada amostra.

A cada boletim de ocorrência presente na base de dados interna da SIAC é atribuído um campo conhecido como *consolidado*, a categoria na qual aquela ocorrência possa ser quantizada e que auxilie na descrição estatística de recortes temporais da base. As classes de consolidados são utilizadas na geração de relatórios que descrevem os índices de segurança pública para os respectivos gestores administrativos do setor, impactando na tomada de decisão em relação à incidência destas classes, e da definição de metas e objetivos a serem especulados para os próximos meses, além dos dados disponibilizados nos portais de transparência da SEGUP e concedidos a instituições e indivíduos interessados. A determinação de qual classe de consolidado para cada boletim não é trivial e requer a leitura e análise de seu respectivo relato e julgamento de um ou mais indivíduos, dentre um grupo de analistas especialmente alocado pela secretaria para esta tarefa. Alguns exemplos de classes de consolidados incluem *Ameaça*, *Comunicação de Óbito*, *Estelionato*, *Estupro*, *Estupro de Vulnerável*, *Furto*, *Homicídio*, *Lesão Corporal*, *Morte por Intervenção de Agente do Estado*, *Roubo*, *Suicídio* e *Tráfico de Drogas*.

O presente trabalho propõe uma solução para este problema de consolidação dos boletins de ocorrência policiais, na qual a SIAC disponibilizou um conjunto de dados para os anos de 2019, 2020 e 2021. Tal solução se apoia na tarefa de classificação de textos, o que demanda a escolha de um atributo textual e de um atributo com as classes-alvo do problema. O relato descritivo é escolhido como o atributo textual de onde o conhecimento será extraído, enquanto que o consolidado é escolhido como o atributo-alvo para o problema de classificação, visto que é o atributo de classes mais preciso a ser encontrado na base de dados.

Mediante esta contextualização, três desafios foram evidenciados como importantes de serem tratados como parte da visão geral do problema, dadas as suas complexidades:

- Alta quantidade de classes: o problema de multi-classificação proposto contém um número considerável de classes, muitas vezes sendo sobrepostas e descrevendo eventos bastantes similares com classes diferentes. Dessa forma, é necessária a escolha de quais classes são interessantes para compor a solução, de forma equilibrada e que consiga abranger a maior quantidade de amostras sem prejudicar a performance.

- Conjunto de dados desbalanceado: a frequência absoluta das classes de consolidados é exposta na Figura 1, contendo um ranking das 25 classes mais frequentes no conjunto de dados. Enquanto que a proporção de registros para a classe mais popular e a 25ª expostas na figura (furto e conflitos familiares, respectivamente) é da ordem de quase 30 vezes de diferença, uma grande quantidade de classes raras do conjunto de dados apresentam apenas um relato, o que compromete significativamente a distribuição das probabilidades de predição.

- Classes dinâmicas: ao longo do tempo, novas classes podem ser adicionadas ou removidas. Um exemplo de adição recente é a *Invasão de Dispositivo Informático*, que passa a ser categorizada a partir de 2016 em razão do aumento do combate aos crimes cibernéticos. Por outro lado, a classe de *Roubo de Veículo*, que começou a ser categorizada em 2015, teve suas amostras de BOPs contabilizados como *Roubo* a partir de 2021, restando a outras colunas a especificação de qual o objeto alvo da ação. Dessa forma, o modelo que representa o conhecimento extraído da base necessita de constante atualização, seja através de ajustes nas abordagens existentes ou pela geração de novas modelagens.

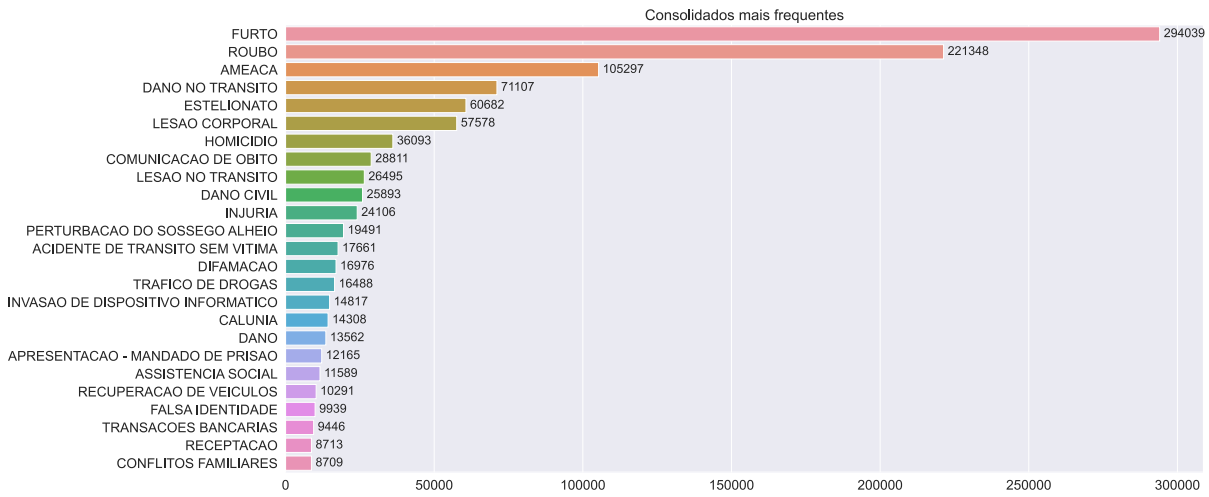


Figura 1 – As 25 classes mais frequentes do conjunto de dados.

A metodologia empregada dedica esforços para a minimização destes problemas.

1.2 Justificativa

Há um consenso cada vez maior entre os gestores da administração pública em aplicar os esforços de policiamento ostensivo apenas onde este é realmente necessário, o que requer um trabalho de inteligência que garanta a possibilidade de uma vigilância preventiva (CASTELLA, 2002). A eficiência de tal serviço de inteligência depende de um volume e qualidade elevada de informações que descrevam o cenário na qual as atividades de investigação estão ambientadas, requisitos estritamente relacionados com o uso inteligente dos recursos disponíveis na busca por soluções cada vez mais atuais e confiáveis.

Uma das principais frentes de investimentos da segurança pública brasileira nos últimos anos é em ciência, tecnologia e inovação. Apesar de ainda enfrentar limitações e carência de uma política nacional que contemple diretrizes a serem seguidas, é possível notar um salto no interesse geral das iniciativas públicas desde a virada do milênio (MIRANDA, 2012). Nesse sentido, a Agência Brasileira de Desenvolvimento Industrial (ABDI, 2010) listou algumas aplicações emergentes de Tecnologia da Informação e Comunicação (TICs) que possam ser investigadas prioritariamente, incluindo:

- Biometria: identificação por impressão digital, íris, voz, traços da face, DNA, palmar, etc.;
- *Radio-Frequency Identification* (RFID): método de identificação automática através de sinais de rádio, recuperando e armazenando dados remotamente;
- Video-monitoramento e câmeras inteligentes: sistemas de vídeo com identificação de caracteres ou câmeras com detecção de movimentos e sensor de posição;
- Softwares de inteligência: softwares que atendem o ciclo de produção de informações estratégicas, contendo ferramentas para coleta, análise (redes de relacionamento, geoprocessamento, mineração de dados, etc) e difusão de informações;
- Conexão ultra-segura: conexões que se utilizam de criptografia com alta dificuldade de quebra mesmo com máquinas de alto desempenho;
- Sistemas de monitoramento e bloqueio de sinais: sistemas que podem monitorar e bloquear radiofrequências para o ambiente prisional;
- Redes integradas de telecomunicações: redes de comunicação de missão crítica integradas entre as instituições de segurança pública baseadas em padrões abertos;
- Sistemas avançados de bancos de dados: propostas de unificação de dados dispersos em diversas bases visando alimentar os sistemas de inteligência;
- Sistemas de detecção e reconhecimento de padrões de vídeo: sistemas que trabalham em conjunto com câmeras de vídeo para detecção de objetos perigosos e reconhecimento de padrões em imagens.

Somado a isto, em relação ao cenário nacional de inteligência e dados, segundo o Anuário Brasileiro de Segurança Pública de 2022 (ABSP, 2022), o estado do Pará está na 5ª posição do ranking de qualidade estimada dos registros estatísticos de mortes violentas intencionais no ano de 2021. Este ranking é o principal critério de comparação da qualidade dos dados estatísticos entre as unidades da federação (UF). Em geral, o Pará apresenta o melhor desempenho entre os estados da região Norte, e figura no grupo de UFs de mais alta qualidade em relação à pontuação final obtida. É importante dar destaque para o desempenho paraense em critérios como o conceito de mínimo de informações sobre os registros[1], onde o Pará recebe pontuação máxima. Quanto à convergência entre as fontes de informação[2], o Pará apresenta o segundo melhor desempenho, apenas atrás de São Paulo. Outros índices apontam para um desempenho razoável na definição do conceito de morte violenta e da transparência dos dados. Por outro lado, a qualidade dos dados de mortes do estado do Pará encontra deficiências na quantidade de informações perdidas entre vítima e fato, além de uma grande proporção de casos indeterminados. Tal desempenho evidencia os esforços da administração pública paraense na manutenção da qualidade dos dados de segurança pública, e mostra quais os pontos que necessitam de melhoria para o próximo ano.

Dessa forma, esse trabalho pretende desenvolver uma ferramenta computacional que modele o conhecimento necessário para a determinação classes de eventos policiais, agindo no eixo de conceitualização dos registros de segurança pública, permitindo assim uma análise determinística, livre de vieses, e que possa manter a qualidade do veredito humano com a celeridade de processos automatizados, desde que seja garantida a sua confiabilidade. Também é desejável que tal abordagem seja capaz de somar esforços com metodologias tradicionais de tratamento dos dados de segurança pública, a fim de auxiliar em tarefas de estatística, reduzir esforços manuais e temporais e contribuir na melhoria da qualidade dos dados formalizados nos sistemas de bancos de dados públicos paraenses.

1.3 Objetivos

1.3.1 Objetivo Geral

O objetivo geral deste trabalho é o desenvolvimento de um classificador de boletins de ocorrências policiais em um conjunto de classes específicas à segurança pública, utilizando técnicas de mineração de dados, como a descoberta de conhecimento em bases de dados e processamento de linguagem natural, e aprendizado de máquina supervisionado, como as redes neurais convolucionais e *word embeddings*.

1.3.2 Objetivos Específicos

Foram estabelecidos 5 objetivos específicos para este trabalho:

[1] Informações sobre a vítima, o fato e o agressor presumível.
[2] Convergência entre os homicídios obtidos e as certidões de óbito emitidas.

- Investigar o uso de mineração de dados na segurança pública;
- Descrever e analisar métricas avaliativas que possibilitem que este trabalho seja comparado a outras abordagens semelhantes de classificação de relatos policiais;
- Auxiliar na automação de processos estatísticos de segurança pública de forma contínua e robusta;
- Implantar tecnologia de mineração de dados em ambientes de produção;
- Desenvolver ferramenta de livre acesso que possa ser usada e aperfeiçoada por indivíduos e instituições interessadas.

1.4 Estrutura do Trabalho

Este trabalho está dividido em 6 capítulos: o Capítulo 1 introduziu o contexto, a justificativa, e os objetivos geral e específicos; Capítulo 2 aborda os referenciais teóricos necessários para uma boa compreensão do trabalho; o Capítulo 3 descreve um levantamento dos trabalhos relacionados que permitiram o entendimento dos avanços recentes no meio onde o trabalho está inserido; o Capítulo 4 descreve as metodologia propostas; o Capítulo 5 apresenta os resultados derivados da aplicação das metodologias; e o Capítulo 6 apresenta as considerações finais e as propostas de trabalhos futuros.

2 REFERENCIAIS TEÓRICOS

Este capítulo cita e fornece um referencial teórico básico para contextualização científica e compreensão satisfatória da metodologia empregada.

2.1 Descoberta de Conhecimento em Bases de Dados

Knowledge Discovery in Databases (KDD) é um campo emergente e multidisciplinar que se preocupa com métodos e técnicas capazes de extrair informações úteis (conhecimento) de dados. Segundo (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996):

> *"Há uma necessidade urgente de geração de teorias computacionais e ferramentas que auxiliem humanos a extrair informações úteis (conhecimento) dos volumes de dados digitais cujo crescimento é cada vez mais acelerado."*

> *"O problema básico enfrentado pelo processo KDD é de mapear dados de baixo nível em outros formatos que possam ser mais compactos, mais abstratos, ou mais úteis."*

Dessa forma, é definido um processo interativo e iterativo, que seja capaz de orientar os interessados na extração de conhecimento através de diversas etapas, ilustradas na Figura 2.

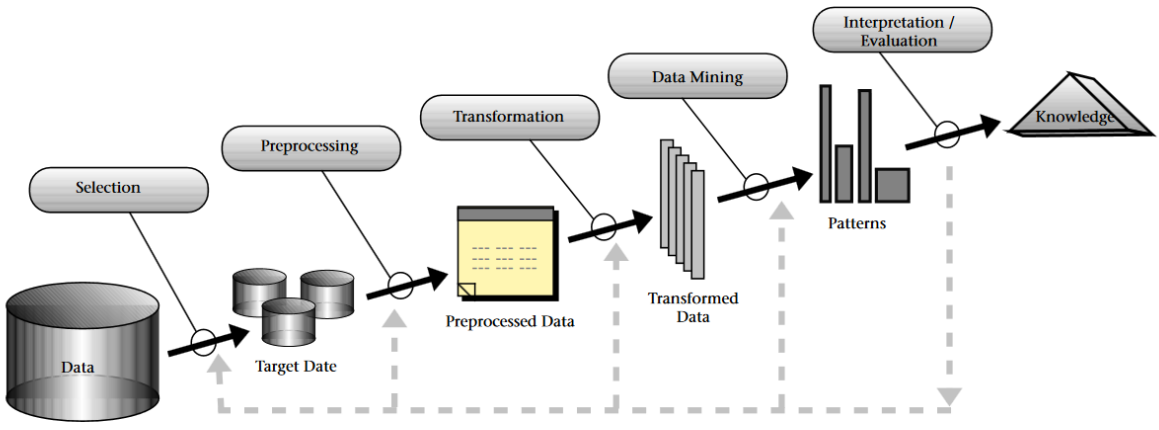


Figura 2 – Visão geral das etapas do processo KDD.

As cinco etapas do processo KDD podem ser descritas da seguinte forma:

1. Seleção: escolha do conjunto de dados alvo e escolha de subconjuntos de variáveis ou amostras na qual a descoberta de conhecimento será performada;

2. Pré-processamento: remoção de ruído e manipulação de dados ausentes e discrepantes;

3. Transformação: destaque dos atributos que melhor representam os dados para a tarefa de descoberta de conhecimento, através de redução de dimensionalidade ou formatação dos dados para entrada em algoritmos de processamento;

4. Mineração de dados: aplicação de representações formais e modelos de representação de dados na procura por padrões de interesse nos dados processados;

5. Interpretação e Avaliação: visualização dos padrões extraídos, dos dados gerados pelos modelos de mineração, e possível retorno a qualquer uma das etapas anteriores.

2.2 CRISP-DM

CRoss-Industry Standard Process for Data Mining (CRISP-DM) é uma abordagem direcionada à indústria para orientação de modelos de processos de mineração de dados, inspirado no processo KDD (IBM Corp, 2011). É comumente dividido em seis fases:

1. Entendimento do negócio: exploração das expectativas da organização/projeto em relação às soluções de mineração de dados, determinando os objetivos do negócio e de mineração, avaliando cenários e produzindo um plano de projeto;

2. Entendimento dos dados: exploração dos dados disponíveis para mineração, através de sua coleta, descrição, e verificação da qualidade;

3. Preparação dos dados: manipulação dos dados a serem modelados, através de sua seleção, limpeza, construção, integração e formatação;

4. Modelagem: geração de modelos que atendam às expectativas do entendimento do negócio, através da seleção de técnicas de modelagem, da geração de cenários de testes, e da construção e avaliação dos modelos;

5. Avaliação: verificação dos critérios de sucesso dos modelos gerados, através da avaliação dos resultados e de processos de revisão das etapas anteriores;

6. Implantação: uso do conhecimento adquirido para melhoria dos processos do negócio, através do planejamento da implantação, do monitoramento e da manutenção, além da produção de um relatório final.

2.3 Processamento de Linguagem Natural

Natural Language Processing (NLP) é uma sub-área da Inteligência Artificial que se preocupa com a aquisição de conhecimento por parte de um agente em exemplos de linguagem

natural (comunicação humana e linguagem escrita), através do uso de modelos de linguagens (RUSSELL; NORVIG, 2010). Em geral, pode ser divida em três segmentos:

- Classificação de Texto: dado um exemplo de texto, trata da decisão de qual categoria este texto pertence em relação a um conjunto pré-definido de classes. Pode ser subdivida em identificação de idioma, classificação de gênero, análise de sentimentos, e detecção de *spam*. Geralmente requer a organização dos elementos textuais (caracteres, palavras e sentenças) em estruturas como *n-gramas* (sequências de elementos delimitada por uma janela deslizante de tamanho *n*), ou em *bag of words* (conjunto não-ordenado de unigramas dos elementos). Quanto a número de classes, pode ser de dois tipos: classificação binária de texto, para duas classes (geralmente positivo ou negativo, sim ou não, 0 ou 1, etc); ou multi-classificação de texto, para mais de duas classes (como uma lista de idiomas, de categorias, de sentimentos, etc);

- Recuperação de Informação: é a busca por documentos relevantes que atendam a uma necessidade de informação. Tal busca é caracterizada pelo conjunto de documentos a ser pesquisado, pela *query* de consulta imposta, pelo conjunto de resultados e por sua consequente apresentação.

- Extração de Informação: é a aquisição de conhecimento obtido através do ato de percorrer um texto e buscar por ocorrências de casos particulares ou relacionamentos entre os elementos textuais. Técnicas comuns incluem o uso de Autômatos de Estado-Finito (usualmente implementados por Expressões Regulares), Modelos Probabilísticos (Modelo Oculto de Markov, por exemplo), Extração de Ontologias (fatos estruturados a partir de exemplos anotados) e Leitura de Máquina (extração automatizada de ontologias).

2.4 Aprendizado de Máquina

Machine Learning é uma sub-área da Inteligência Artificial derivada da abordagem de Aprendizado a partir de Exemplos, que envolve o ajuste automatizado de modelos computacionais de acordo com a sua exposição a informações contidas em um conjunto de dados (AGGARWAL, 2015). Pode ser dividido em:

- Aprendizado Supervisionado: algoritmos cujo aprendizado é obtido a partir da exposição a dados rotulados que generalizem qualquer tipo de entrada esperada. Alguns exemplos de técnicas supervisionadas incluem Regressão Logística, Máquinas de Vetor de Suporte, Árvores de Decisão e Florestas Aleatórias.

- Aprendizado Não-Supervisionado: algoritmos capazes de aprender padrões de dados não-rotulados, geralmente explorando critérios estatísticos, geométricos ou de similaridade. O exemplo mais comum deste aprendizado é o *K-Means Clustering*.

- Aprendizado de Reforço: algoritmos que avaliam a qualidade de uma solução e baseiam seu aprendizado do reforço positivo ou negativo de suas avaliações, através da exploração de um espaço de soluções. Alguns exemplos incluem os Métodos de Monte Carlo, Força Bruta e *Q-Learning*.

Quanto aos problemas de classificação em aprendizado supervisionado, que é o principal foco de interesse do presente trabalho, um conjunto de atributos numéricos ou categóricos é utilizado para descrever uma amostra do conjunto de aprendizado, onde pelo menos um atributo de classe indica qual o nível de problema de classificação:

- Classificação binária: apenas um atributo de classe que possui duas possibilidades de valores, como 0 ou 1, sim ou não, correto ou incorreto, etc.
- Classificação multi-classe: apenas um atributo de classe que possui mais que duas possibilidades de valores, como a sequência de dígitos de 0 a 9, uma lista de sentimentos que um filme pode despertar, as tipologias criminais de uma base de dados de segurança pública, etc.
- Classificação multi-classe multi-saída: mais que um atributo de classe com mais que duas possibilidade de valores, como os estilos de cada peça de roupa de um vestuário, tipos de animais presentes em uma enciclopédia, as especificações de *modus operandis* de uma base de dados de segurança pública, etc.

A avaliação de uma solução para um ploblema de classificação passa pela comparação entre as classes esperadas para os exemplos de aprendizado e as classes obtidas pelo modelo ajustado, obtidas por um processo conhecido como predição. Um conjunto extensivo de métricas pode ser utilizada para medir o quão próximo as predições estão do que é esperado para os exemplos de aprendizado, incluindo:

- *True Positives* (TP): ou verdadeiros positivos, são as amostras de uma classe que foram corretamente preditas como pertencendo àquela classe;
- *False Negatives* (FN): ou falsos negativos, são as amostras de uma classe que foram incorretamente preditas como não pertencendo àquela classe;
- *False Positives* (FP): ou falsos positivos, são as amostras que não pertencem a uma classe que foram incorretamente preditas como pertencendo àquela classe;
- *True Negatives* (TN): ou falsos negativos, são as amostras que não pertencem a uma classe que foram corretamente preditas como não pertencendo àquela classe;

- *Accuracy* (ACC): ou acurácia, é proporção de classificações corretas (*True Positives* e *True Negatives*), em relação numero total de amostras de teste;

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

- *True Positive Rate* (TPR): ou sensibilidade, é a proporção de classificações positivas corretas (*True Positives*) dentre todas as amostras verdadeiramente positivas;

$$TNR = \frac{TP}{TP + FN}$$

- *True Negative Rate* (TNR): ou especificidade, é a proporção de classificações negativas corretas (*True Negatives*) dentre todas as amostras verdadeiramente negativas;

$$TNR = \frac{TN}{TN + FP}$$

- *Positive Predictive Value* (PPV): ou precisão, são os acertos dentre as predições verdadeiras;

$$PPV = \frac{TP}{TP + FP}$$

- *F1-score*: é a média harmônica entre a precisão e a sensibilidade;

$$F1 = \frac{2 \times PPV \times TPR}{PPV + TPR}$$

- *Matthews Correlation Coefficient* (MCC): ou Coeficiente de Correlação de Matthews, é uma métrica importante para problemas de classificação multi-classe em conjuntos de dados desbalanceados, fornecendo um cálculo uniforme ao longo de todos os quatro índices das matrizes de confusão individuais;

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

2.4.1 Redes Neurais Artificiais

Artificial Neural Networks (ANN) são sistemas computacionais inspirados nos neurônios biológicos e na estrutura do cérebro, com foco no conhecimento a partir de exemplos, objetivando a sua aquisição mediante aprendizado, a sua representação através dos valores numéricos dos neurônios artificiais, e a sua aplicação em diversas áreas. É formalmente definida como (HAYKIN, 2001):

> *"Uma rede neural é um processador maciçamente paralelamente distribuído constituído de unidades de processamento simples, que têm a propensão natural para armazenar conhecimento experimental e torná-lo disponível para o uso. Ela se assemelha ao cérebro em dois aspectos:*

1. *O conhecimento é adquirido pela rede a partir de seu ambiente através de um processo de aprendizagem.*
2. *Forças de conexão entre neurônios, conhecidas como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido."*

A Figura 3 ilustra o *Perceptron*, a unidade de processamento de uma rede neural proposto para aprendizado, também conhecido como neurônio artificial (ROSENBLATT, 1957).

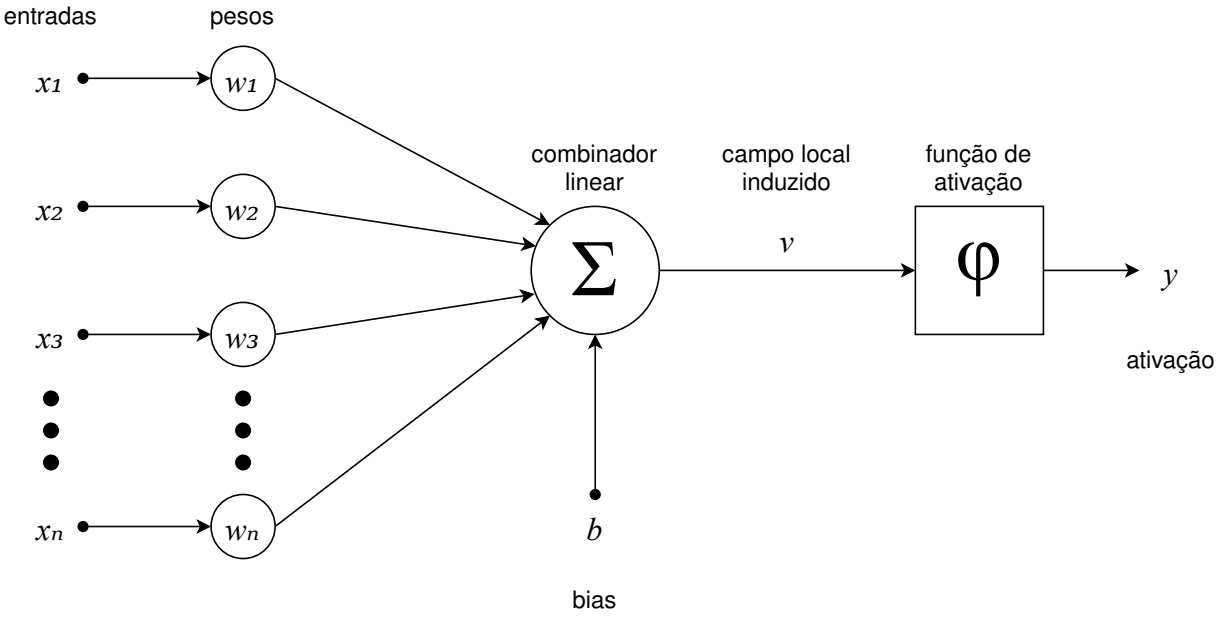


Figura 3 – Perceptron, o neurônio artificial de Rosenblatt.

Cada neurônio artificial j computa uma soma ponderada de suas n entradas x, gerando um campo local induzido v_j:

$$v_j = \sum_{i=0}^{n} w_{i,j} x_i$$

Esta soma é então aplicada a uma função de ativação ϕ, em conjunto com um *bias* b, gerando o sinal de saída y_j do neurônio:

$$y_j = \phi(v_j + b_j)$$

A função de ativação é uma função matemática diferenciável que age como um limitador que garante a propriedade da rede neural de representar uma função não-linear. Dentre os tipos mais comuns de funções de ativação podemos citar:

- Função *limiar*: retorna 1 para entradas positivas e 0 para negativas;
- Função *sigmóide*: retorna valores reais entre 0 e 1;

- Função *ReLU*: retorna o valor de entrada, caso esta seja positiva, ou 0, caso a entrada seja negativa. Geralmente usada em redes neurais profundas.

Em resumo, o perceptron é um combinador linear dos pesos sinápticos w das entradas x aplicadas em conjunto com um *bias* b, cujo campo induzido local resultante v é aplicado a uma função de ativação ϕ para gerar uma saída y.

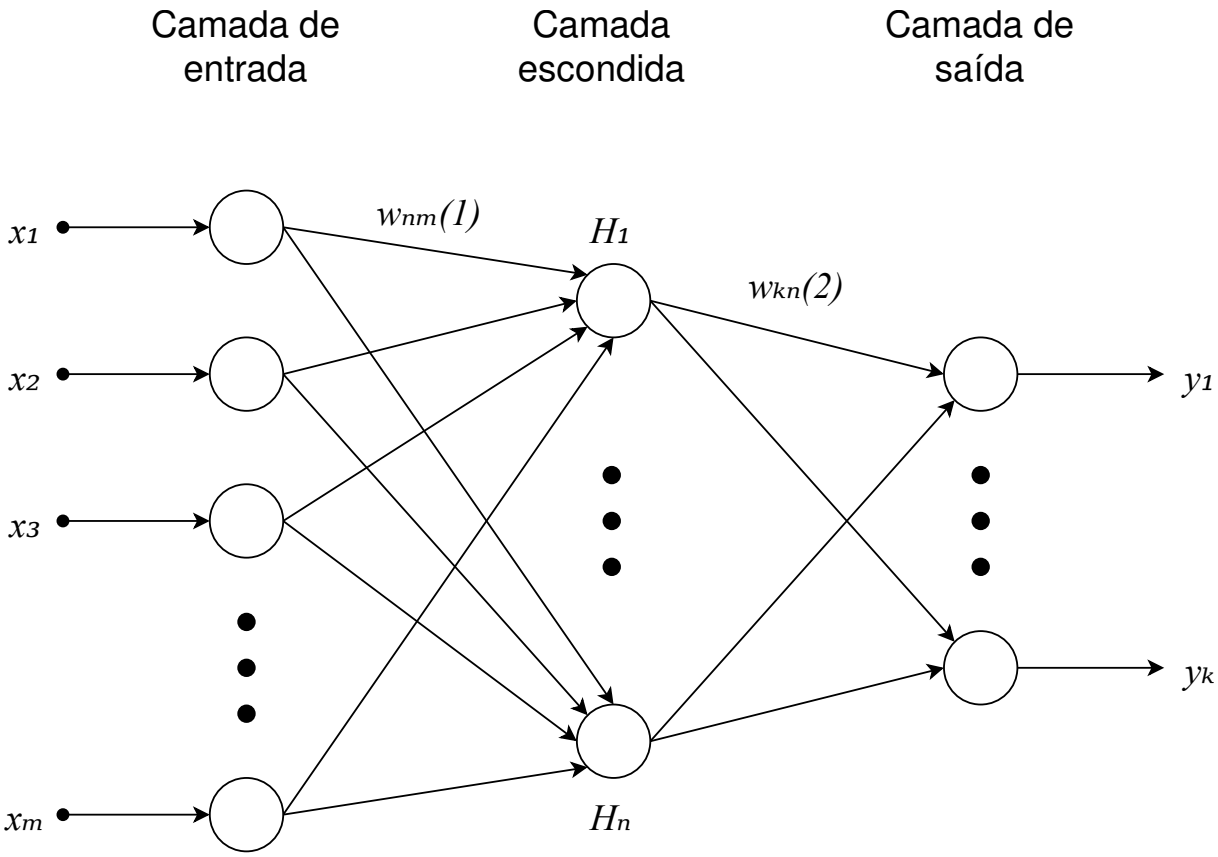


Figura 4 – Uma rede neural de única camada escondida.

Uma rede neural artificial é composta por vários neurônios artificiais, interconectados e agrupados em camadas. A Figura 4 ilustra uma rede neural de múltiplas camadas, composta de uma camada de entrada, uma camada escondida e uma camada de saída.

A extração de conhecimento por uma rede neural de múltiplas camadas é conduzida através da tarefa de treinamento, computada por um algoritmo de aprendizado. *Backpropagation* (retropropagação) é o algoritmo mais consagrado para treinamento, consistindo de duas fases:

- Fase de propagação: os sinais de entrada são propagado ao longo da rede, passando por cada camada com respectivos pesos fixados, até gerar uma saída y_j.
- Fase de retropropagação: uma comparação entre a saída da propagação e a saída esperada d_j gera um sinal de erro $e_j = d_j - y_j$, propagado ao longo da rede em sentido inverso, atualizando os pesos sinápticos.

Cada iteração do treinamento é chamada de época, e consiste na execução das duas fases do algoritmo. São necessárias várias épocas de exploração das entradas para realizar o ajuste dos pesos sinápticos que melhor representem os dados a serem modelados pela rede. Encontrar o número de épocas de treinamento ideal consiste na medição da performance da rede em termos de capacidade preditiva, o que demanda testes de diferentes valores de épocas. Existem diversos outros parâmetros a serem configurados além do número de épocas, como a quantidade de camadas escondidas, a quantidade de neurônios por camada, as funções de ativação envolvidas, os algoritmos de cálculo de erro, etc. A combinação destes parâmetros para construção de um modelo gera uma arquitetura de rede neural.

2.4.2 Aprendizado Profundo

Deep Learning se refere a arquiteturas de Redes Neurais Artificiais multi-camadas, para processamento de unidades não-lineares (BLUM; HOPCROFT; KANNAN, 2015). Cada camada adiciona uma ênfase no aprendizado de representações dos dados cada vez mais significativas, geralmente variando entre 2 até 100 ou mais camadas escondidas (CHOLLET, 2018). A figura 5 mostra como uma Rede Neural Profunda consegue processar os dados originais de um vetor que representa um relato descritivo de um boletim de ocorrência para multi-classificação do texto em uma classe dentre 8 possibilidades de categorias criminais.

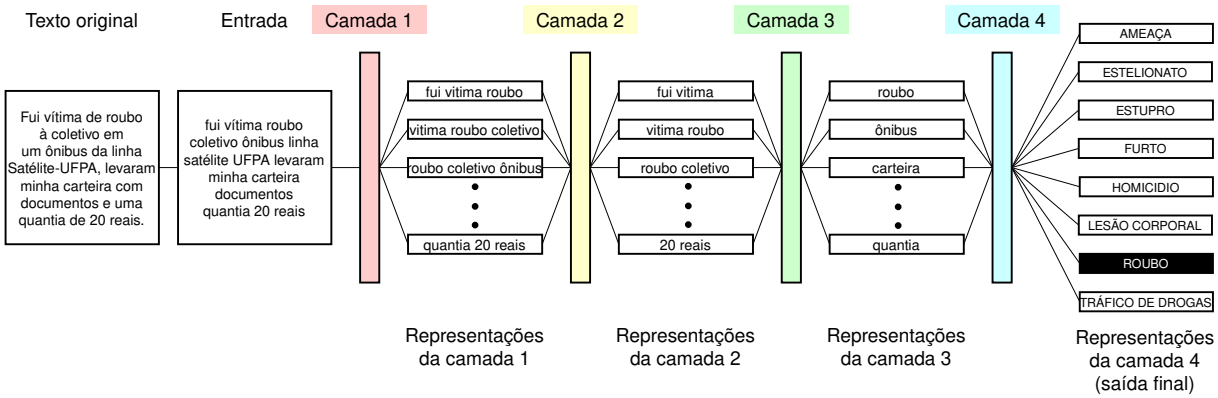


Figura 5 – Uma rede neural profunda para classificação de texto.

Este exemplo mostra que a rede transforma a entrada em representações cada vez mais informativas, performando filtragens dos dados ao longo das camadas e gerando um resultado que responda a problemática apresentada, que neste caso é de classificação multi-classe de texto. Redes neurais profundas possuem um número considerável de parâmetros a serem configurados, e um número ainda maior de unidades de processamento a serem computadas, exigindo o uso de máquinas com alta capacidade computacional, especialmente na etapa de treinamento.

Algumas arquiteturas consagradas na literatura são de importante destaque para a compreensão das capacidades do aprendizado profundo e suas aplicações, incluindo:

- *Convolutional Neural Network* (CNN): redes amplamente usadas para tarefas de visão

computacional, cujo funcionamento é centrado na operação de convolução, que consiste na extração de trechos do vetor de entrada que passam a sofrer uma mesma transformação a fim de gerar uma saída contendo um mapa de características. Em aplicações de processamento de imagens tem um formato de representação bidimensional, mas também se apresenta como tridimensional no processamento de vídeos, ou até mesmo unidimensional para processamento de sequências de texto;

- *Recurrent Neural Network* (RNN): redes amplamente usadas para Processamento de Linguagem Natural, onde as sequências de entrada processadas possuem estados que contém informações do que foi previamente processado, abstraindo a forma como capturamos informações ao ler um texto, com um golpe de vista de cada vez enquanto mantém memórias do que foi recentemente lido. Derivam desta arquitetura a *Long Short-Term Memory* (LSTM) e a *Hierarchical Attention Network* (HAN);
- *Generative Adversarial Network* (GAN): é um modelo de aprendizado não-supervisionado que gerencia o confronto entre dois modelos supervisionados: o gerador, que gera novos exemplos, e o discriminador, que verifica se os exemplos criados são parte do domínio do problema. O confronto termina quando o discriminador é enganado em pelo menos metade dos exemplos gerados, indicando a existência de um conjunto de exemplos plausíveis. São usadas para gerar arte, anúncios e mídia interativa e para remasterização de jogos;
- *Self-Organizing Map* (SOM): é um modelo de aprendizado não-supervisionado que produz representações da estrutura topológica dos dados em dimensões reduzidas. São aplicadas na redução de dados de entrada, aceleração de aprendizado não-linear e na compressão de informação.

Alguns dos problemas que envolvem soluções de redes neurais profundas incluem:

- Classificação de imagens;
- Entendimento de linguagem natural;
- Reconhecimento de fala;
- Transcrição de texto manuscrito;
- Tradução de máquina avançada;
- Direção autônoma;
- Direcionamento personalizado de anúncios;
- Habilidade de jogar Go, Xadrez e outros jogos competitivos;
- Raciocínio lógico formal.

3 TRABALHOS RELACIONADOS

O levantamento dos trabalhos relacionados foi restrito para o período entre janeiro de 2018 a junho de 2022, onde foram separados 21 resultados importantes pela sua relevância em relação ao problema proposto. Além disso foram listados trabalhos encontrados em referências bibliográficas dos resultados da busca e trabalhos publicados pelo autor em conferências, listados no Apêndice A - Publicações.

Os resultados foram divididos em duas subseções, a primeira referente a trabalhos com viés estatístico e de ciência de dados clássica, enquanto que a segunda foca em trabalhos relacionados a algoritmos de aprendizagem de máquina.

3.1 Ciência de dados

A seguir serão descritos trabalhos que relacionam estatística e ciência de dados à segurança pública, totalizando 9 trabalhos (2 de autoria própria e 7 encontrados em mecanismos de buscas).

O presente autor participou da publicação de dois trabalhos onde é realiza a aplicação de ferramentas de ciência e minerações de dados em bases públicas de segurança fornecidos pela SIAC para os anos de 2019, 2020 e 2021 na capital do Estado do Pará, Belém, em relação a crimes violentos (mortes, roubos e furtos) (SOUZA et al., 2022a; SOUZA et al., 2022b). Com o apoio da metodologia CRISP-DM, é proposta a aplicação de uma análise exploratória dos dados em conjunto com a geração de regras associativas, fornecendo um panorama dos índices criminais no período fornecido que auxilie as autoridades em processos de tomada de decisão. Os resultados da análise exploratória indicam: uma queda de cerca de 41% dos furtos e de 25% dos roubos para o período entre 2019 e 2020 (marcado pelo início da pandemia de COVID-19); os impactos dos 8 primeiros meses da pandemia no Brasil (abril a dezembro de 2020) possuem uma correlação com baixos índices de registros de crimes violentos; furtos tem grande incidência durante a manhã, enquanto roubos e homicídios são mais comuns durante a noite; os bairros da capital com maior número de incidências incluem Marco, Guamá e São Brás; a faixa-etária mais frequente dentre as vítimas está entre 35 e 64 anos de idade, predominantemente homens. Quanto às regras associativas, 6 se destacam com um valor de confiância acima de 0,7, delineando alta correlação do crime de furto em relação a: ausência de instrumento empregado para consumação do crime; vítimas tendem a serem homens entre 35 e 64 anos; as vias públicas do bairro do Marco nas manhãs de Outubro de 2019.

Foi concebido um estudo da correlação entre as investigações de 256 registros de homicídio em relação ao perfil das vítimas desse crime durante o primeiro semestre de 2019 em Belém, estado do Pará (COSTA et al., 2020). Uma análise exploratória de dados revelou que: 89% das vítimas envolvidas eram do sexo masculino; a faixa etária predominante varia entre

18 a 34 anos em 61,6% dos casos; majoritariamente, um grau de escolaridade de estudo até o ensino médio está presente em quase 99% dos casos de vítimas; em relação a cor, 98,2% das vítimas são da raça negra (pretos e pardos); com metade da distribuição espacial concentrada em apenas 10 bairros, predominantemente Guamá, Cabanagem e Pedreira. Por fim um estudo da elucidação das investigações em relação a estes registros foi traçado, onde houve elucidação em: 44% dos casos cuja vítima era do sexo feminino, em comparação com uma taxa de apenas 19% para vítimas do sexo masculino; 17% das vítimas com grau de escolaridade até o ensino fundamental; uma concentração de 33% para vítimas de cor branca, em comparação com 23% e 0% para as cores parda e preta, respectivamente.

Um estudo regional propõe verificar a relação do crime de homicídio com fatores de vulnerabilidade social entre o pública de jovens na faixa de 15 a 19 anos para o período de 2008 a 2019, na Região Metropolitana de Belém (TRINDADE, 2019). Através de uma análise quantitativa - baseada em recursos estatísticos que exponham dados de indicadores e variáveis a serem observadas, tais como Análise de Correspondência e Análise de Componentes Principais - e de uma análise qualitativa - que permita interpretar conceitos e fenômenos existentes nos dados coletados -, foi inferida uma ausência de integração de esforços nas áreas da saúde, educação e segurança pública, além de apontar as causas e fatores de risco para a ocorrência de homicídios, sendo diretamente afetados por políticas públicas de educação, saúde, habitação, e emprego. Em especial na Região Metropolitana de Belém, ficou evidenciado pelos resultados que o público jovem do sexo masculino, solteiros com baixa escolaridade, estudantes, autônomos, sem ocupação, ou que fazem parte do mercado informal é mais propenso a ser atingido por crimes de caráter homicida.

Um estudo quantitativo exploratório foi conduzido para avaliar a criminalidade no Estado do Pará, nos anos entre 2017 e 2019 (REGATEIRO et al., 2021). Os boletins de ocorrência referentes aos crimes de furto, roubo, roubo de veículo, homicídio, latrocínio, lesão corporal seguida de morte, foram coletados e submetidos a uma análise exploratória que permitiu a visualização da informação através de gráficos, tabelas e medidas de síntese. Foram propostos então a Taxa de Criminalidade Média Bayesiana Ponderada Padronizada para Município (TCMBPPM) - objetivando a obtenção de indicadores de criminalidade por tipologia criminal - e o Índice de Criminalidade Média Bayesiana Duplamente Ponderada Padronizada para Município (ICMBDPPM) - objetivando mensurar criminalidade dos municípios. Resultados apontam que em 2019 grande parte dos municípios apresentaram baixos índices de criminalidade, sendo que em 2017 e 2018 os 5 primeiros colocados no ranking apresentaram índices muito altos. Todos estes municípios com alta taxa de crimes violentos possuem características socio-econômicas deficitárias, como em relação ao saneamento básico, urbanização, Índice de Desenvolvimento Humano Municipal (IDHM), e taxa de ocupação. O estudo também gerou visualizações geográficas a partir dos índices coletados.

A aplicação de ciência de dados na segurança pública no Estado do Rio de Janeiro é

executada sobre um conjunto de dados de evolução mensal das estatísticas por circunscrição das delegacias da Polícia Civil fluminense entre 2003 a 2018 (SOUZA, 2018). A metodologia consiste na integração, limpeza, e engenharia de atributos, além da análise exploratória dos dados que levantou indicadores para crimes violentos ao longo de todo o período - com especial atenção ao comportamento obtido durante a administração do então governo do estado. Também foram gerados gráficos de distribuição de crimes por municípios e regiões, além da visualização da correlação entre o total de crimes e a população.

Uma análise exploratória de dados conduzida sobre boletins de ocorrência restritos ao crime de roubo de celular na cidade de São Paulo entre 2010 e 2018 buscou encontrar padrões através de observações estatísticas (VARGAS, 2019). Também foram realizadas análises com enfoque inferencial, preditivo, e espacial. Foram coletados resultados que indicam que em ocorrências noturnas o autor do delito se locomove de moto, explicando cerca de 84% da variância do total de ocorrências de roubo de celular na cidade. Dentre os algoritmos de Regressão Logística, Árvore de Decisão, SVM, Rede Neural, Floresta Aleatória e Bagging, os dois últimos conseguem prever com sucesso cerca de 60,5% dos crimes em flagrante delito.

Uma proposta de uma abordagem de ciência de dados centrada na interação humana é capaz de extrair possíveis associações entre os padrões dos crimes, organizar *clusters* de crimes similares, e identificar redes de autores de crimes e listas de suspeitos baseadas em similaridades espaço-temporais e de *modus operandi* (QAZI; WONG, 2019). Foi utilizado um conjunto de dados de roubos, com cerca de 1,6 milhões de registros que incluem informações de autores e vítimas, coleto junto à agências de segurança pública do Reino Unido. Os dados de informações pessoais como nomes, números de referência, localização, e tempo, foram anonimizados através de técnicas de encriptação, sendo impossível ligar uma pessoa a amostra dos dados. Através da análise proposta é possível visualizar os padrões dos crimes em um espaço 2D com seleção dinâmica de atributos. Também é importante destacar que padrões relacionados a "porta de polivinil" e "janelas" é um *modus operandi* comum para crimes de roubo.

Um estudo sobre as relações entre os fatores socio-demográficos que impactam diretamente os crimes de assalto e arrombamento foi realizado em Kuala Lumpur, Malásia (CHIEW; AMERUDIN; YUSOF, 2020). Os dados incluem registros espaço-temporais de localização, data, horário, e perdas estimadas para os anos entre 2011 e 2016. Com o auxílio de ferramentas e provedores de geolocalização locais, esses dados foram complementados com atributos descritivos das rodovias e demarcação de terras. Utilizando o software de manipulação e análise geográfica ArcGIS, foi possível destacar um maior índice de arrombamentos em áreas residenciais da periferia, em comparação com apartamentos e condomínios, assim como em áreas distantes de delegacias, e onde há uma alta concentração de residentes com alto grau de educação e consequente maior acúmulo de bens, dentre outros fatores de raça, classe trabalhadora, fatores de imigração, e índices de residentes idosos.

3.2 Aprendizado de máquina

A seguir serão descritos trabalhos relacionados a algoritmos de aprendizado de máquina aplicados à segurança pública, totalizando 11 trabalhos encontrados em mecanismos de buscas.

A capital paraense, Belém, foi alvo da aplicação de mineração de dados para classificar crimes usando dados provenientes de redes sociais (FURTADO; SOUZA, 2019). A partir da extração, filtragem e análise estatísticas dos dados provenientes do Twitter entre janeiro de 2017 e janeiro de 2018, pré-processados ao longo das etapas de limpeza, integração, transformação, e redução dos dados. Uma análise exploratória consistindo de distribuição do quadro de crimes, distribuição de períodos do dia, distribuição ao longo dos dias da semana e dos meses, e distribuição ao longo dos bairros foi realizada em paralelo. Através da estratificação das classes de roubo, acidentes e homicídios, divisão de 80% para treino e 20% para teste, e uso da técnica de validação cruzada, o algoritmo de *Classification and Regression Tree* (CART) gerou 43,23% de acurácia na base de teste, além da geração visual das regras da Árvore de Decisão.

A tipificação de ocorrências policiais foi desenvolvida através da implementação e consequente comparação entre diversos algoritmos de aprendizado de máquina: C4.5, CART, *k-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM), Rede Neural, *Repeated Incremental Pruning to Produce Error Reduction* (Ripper), e Floresta Aleatória utilizando o ambiente computacional R (AMORIM; PEREIRA, 2019). O conjunto de dados consiste em 465376 amostras e 17 atributos, distribuídos através de 381 classes únicas, para uma base de ocorrências no estado da Bahia. Um pré-processamento que consistiu na remoção de símbolos especiais, remoção de *stopwords* (palavras de parada), remoção de valores ausentes, remoção de acentos, e transformação para caixa baixa foi realizada para padronizar a base. A avaliação dos algoritmos se deu pela técnica de *Hold-out* para diversos testes que variavam a seleção das amostras e dos atributos, obtendo picos de 99% de acurácia para o C4.5 e *Ripper*.

Desenvolvido como um mecanismo de visualização da aplicação de modelos de aprendizagem de máquina sobre boletins de ocorrência do Estado de São Paulo, foram levantados padrões sobre regiões mais expostas a manchas criminais, além de períodos do dia, tipos de veículos, locais, e marcas de celulares mais presentes em registros criminais (ASSAD; CHAGAS, 2019). Através da análise dos dados de roubo de celular, furto de veículo, roubo de veículo, homicídio doloso, lesão corporal, latrocínio, e intervenção policial no ano de 2018, com os modelos de Árvore de Decisão e Regressão Logística construídos no software *Orange*, foram extraídas diversas regras associativas categorizadas por crimes e regiões geográficas da cidade.

Com o objetivo de entender como se relacionam os indicadores de ocorrências de crimes contra a mulher, foi proposta uma solução através da descoberta de conhecimento em bancos de dados para o estado do Rio Grande do Sul (VINHOLES, 2019). Após as etapas de limpeza, integração, redimensionamento, e transformação, dois conjuntos de dados referentes a crimes contra a mulher e ao quantitativa de população feminina estadual foram mesclados e submetidos

ao algoritmo Apriori, para descoberta de regras de associação. Os resultados apontam relações entre índices de criminalidade contra a mulher: em 61% dos períodos onde as ocorrências de lesões foram baixas, as ameaças foram baixas; em 52% dos períodos em que houveram muitos estupros, ameaças foram baixas; em 50% dos períodos em que lesões corporais foram altas, ameaças foram baixas.

Outra análise comparativa entre algoritmos de aprendizagem de máquina utiliza fontes de dados heterogêneas (de diferentes fontes) para predição da tendência e número de ocorrências de crimes por tipo e por região geográfica (CASTRO, 2020). Os dados dos crimes de furto e roubo coletados para o Estado de Minas Gerais, foram submetidos às etapas de seleção, transformação, e análise de atributos, para então serem aplicados aos algoritmos de aprendizado *k-Nearest Neighbor* (k-NN), *Support Vector Machines* (SVM), Florestas Aleatórias, eXtreme Gradient Boosting (XGBoost) e *Long Short Term Memory* (LSTM). Os experimentos mostraram que a rede neural LSTM apresenta uma pequena vantagem sobre os demais, alcançando 91% de acurácia.

Afim de realizar uma avaliação sobre dados governamentais de segurança pública, foram realizados dois experimentos para bases de dados em âmbito nacional e no Estado de Minas Gerais, respectivamente (PRADO, 2020). Foram desenvolvidas uma arquitetura centralizada para as tarefas de ETL (Extração, Transformação e Carga) e mineração de dados, duas aplicações públicas para cada âmbito abordado no problema de pesquisa, além da extração de regras associativas entre crimes e estados, entre crimes e municípios, entre crimes e RISPs, entre alvos de roubo e municípios, e entre alvos de roubo e RISPs.

Situada no contexto do Estado de Pernambuco, uma abordagem não supervisionada para criação de *clusters* entre os municípios do estado foi desenvolvida a partir da análise de variáveis representativas de criminalidade (COSTA, 2020). Ao obter dados de ocorrências criminais de 2018 para os 185 municípios do estado, foi conduzido um estudo de redução de dimensionalidade através da técnica de Análise de Componentes Principais. De posse dos dados com dimensionalidade reduzida, o emprego do algoritmo *k-means* através da biblioteca *cluster* da linguagem R, permitiu obter a agregação dos municípios para diversos valores de k, onde foi observado um comportamento satisfatório para $k = 26$, uma coincidente referência às 26 Áreas Integradas de Segurança (AIS) do estado. Com isso foi possível destacar os três crimes mais representativos para cada *cluster* obtido, através da média móvel das ocorrências de cada crime para todos os municípios do *cluster*.

Um estudo de incidentes reportados entre 2013 e 2017 na cidade de Chicago-EUA foi realizado através da aplicação das técnicas de Naïve Bayes e Árvores de Decisão (ALDOSSARI et al., 2020). O experimento consistiu nas etapas de limpeza, extração de atributos - através do algoritmo de avaliação do ganho de informação em atributos -, e modelagem sob as técnicas citadas. A análise resultou em uma acurácia de 83,33% para Naïve Bayes e de 91,59% para Árvore de Decisão, indicando uma melhor performance desta última.

Utilizando um conjunto de dados dos crimes da cidade de Denver-EUA no período entre janeiro de 2014 e maio de 2019, foi realizado um estudo comparativo entre algoritmos de aprendizado de máquina (RATUL, 2020). Foram realizadas as etapas de limpeza, redução, integração, conversão, randomização, normalização, amostragem, e seleção de atributos dos dados. Quanto aos algoritmos utilizados, Floresta Aleatória, Árvore de Decisão, *K-Neighbor Classifier* (KNN), Análise de Discriminante Linear (LDA), *AdaBoost*, *ExtraTrees*, e quatro modelos diferentes de *Ensemble* foram avaliados por três estratégias diferentes: divisão treino-teste, validação cruzada, e teste t pareado. Os resultados indicaram que todos os algoritmos alcançaram acurácia satisfatória acima de 90%, exceto pelo *AdaBoost*. O destaque ficou para um dos métodos de *Ensemble* que manteve a acurácia citada para o maior número de classes utilizada, 15.

Uma comparação empírica entre um modelo de *ensemble* proposto e algoritmos de classificação unitários foi realizada sobre conjuntos de dados de crimes violentos na Índia (KSHATRI et al., 2021). Os dados consistem na coleta de registros de assassinatos, estupros, roubos, e outros crimes violentos, entre 2001 e 2015, submetidos às etapas de extração, transformação, e redução. O desempenho dos algoritmos unitários J48, SMO, Naïve Bayes, floresta aleatória e *bagging ensemble* foi comparado com o classificador *Modified Ensemble Stacking* proposto. Dentre os algoritmos unitários, *bagging ensemble* foi o melhor, alcançando acurácia de 95,55% na base de testes. Já o modelo de *staking* proposto alcançou uma acurácia de 99,5%, provando a eficácia de modelos *ensemble* sobre os unitários.

Situado em Pequim, China, um experimento para extração de atributos negligenciados em registros policiais buscou identificar potenciais padrões ofensivos (CHEN; KURLAND, 2018). A metodologia consiste na delineação de três variáveis de fundamental importância no campo da criminalística: tempo, local, e *modus operandi*. Ao analisar 363 registros de roubos de bicicletas, para os atributos de contexto, modus operandi e tempo, através do algoritmo Apriori, foram extraídas duas regras predominantes. A primeira descreve que de todos os crimes de roubo de bicicletas, 4% ocorreram pela manhã, sob o contexto de ruas, nas quais 67% destes foram cometidos com arrombamento de cadeados. A segunda regra descreve que de todos os crimes de roubo de bicicletas, 4,6% ocorreram durante a noite, com subtração total do objeto, onde 60% destes ocorrem em bairros residenciais.

4 METODOLOGIA

A seguir será descrita a metodologia de trabalho proposta. Serão utilizadas técnicas descritas pelo referencial teórico listado no Capítulo 2, em conjunto com a especificação dos recursos, parâmetros e etapas que foram seguidas no desenvolvimento do trabalho, a fim de serem gerados os resultados que atendam aos objetivos listados.

O desenvolvimento do classificador é apoiado na metodologia CRISP-DM, através de adaptações do modelo de referência original. Apesar da maioria dos guias descreverem a operação de cada fase, o modelo foi idealizado para ser flexível e personalizável, possibilitando com que suas fases sejam realocadas, priorizadas ou simplesmente omitidas de acordo com as necessidades do projeto.

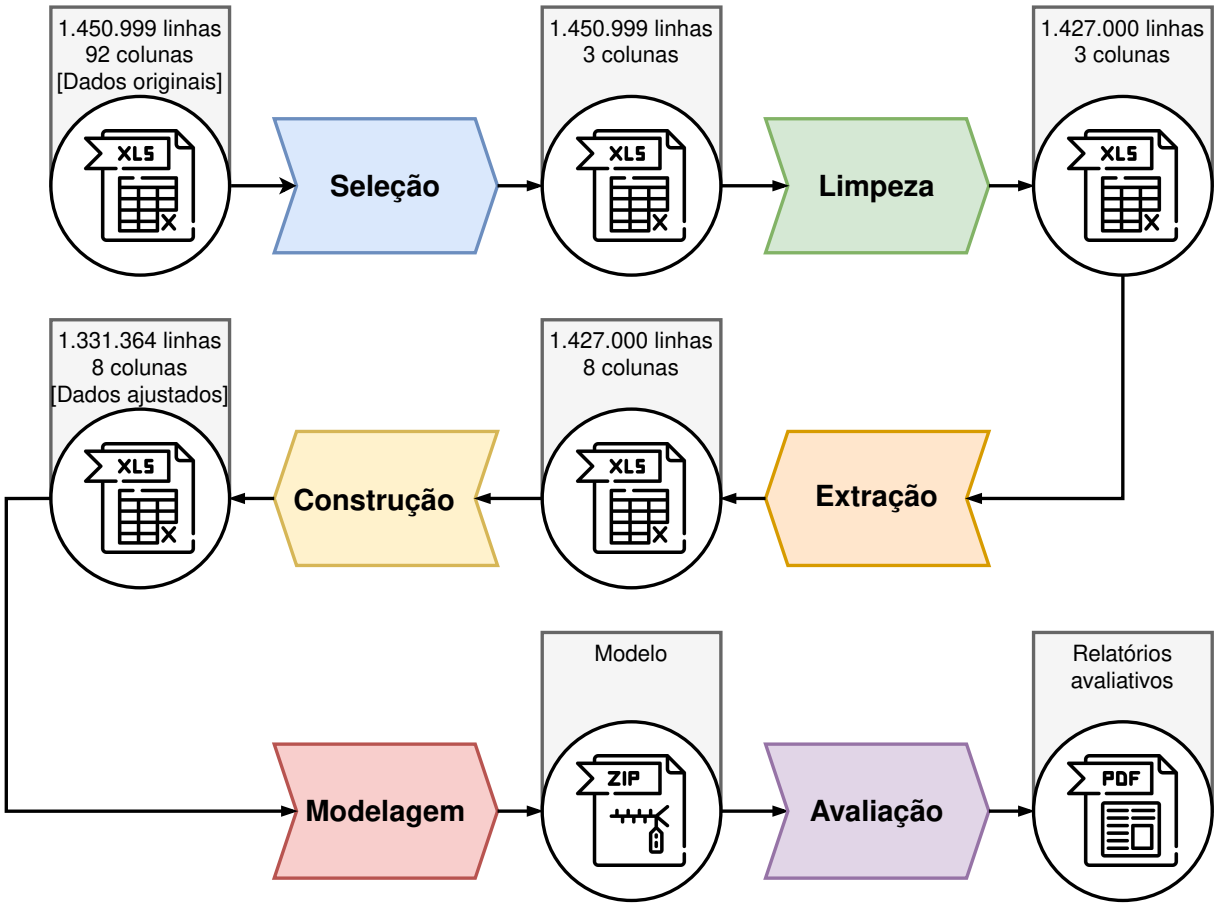


Figura 6 – Fases do modelo CRISP-DM adaptadas para o desenvolvimento do classificador.

A Figura 6 ilustra como o modelo CRISP-DM foi adaptado para se ajustar aos objetivos do trabalho, consistindo na execução sequencial de seis etapas que focam nas fases de preparação, modelagem e avaliação, originais do modelo citado. A fase de preparação foca na manipulação do conjunto de dados para atender os requisitos necessários para a modelagem e foi subdivida em seleção, limpeza, extração e construção, o que aumentou a modularidade da ferramenta. A modelagem realiza o ajuste dos modelos de classificação propostos para extrair o conhecimento do conjunto de dados processada, gerando arquivos com os modelos ajustados e outros arquivos

necessários para a reprodução e predição de novos resultados, como os objetos que representam as lista de classes aprendidas e o *tokenizador*, responsável pela vetorização das palavras em sequências de números. Por fim, a avaliação gera os relatórios avaliativos para aferição de performance dos modelos gerados.

O classificador foi desenvolvido com a linguagem de programação Python, versão 3.8.11, uma das tecnologias mais usadas em tarefas de manipulação e mineração de dados. Dentre as bibliotecas desta linguagens utilizadas para a implementação do classificador, é importante destacar:

- JupyterLab (v. 3.1.18): plataforma de desenvolvimento de documentos que combinam código-fonte com linguagem de marcação, formando ambientes interativos para a exploração de técnicas de ciência de dados e computação científica;
- Keras (v. 2.7.0): biblioteca que oferece uma interface modular, extensível e simplificada para a biblioteca TensorFlow;
- Matplotlib (v. 3.4.3): biblioteca e visualização da informação e criação de gráficos;
- NLTK (v. 3.6.5): biblioteca de suporte a recursos de Processamento de Linguagem Natural;
- NumPy (v. 1.21.2): biblioteca de análise numérica e suporte a arrays multi-dimensionais;
- Pandas (v. 1.3.3): biblioteca de análise e manipulação de dados. Oferece estruturas de dados e recursos de visualização de dados bidimensionais, com alta compatibilidade com a biblioteca Numpy;
- Plotly (v. 5.8.0): biblioteca de plotagem e visualização da informação, com foco na geração de gráficos interativos para aplicações web;
- Scikit-Learn (v. 1.0): biblioteca para aprendizado de máquina. Oferece módulos e algoritmos clássicos e extensíveis para pré-processamento e modelagem para tarefas de aprendizado;
- spaCy (v. 3.2.0): biblioteca de suporte a recursos avançados de Processamento de Linguagem Natural;
- TensorFlow (v. 2.7.0): biblioteca para aprendizado de máquina e inteligência artificial. Apesar de disponibilizar ferramentas clássicas para tarefas de inteligência computacional, possui ênfase na implementação de Redes Neurais Profundas.

Em relação ao hardware, foi utilizada uma estação com as seguintes especificações:

- Sistema Operacional: Windows Server 2008 R2 Standard (64 bits)

- CPU: 24x Intel(R) Xeon(R) CPU E5-2620 0 @2.00GHz
- RAM: 36 GB

A seguir serão descritas a forma de coleta dos dados e as técnicas aplicadas nas seis fases adaptadas do modelo de referência para o desenvolvimento do classificador.

4.1 Coleta dos dados

Quanto à colheita dos dados usados na análise, a SIAC disponibilizou um conjunto de registros policiais para os anos de 2019, 2020 e 2021. Armazenados em arquivos de formato tabular que totalizam cerca de 1,06 GigaBytes, os dados originais combinados contém 1.450.999 amostras e 92 atributos. Estes são formados principalmente por colunas de identificação do registro, tempo, localização, descrição do *modus operandi*, informações sobre vítimas e autores, e o relato descritivo do evento. Quanto às colunas de tipologia criminal, é importante destacar os seguintes atributos:

- "registros": classes atribuídas nas delegacias no momento do registro do boletim de ocorrência. Devido ao seu preenchimento emergencial, nem sempre reflete com precisão o evento presente no relato descritivo, nem são normalizadas, tendo casos onde as classes remetem ao Código Penal Brasileiro (ex. "ART. 147 - AMEAÇA"), e até mesmo subconjuntos de classes com erro ou duplicidade de grafia que podem ser agrupadas em uma única (ex. "ART. 157 - FURTO", "ARTIGO 157 - FURTO", "FUERTO", "FURTO", etc).
- "consolidado": classes atribuídas na base de dados da SIAC, como uma alternativa mais precisa e confiável em relação às classes de registros. Usado principalmente na geração de relatórios estatísticos que chegam nos gestores da segurança pública no estado do Pará, na mídia interessada nos dados quantitativos e nos portais de transparência da secretaria.

Em relação à operacionalidade interna da SIAC, a coluna de consolidados surge como uma solução para a baixa consistência da coluna de registros, cujo uso direto poderia comprometer as demandas estatísticas da secretaria. O departamento de estatística da SIAC aloca um grupo de 15 analistas criminais para ler e rotular os relatos em uma classe específica, de acordo com seus conhecimentos empíricos da legislação brasileira e, em casos complexos, do julgamento coletivo destes analistas, gerando então as classes de consolidados. Atualmente, essa análise é feita sobre um conjunto selecionado de leitura de relatos, focando principalmente em crimes violentos (homicídio, roubo, estupro, etc.), devido ao grande volume de dados diários gerados nas delegacias. Essa metodologia indica que cerca de 10% das classes de registros não são compatíveis com as classes finais de consolidados, uma margem de erro que é o principal foco de alocação de esforços para melhoria da qualidade dos dados.

Como resultado das demandas da secretaria por uma ferramenta automática de confirmação das classes de eventos criminais para processamento do volume massivo de registros policiais diários, em conjunto com a necessidade de definir um conjunto descritivo das classes de relatos para o problema de pesquisa, o atributo de consolidado foi escolhido como a coluna-alvo.

Informações mais completas sobre os 92 atributos da base de dados original estão presentes no Apêndice B - Dicionário de Dados.

4.2 Fases de desenvolvimento do classificador

4.2.1 Seleção

A etapa de seleção decide quais dados serão usados para análise, baseado na relevância destes para os objetivos do projeto, através da seleção de atributos (colunas) e amostras (linhas). Apenas três colunas foram julgadas como importantes para o desenvolvimento do classificador:

- "nro_bop": é o identificador de cada boletim, que será usado durante a etapa de limpeza dos dados. Possui dois padrões de formatos, para os dois sistemas de bancos de dados do Sistema Integrado de Segurança Pública (SISP) do estado do Pará, 00000/0000.000000-0 ou 000/0000.000000-0, onde os dígitos antes da barra identificam a unidade seccional de registro do boletim, (cujo comprimento depende do banco de dados de armazenamento), os quatro dígitos seguintes identificam o ano de registro, e os últimos sete dígitos são gerados pelo sistema interno de registro. Apesar de ser um identificador das instâncias, não é um valor único, já que pode ser duplicado para crimes coletivos[1];
- "relato": é a descrição completa do evento, sendo a principal fonte de extração de conhecimento da base de dados, através da modelagem das palavras presentes. A quantidade de palavras (*tokens*) de cada relato foi calculada através do fatiamento do texto em cada caractere de espaço único, restando portanto contar a quantidade de elementos da coleção resultante, o que gerou as seguintes estatísticas descritivas:
  - Média: $\bar{x}_{tokens} = 121,8$;
  - Desvio padrão: $\sigma_{tokens} = 78,5$;
  - Quartis: $Q_1 = 72$, $Q_2 = 101$ e $Q_3 = 148$;
  - Relato mínimo[2]: $min_{tokens} = 1$;
  - Relato máximo: $max_{tokens} = 2244$.

[1] Exemplos de crimes coletivos: roubos, motins, estupros e ameaças. Muitos desses casos coletivos são mapeáveis por características de tempo, local e autoria.

[2] Alguns dos boletins duplicados de crimes de natureza coletiva recebem apenas um caractere no seu campo de relato, geralmente um ponto (".").

- "consolidado": é a classe dada para a amostra por um analista criminal após a leitura de seu relato. É o atributo-alvo para o problema de classificação, devido à sua descrição precisa e sucinta dos relatos da base. A base de dados original contém 977 classes, cuja distribuição é desbalanceada, um problema que é minimizado através das próximas etapas de preparação dos dados.

Ao final da seleção, os dados processados contém 1.450.999 instâncias e 3 atributos.

4.2.2 Limpeza

A etapa de limpeza eleva a qualidade dos dados para tornar a aplicação de técnicas de mineração de dados viável. Consiste da seleção de subconjuntos limpos de dados através de:

- Remoção de *outliers*: os dados defeituosos são definidos de duas formas. Primeiro é feita a exclusão de amostras cujo valor de "nro_bop" não siga os padrões dos identificadores únicos, já que alguns dos caracteres de separação dos dígitos podem estar ausentes ou erroneamente posicionados, dificultando suas identificações. Em seguida são calculados os comprimentos de cada relato e armazenados em uma coluna chamada "token_count", armazenada em uma segunda tabela com os respectivos "nro_bop", a fim de estabelecer uma quantidade mínima de palavras suficiente para a extração de conhecimento. Considerando que a média da quantidade de palavras $\bar{x}_{tokens}$ e seu desvio padrão σ_{tokens} são conhecidos, foi estabelecido que ao menos 20 palavras são necessárias para uma boa compreensão do evento, fazendo com que as amostras abaixo desse limite fossem eliminadas da modelagem. Este limite foi calculado a partir da seguinte expressão:

$$tokens\_necessários = \bar{x}_{tokens} - \sigma_{tokens} \times 1,3 = 121.8 - 78.5 \times 1,3 \approx 20$$

- Remoção de amostras duplicadas: a base de dados possui valores duplicados, onde crimes com múltiplas vítimas/autores de um único evento possam ser encontrados em boletins diferentes. A coluna "nro_bop" é o identificador único dos boletins de ocorrência e permite que as duplicatas sejam eliminadas, mantendo a primeira ocorrência. Dessa forma, duplicidades nas colunas "nro_bop" e "relato" são removidas, mantendo apenas a primeira amostra encontrada.

- Remoção de classes com poucas amostras: considerando a proposta de dividir a base nos subconjuntos de treino, validação e teste durante a etapa de modelagem, a quantidade mínima de amostras suficientes para a construção do classificador foi fixada em 6, onde 4 serviriam para treino e as demais para validação e teste. Tal limite foi mantido baixo o suficiente para acomodar a maior quantidade de classes, mesmo que prejudicando o desempenho destas raras, em vista da dificuldade de determinar quais classes seriam realmente importantes apesar de suas baixas frequências.

Ao final da limpeza, os dados processados contém 1.427.000 instâncias e 3 atributos.

4.2.3 Extração

A etapa de extração aplica técnicas de engenharia de atributos para descoberta de características importantes nos dados existentes. O principal indicativo da significância dos relatos é a quantização de seus elementos textuais, tais como palavras, caracteres e sentenças, além de métricas estatísticas destas contagens. O intuito dessa quantização é separar relatos altamente descritivos - geralmente crimes violentos - daqueles que necessitam de poucas palavras. Ao final desta etapa, foram adicionadas as seguintes colunas (considerando que "token_count" também é calculada na limpeza):

- "token_count": quantidade de palavras no relato.
- "char_count": quantidade de caracteres no relato.
- "avg_token_length": média de caracteres por palavra, dada por $\frac{char\_count}{token\_count}$.
- "seq_count": quantidade de sentenças no relato, separadas por vírgula ou ponto.
- "avg_seq_length": média de palavras por sentença, dada por $\frac{token\_count}{seq\_count}$.

Ao final da extração, os dados processados contém 1.427.000 instâncias e 8 atributos.

4.2.4 Construção

A etapa de construção conclui a preparação dos dados, produzindo atributos derivados, novos registros, ou valores transformados para atributos existentes, gerando a forma final dos dados submetidos à etapa de modelagem. As alterações sobre a coluna de relatos incluem:

- Transformação dos relatos para caixa baixa, normalizando todos os caracteres para minúsculo;
- Remoção de *tags* HTML: o registro dos relatos na base de dados original contém elementos de marcação provenientes da página web onde o sistema é hospedado, o que acumula muitos caracteres indesejados. Tais padrões de *tags* HTML foram identificados e eliminados com o uso de expressões regulares.
- Remoção de espaços múltiplos;
- Remoção de pontuação e caracteres especiais;
- Remoção de acentos e sinais diacríticos;

- Remoção de palavras de parada (*stopwords*), uma lista de palavras que não trazem significância ao contexto de segurança pública, tais como artigos, conjunções, preposições, etc., restando então apenas palavras funcionais como substantivos e verbos;

- Codificação de informações sensíveis: dados de identificação pessoal de vítimas e autores, instituições e localidades podem trazer vieses sobre a análise das palavras do relato, como relacionar a ocorrência de um crime a um nome próprio específico ou associar uma instituição ou logradouro como determinante na classificação de certas classes. A identificação destas informações em uma amostra de texto pode ser feita através de expressões regulares, ao procurar padrões conhecidos (ex. RG, CPF, número de telefone celular, e-mail, placa de carro, etc), ou por análise gramatical provida por módulos de processamento de linguagem natural capazes de especificar a classe gramatical de cada palavra (ex. palavras classificadas como substantivos próprios geralmente designam o nome de uma pessoa ou instituição). Uma remoção destas palavras eliminaria o problema de atributos sensíveis determinísticos, porém acarretaria em lacunas nas estruturas semânticas e sintáticas do relato, através da quebra de relações com palavras vizinhas. Dessa forma, os termos identificados como sensíveis foram substituídos por etiquetas equivalentes e padronizadas para minimizar o efeito da ausência dessas palavras (ex.: RG, CPF, LICENSEPLATE, INST, LOCAL, etc);

- Lematização: transformação de cada palavra em seu radical morfológico (ex. “correndo”, “corre”, “correu”, etc., se reduzem a “correr”). Tal transformação é importante para a tarefa de vetorização de palavras da modelagem, onde um conjunto de flexões reduzidas a uma única palavra é mapeada para um único índice, reduzindo portanto a variação dos atributos numéricos a serem explorados.

Já as alterações sobre a coluna de consolidados incluem:

- Remoção de classes indesejadas: uma lista de classes indesejadas é fornecida, incluindo amostras como “JA LANCADO” (amostras duplicadas de um mesmo evento já consolidado com identificador diferente), “PREJUDICADO” (para quaisquer relatos defeituosos), e aquelas com grafia incorreta e sem possibilidade de ajuste, por exemplo.

- Agrupamento de classes similares: um mapeamento de classes similares é fornecido, para substituição de uma classe por outra, como por exemplo é o caso de “ROUBO DE VEICULO”, que a partir de 2022 passa a ser consolidado como “ROUBO” pela base de dados da SIAC (especificidades do objeto roubado são listadas em outras colunas);

Ao final da construção, os dados processados contém 1.331.364 instâncias e 8 atributos, divididos em 463 classes de consolidados. Uma lista completa das classes geradas na etapa de construção e que foram utilizadas na modelagem está presente no Anexo A - Lista de Consolidados.

4.2.5 Modelagem

A etapa de modelagem é responsável pela execução da ferramenta de modelagem escolhida sobre o conjunto de dados processados para criação de modelos preditivos. Um objeto de "*tokenização*" captura todas as palavras presentes nos dados e mapeia cada uma para um índice único. Em seguida, os textos são transformados em sequências de números limitados em um comprimento máximo de sequência, a fim de permitir sua vetorização. Já as classes de consolidados são transformadas em colunas-alvo identificadas por um único número, através da técnica de *One-Hot Encoding*. As 1.331.364 amostras são então dividas em 3 subconjuntos: treino (70%), validação (15%) e teste (15%), estratificando as classes ao longo dos três conjuntos. Finalmente, o modelo proposto, ilustrado na Figura 7, foi ajustado para os dados de treino, iniciando a captura de conhecimento.

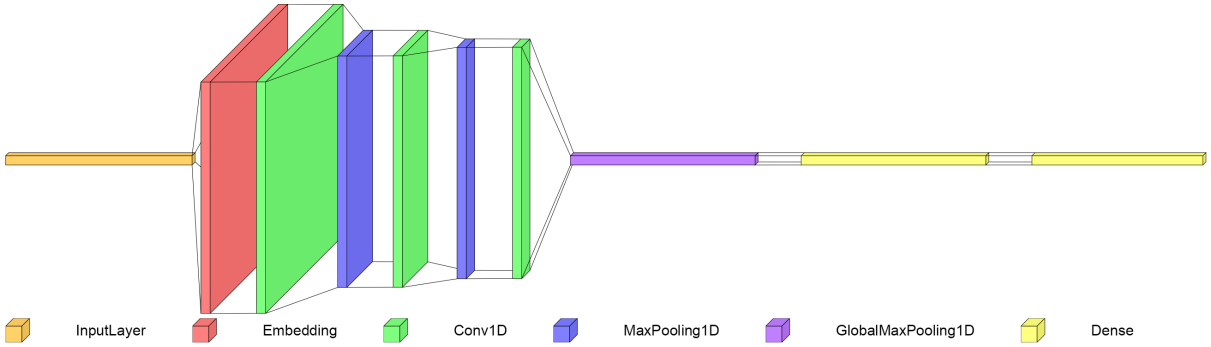


Figura 7 – Camadas do modelo de aprendizado profundo proposto.

Quanto à escolha das camadas utilizadas, (KIM, 2014) propõe o uso de Redes Neurais Convolucionais (*Convolutional Neural Networks* - CNN) na classificação de sequências de texto, onde uma operação de convolução é aplicada por um filtro sobre janelas locais de dados em um vetor de palavras, gerando um mapa de características de padrões cujo valor máximo pode ser passado a uma camada de ativação, produzindo os vetores de saída para os propósitos da classificação. De forma similar, (CHOLLET, 2018) aborda como as arquiteturas CNN são computacionalmente mais baratas e competitivas em comparação com modelos usualmente encontrados em esquemas de classificação de texto, como por exemplo as Redes Neurais Recorrentes (*Recurrent Neural Networks* - RNN). Este último autor também destaca as vantagens de usar *word embeddings* sobre métodos tradicionais de vetorização de dados textuais (como o *One-Hot Encoding*), criando densos vetores de palavras usando uma abordagem semanticamente coerente, onde palavras similares estão próximas. Dessa forma, a arquitetura proposta é composta do seguinte esquema, cujos parâmetros estão expostos na Tabela 1:

- Uma camada de *embedding* que redimensiona os dados de entrada em um tensor denso tridimensional, capaz de entender as relações semânticas entre as palavras, configurada com uma dimensão de ordem 500 e limitada a 500 palavras por sentença. Dada a natureza do problema proposto, onde um grande número de palavras são exclusivas ao ambiente

policial, a camada de *embedding* foi ajustada do zero, ao invés de utilizar modelos pré-treinados, tais como o GloVe e word2vec. A comparação entre as *embeddings* treinadas e pré-treinadas pode ser explorada em publicações posteriores.

- Uma sequência de camadas de convolução de dimensão única e camadas de *max pooling* foram capazes de extrair mapas de características na janela deslizante de valores do vetor de *word embedding*, em conjunto com operações de redução de dimensionalidade.
- Um par de camadas densamente conectadas ativam os neurônios correspondentes a cada uma das 463 classes, gerando a distribuição de probabilidades das predições.

Tabela 1 – Camadas e parâmetros da arquitetura proposta.

Camada	Dimensão de saída	Parâmetros
InputLayer	[(None, 505)]	0
Embedding	(None, 505, 500)	250.000.000
Conv1D #1	(None, 503, 500)	750.500
MaxPooling1D #1	(None, 167, 500)	0
Conv1D #2	(None, 165, 500)	750.500
MaxPooling1D #2	(None, 55, 500)	0
Conv1D #3	(None, 53, 500)	750.500
GlobalMaxPooling1D	(None, 500)	0
Dense #1	(None, 500)	250.500
Dense #2	(None, 463)	231.963
Parâmetros totais: 252.733.963		

O aprendizado foi configurado para executar durante uma única execução de 30 épocas, com uma opção de interrupção após decorridas 10 épocas sem melhorias na perda de validação.

4.2.6 Avaliação

Com a geração dos modelos preditivos, a etapa de avaliação é responsável pela estimação do nível em que os modelos atendem aos objetivos do projeto. Os resultados da avaliação serão analisados no capítulo seguinte, sob as perspectivas de treinamento, conjunto de teste e produção.

4.2.6.1 Treinamento

Consiste na análise do ajuste do modelo proposto para a base de dados processada sobre as 30 épocas configuradas na fase de modelagem. Este modelo foi ajustado para uma base de treinamento contendo 931.954 instâncias - 70% dos dados pós-construção - e 505 atributos - 500 relativos ao comprimento máximo das sequências de palavras vetorizadas e 5 relativos aos atributos estatísticos extraídos dos relatos. Ao fim de cada época, o estado do modelo foi avaliado pelo conjunto de validação, composto por 199.705 instâncias (15% dos dados pós-construção) e 505 atributos. A métrica de perda de validação (*validation loss*) foi utilizada como parâmetro de determinação do melhor modelo gerado dentre as épocas de treinamento.

4.2.6.2 Conjunto de testes

Consiste na avaliação do melhor modelo gerado sobre um conjunto de testes, criado durante a fase de modelagem, produzindo os indicativos de desempenho da metodologia proposta, através do confronto entre os valores da coluna-alvo esperados e as predições do classificador. Os indicativos observados descrevem o comportamento geral da avaliação dos testes ou o comportamento individual de cada classe para as métricas descritas na Seção 2.4. Uma análise sobre as classificações incorretas também é levantada, buscando encontrar os motivos que levam uma instância a ser predita como uma classe que não era a esperada.

4.2.6.3 Produção

Consiste na análise do comportamento do melhor modelo gerado em um ambiente de produção, através da implantação de uma ferramenta de consolidação de boletins de ocorrência na SIAC, consistindo do processamento dos relatos, da geração da coluna de consolidado, e da confirmação dos dados rotulados pelas delegacias policiais em relação aos relatos textuais.

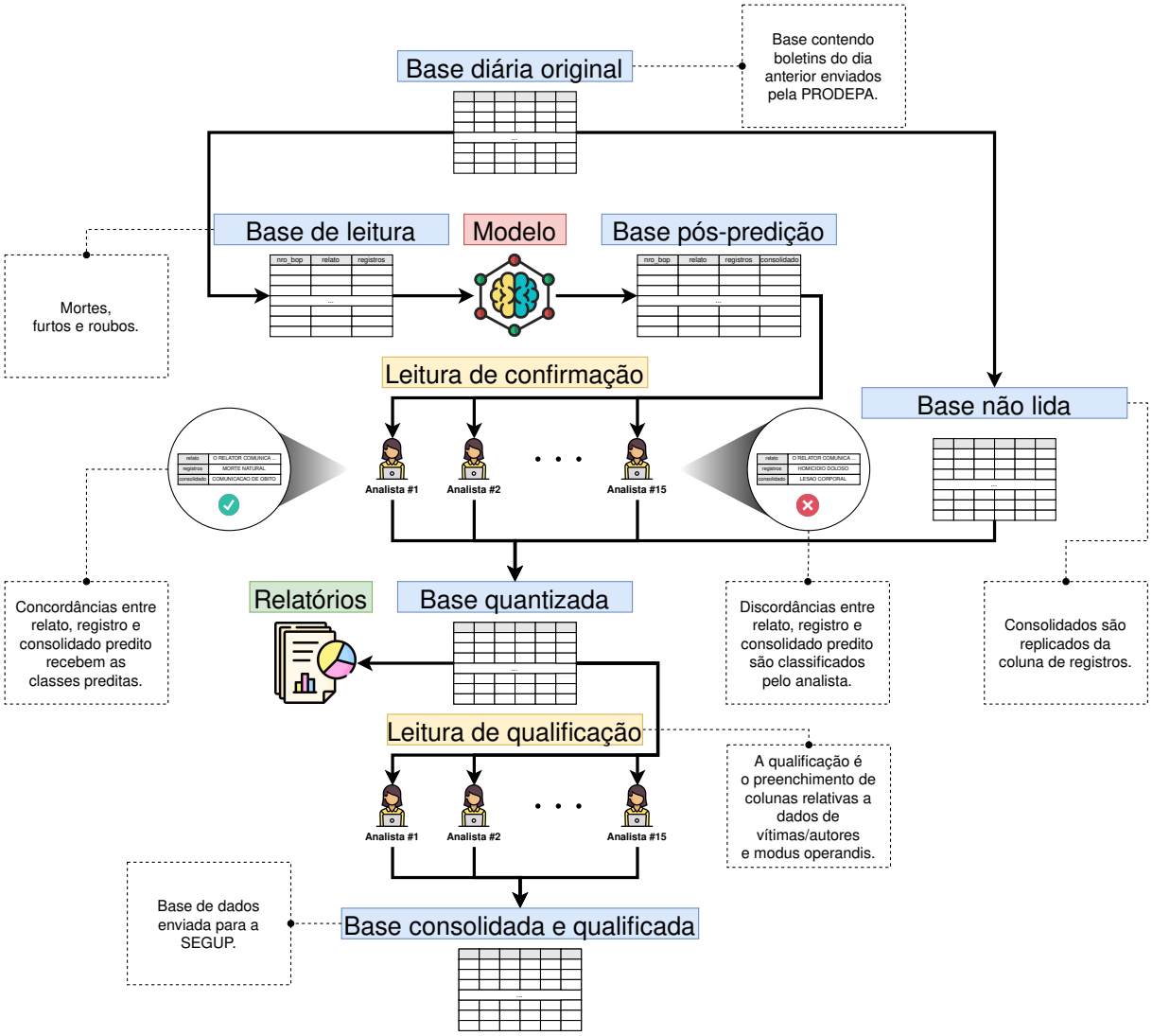


Figura 8 – Metodologia de produção completa.

A metodologia de implantação do classificador combinada com a consolidação dos dados pela SIAC está ilustrada na Figura 8. Apenas o conjunto de leitura passado para análise humana é passado para predição, consistindo em boletins cujas colunas de "relatos" e "registros" estejam associados à mortes em geral[3], furtos e roubos (selecionados previamente através de filtros e ferramentas de pesquisa de strings desenvolvidos pela secretaria), dada a baixa precisão do modelo para a maioria das amostras não lidas pelos analistas.

Uma comparação entre as predições do classificador e a classe de registro indica um conjunto de predições incorretas, que são então passadas para leitura dos analistas. Já para as predições corretas, a maioria dos relatos são atribuídos para a classe confirmada pelo classificador, exceto para as mortes em geral que requerem uma análise humana mais profunda. Dessa forma, o modelo estima uma quantidade de amostras prontas para serem consolidadas e destaca as instâncias que necessitam de atenção. Essa base cujos consolidados podem ser quantizados são passados para a divisão de estatística, a fim de iniciar os trabalhos que necessitam destes índices. Por fim, os analistas retomam a leitura para qualificar os dados, preenchendo informações de vítimas e/ou autores e *modus operandis*.

[3] *Homicídio*, *Homicídio Culposo*, *Comunicação de Óbito*, *Morte no Trânsito*, *Suicídio*, *Morte por Intervenção de Agente do Estado*, *Lesão Corporal Seguida de Morte*, *Latrocínio*, *Morte a Esclarecer com Indício de Crime*, *Morte a Esclarecer sem Indício de Crime*, ou qualquer outro consolidado que caracterize a morte de uma pessoa.

5 RESULTADOS

A seguir, serão expostos os resultados dos três formatos de avaliação abordados na metodologia, juntamente com suas discussões.

5.1 Treinamento

Dentre as diversas arquiteturas e parâmetros testados, o modelo com melhor performance empírica tem a evolução de seu treinamento mostrada na Figura 9, onde o melhor valor da perda de validação foi encontrado já na segunda época, com perda de validação $val\_loss = 0,7042$ e acurácia de validação $val\_accuracy = 0,8105$. Como o ajuste foi configurado para ser interrompido após 10 épocas sem melhorias na perda de validação, o treinamento foi interrompido na 12ª época. As 931.954 amostras do conjunto de dados foram divididos em *batches* de 256 amostras cada, explorados ao longo de 3641 passos, cada um sendo processado em média dentro de 13 segundos. Desta forma, cada época demorou em média 12 horas e 50 minutos para ser completada, totalizando 6 dias, 10 horas e 9 minutos para as 12 épocas.

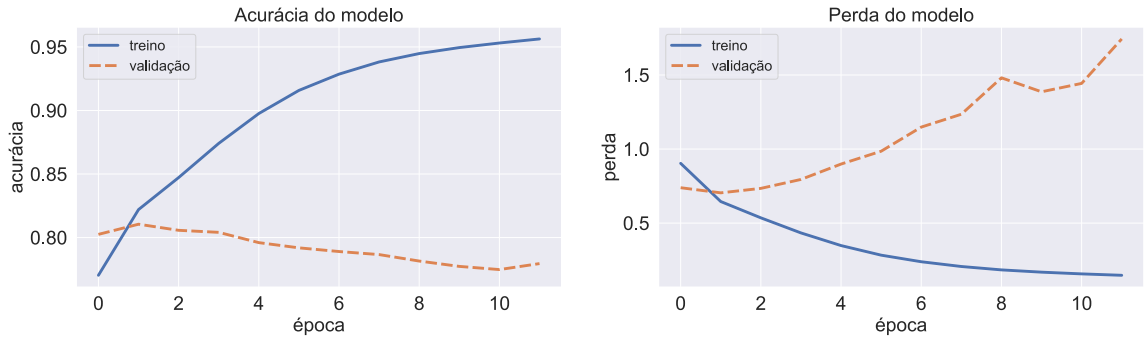


Figura 9 – Evolução da acurácia e perda através das épocas de treinamento.

5.2 Avaliação do conjunto de teste

O melhor modelo foi avaliado usando um conjunto de teste de 199.705 amostras (15% dos dados pós-construção), alcançando uma acurácia geral de 77,89%. Uma matriz de confusão de 463×463 descreve o comportamento geral das predições para o conjunto de teste. Dada a alta dimensão desta matriz e a natureza desbalanceada dos dados, a exposição dos resultados se torna bastante poluída.

A Figura 10 é composta das matrizes de confusão individuais para 16 classes com interesse prioritário para os gestores de segurança pública paraenses, especialmente relacionados a crimes contra a pessoa, contra o patrimônio e contra a dignidade sexual: *Comunicação de Óbito*, *Homicídio*, *Tentativa de Homicídio*, *Homicídio no Trânsito*, *Morte no Trânsito*, *Lesão Corporal Seguida de Morte*, *Morte por Intervenção de Agente do Estado*, *Suicídio*, *Furto*, *Tentativa de Furto*, *Roubo*, *Tentativa de Roubo*, *Ameaça*, *Lesão Corporal*, *Estupro* e *Estupro de Vulnerável*.

Tomando como exemplo a primeira matriz, para *Comunicação de Óbito*, são observados dois tipos de predições dentre seus 1109 exemplos positivos, presentes nos dois primeiros quadrantes:

- *True Positives* (TP): amostras de *Comunicação de Óbito* corretamente classificadas como *Comunicação de Óbito*, totalizando cerca de 96,4% das instâncias positivas;
- *False Negatives* (FN): amostras de *Comunicação de Óbito* incorretamente classificadas como qualquer outra classe, totalizando cerca de 3,6% das instâncias positivas.

Ainda em relação à *Comunicação de Óbito*, são observados dois tipos de predições para os demais 198.596 exemplos negativos, expostos nos dois últimos quadrantes:

Figura 10 – Matrizes de confusão individuais para 16 classes de interesse.

- *False Positives* (FP): amostras de quaisquer outra classe incorretamente classificadas como *Comunicação de Óbito*, totalizando cerca de 0,1% das instâncias negativas;
- *True Negatives* (TN): amostras de quaisquer outra classe corretamente classificadas como quaisquer outra classe, totalizando cerca de 99,9% das instâncias negativas.

Dentre as 16 classes de interesse, *Comunicação de Óbito*, *Furto*, *Roubo* tiveram uma alta taxa de TP, alavancados principalmente pelas suas grandes quantidades de exemplos de treino, enquanto que *Suicídio* é altamente separável dentre as classes relativas a morte, dada a condição de dano auto-infligido. Por outro lado, *Lesão Corporal Seguida de Morte* teve todas as suas amostras incorretamente classificadas, evidenciando tanto o conjunto de palavras que esta classe tem em comum com *Homicídio* e *Lesão Corporal*, quanto os poucos exemplos explorados no treino. *Tentativa de Homicídio*, *Tentativa de Furto* e *Tentativa de Roubo* tiveram baixo desempenho, destacando uma dificuldade em identificar a falta de consumação destes atos. *Estupro* e *Estupro de Vulnerável* possuem relatos similares, sendo diferenciados por fatores como idade ou discernimento para a prática do ato. As demais classes representam crimes contra a pessoa, apresentando subconjuntos de palavras comuns que podem limitar sua separação.

Tabela 2 – Métricas de avaliação para as classes de interesse.

Classe	ACC	TPR	TNR	PPV	F1-score	MCC
Ameaça	0.971167	0.854725	0.981156	0.795528	0.824065	0.809002
Comunicação de Óbito	0.992744	0.843945	0.996033	0.824661	0.834192	0.830541
Estupro	0.998363	0.322086	0.999468	0.49763	0.391061	0.399573
Estupro de Vulnerável	0.997011	0.707071	0.998746	0.77135	0.737813	0.737017
Furto	0.969019	0.933444	0.979066	0.926429	0.929923	0.910048
Homicídio	0.993986	0.877242	0.997236	0.898334	0.887663	0.884639
Homicídio no Trânsito	0.994517	0.444531	0.99809	0.601891	0.511379	0.514601
Lesão Corporal	0.980171	0.785723	0.988937	0.762017	0.773688	0.76342
LCSM[1]	0.999569	0.185185	0.99979	0.192308	0.188679	0.188498
Morte no Trânsito	0.993876	0.544631	0.995622	0.325851	0.407748	0.41843
MIAE[2]	0.999004	0.771493	0.999508	0.776765	0.77412	0.773625
Roubo	0.98272	0.948671	0.989503	0.947386	0.948028	0.937665
Suicídio	0.998989	0.773707	0.999513	0.787281	0.780435	0.779958
Tentativa de Furto	0.998768	0.247525	0.999529	0.347222	0.289017	0.292565
Tentativa de Homicídio	0.996685	0.52381	0.997932	0.400291	0.453795	0.456282
Tentativa de Furto	0.998052	0.388571	0.999122	0.437299	0.411498	0.411244

Um grupo de métricas de avaliação para as classes de interesse foram coletadas da matriz de confusão geral, expostas na Tabela 2. *Roubo*, *Furto*, e *Homicídio* estão entre as classes com melhor performance preditiva, como indicado por seus valores de sensibilidade, especificidade e precisão sendo superior a 85%. Os índices de MCC também indicam que as mesmas três classes citadas apresentam uma performance superior em relação às demais.

[1] Lesão Corporal Seguida de Morte
[2] Morte por Intervenção de Agente do Estado

Uma visualização alternativa da matriz de confusão geral é ilustrada na Figura 11. Um diagrama de cordas expressa relações de muitos-para-muitos, onde cada entidade é representada por um arco de circunferência e suas relações são ligadas por laços. O diagrama expressa predições incorretas através dos relacionamentos entre pares de classes, onde vale destacar:

- *Roubo* e *Furto*: crimes contra o patrimônio, dependem do grau de violência;
- *Ameaça* e *Lesão Corporal*: crimes contra a pessoa, dependem do grau de violência;
- *Calúnia*, *Injúria* e *Difamação*: crimes contra a honra, dependem do tipo de ofensa;
- *Consumo Pessoal de Drogas* e *Tráfico de Drogas*: dependem da comercialização da droga.

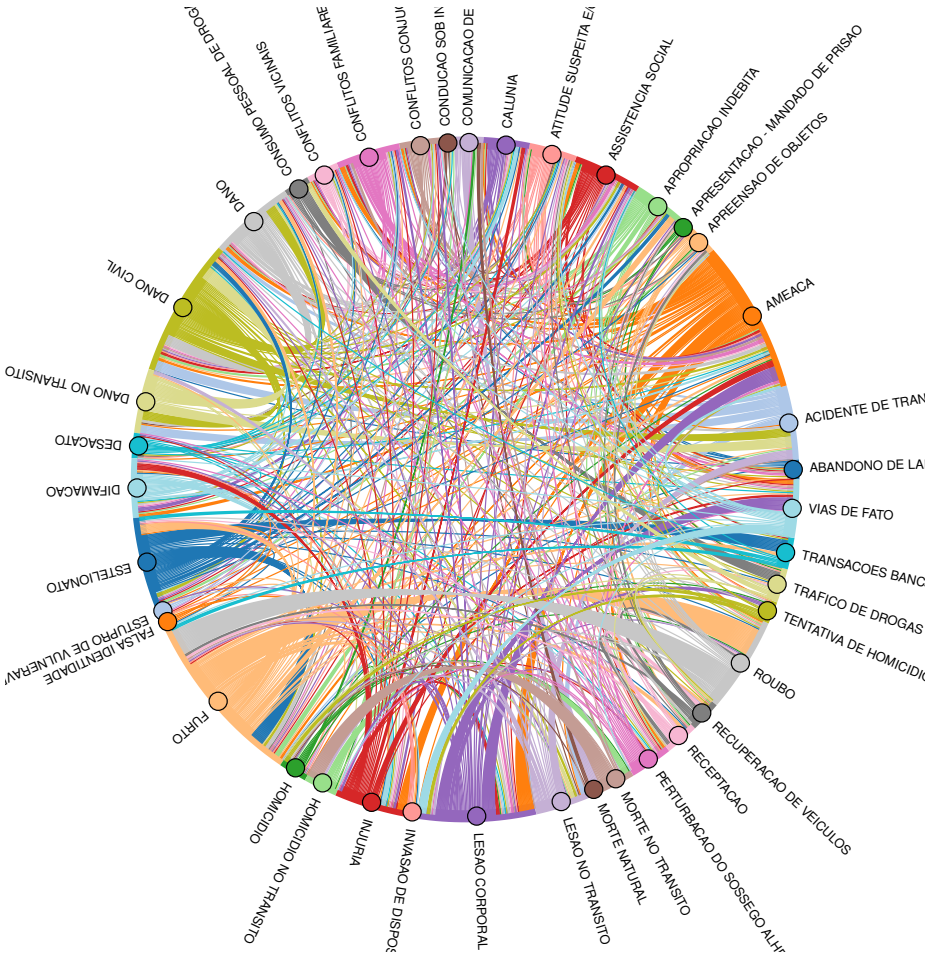
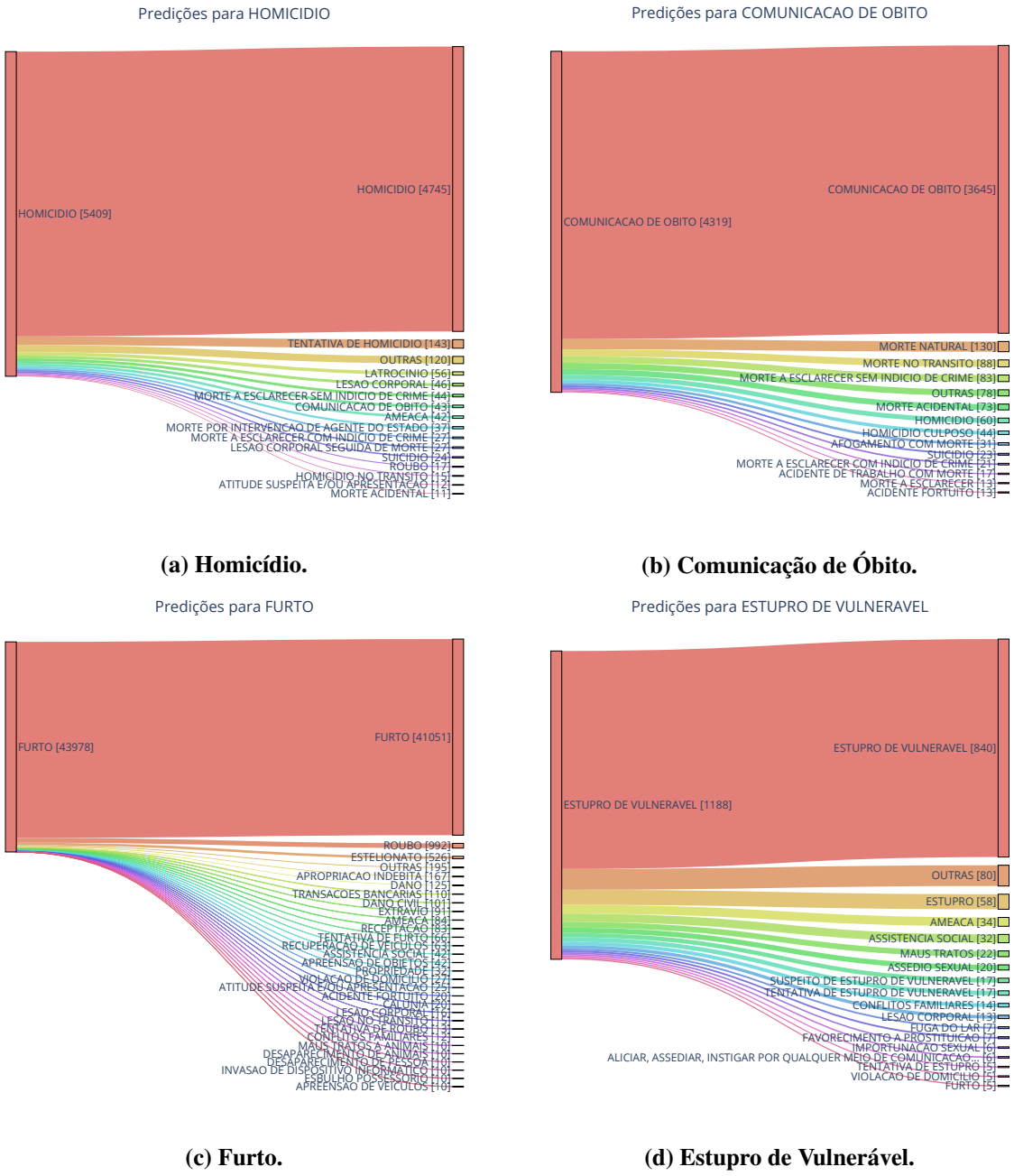


Figura 11 – Predições incorretas entre classes frequentes.

Outra visualização alternativa, agora para as classes individuais, é expressa na Figura 12. Um diagrama *Sankey* representa os fluxos de dados entre um nó-fonte e um nó-alvo. Dessa forma, cada diagrama ilustra os fluxos de predição para uma determinada classe. Essa visualização destaca não apenas as proporções de relatos corretamente classificados, mas também as classes que usualmente são encontradas em predições incorretas, o que implica nas chances de os relatos

desta classe ter palavras em comum com os relatos da classe esperada. Por exemplo, *Homicídio* é associado com crimes violentos contra a pessoa, *Comunicação de Óbito* com mortes por causas naturais e acidentes, *Furto* com crimes contra o patrimônio, assim como *Estupro de Vulnerável* com crimes contra a dignidade sexual. Para efeitos de compactação de tamanho dos diagramas, as classes com poucas ocorrências foram agrupadas em uma classe chamada *Outras*.



(a) Homicídio. **(b) Comunicação de Óbito.**

(c) Furto. **(d) Estupro de Vulnerável.**

Figura 12 – Fluxo de predições para quatro classes de consolidados.

5.3 Avaliação do modelo em produção

Após a construção de um modelo computacional robusto e capaz de classificar 463 classes distintas com uma acurácia de 77,89%, um acordo com o departamento de estatística e

análise criminal da SIAC foi estabelecido, a fim de utilizar o modelo proposto como parte de uma ferramenta de consolidação automática dos relatos diários.

Um conjunto de teste de produção foi selecionado para o período entre 16 de março e 5 de outubro de 2022, contendo 266.595 amostras. É importante mencionar que este conjunto já possui consolidados atribuídos, restando então conduzir uma análise da comparação direta (acertos e erros baseados no casamento de strings) entre as classes preditas e esperadas. É importante reiterar que as classes esperadas podem ser de dois tipos:

- *Registros*: são as classes atribuídas pelas delegacias. São o único fator de comparação presente antes da leitura humana, logo são as classes confrontadas pelo classificador nas predições diárias.
- *Consolidados*: são as classes atribuídas pela secretaria. Estas classes só se fazem presentes no fechamento da base, logo é impossível fazer a sua comparação com as predições do classificador no início da produção diária.

Foram identificados 6 subconjuntos de amostras no conjunto de produção, diferenciados pela presença de seus consolidados na lista de 463 classes aprendidas pelo classificador, e pela presença dessas amostras no conjunto de leitura diário da secretaria [3], cuja representação gráfica e quantidades de amostras estão expressas na Figura 13.

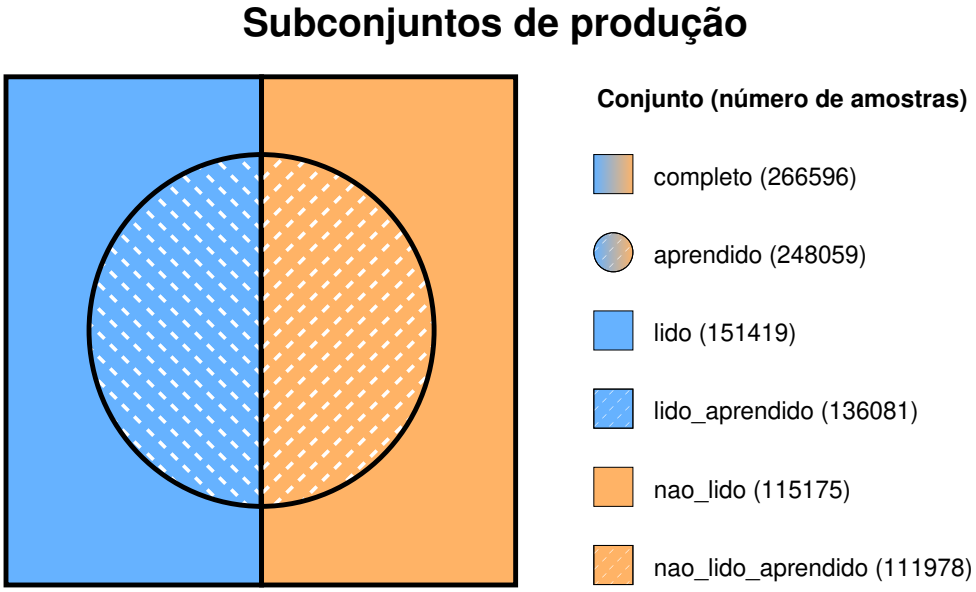
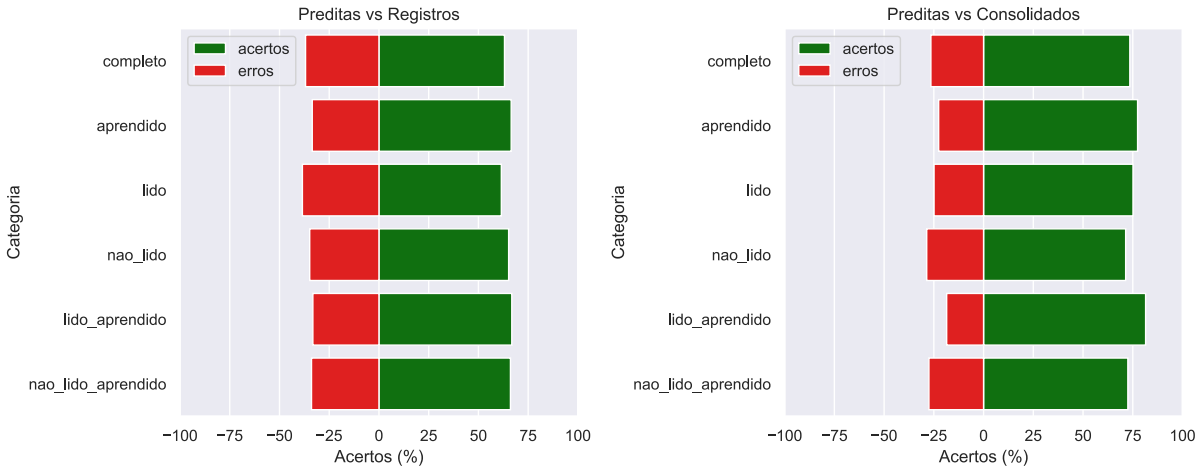


Figura 13 – Grupos de amostras de produção de acordo com aprendizado de suas classes ou consolidação através de leitura humana.

As taxas de acerto das comparações entre as classes *Preditas* em relação às de *Registros* e *Consolidados* e as descrições de cada conjunto são detalhadas a seguir e na Figura 14:

[3] Mortes em geral, furtos e roubos.

- "completo" ($Registros = 63,1\%$ e $Consolidados = 73,5\%$) considera todas as amostras disponíveis, sem restrição;
- "aprendido" ($Registros = 66,4\%$ e $Consolidados = 77,4\%$) considera apenas as amostras cuja classe de consolidado foi aprendida pelo classificador;
- "lido" ($Registros = 61,5\%$ e $Consolidados = 75,1\%$) considera apenas as amostras lidas, sendo então as amostras mais corretas possíveis na base de dados, cujas classes não são necessariamente aprendidas pelo classificador. Este é o conjunto usado nas predições diárias, já que sua posterior análise humana permite sua avaliação;
- "lido_aprendido" ($Registros = 66,7\%$ e $Consolidados = 81,4\%$) considera apenas as amostras lidas cuja classe de consolidado foi aprendida pelo classificador;
- "nao_lido" ($Registros = 65,1\%$ e $Consolidados = 71,4\%$) considera apenas as amostras não lidas, cujas classes são diretamente adaptadas da coluna de registros e que geralmente apresentam discordâncias na comparação;
- "nao_lido_aprendido" ($Registros = 63,1\%$ e $Consolidados = 73,5\%$) considera apenas as amostras não lidas cuja classe de consolidado foi aprendida pelo classificador;



(a) Preditas vs Registros. **(b) Preditas vs Consolidados.**

Figura 14 – Proporção de acertos para a comparação entre classes preditas e esperadas.

Os resultados indicam que o classificador pode reduzir os esforços de leitura para consolidação dos relatos em até 61,5%, a proporção de amostras lidas pelos analistas com classes que não necessariamente foram aprendidas pelo classificador, em comparação com as classes disponíveis antes da análise humana. Na prática, essa proporção de redução de esforços tende a ser menor, já que inclui classes de mortes violentas que sempre são verificadas por analistas. Já para as amostras lidas com consolidados aprendidos, a alta quantidade de acertos indica a importância da escolha do conjunto de classes as serem aprendidas, um dos principais desafios do problema proposto, dado o conjunto de dados desbalanceado.

Ambos os conjuntos de amostras não lidas por analistas consideram a maioria das classes raras que não são normalizadas, ou não são priorizadas no processo de leitura, ou foram registradas antes de 2019 (o primeiro do corte trienal do conjunto de dados), ou não são mais usadas pela secretaria. Uma análise das predições incorretas destes conjuntos por parte dos analistas tem o potencial de indicar uma quantidade precisa de amostras cujas classes de consolidados seriam ajustadas caso o classificador atuasse verdadeiramente sobre estas, gerando mapeamentos mais precisos a serem usados nos próximos modelos de classificação.

Além destes pontos, a comparação entre as classes de *Predições* e de *Consolidados* possuem taxas de acertos sempre superiores às comparações entre as classes de *Predições* e de *Registros*, o que é de ser esperado. O principal motivo para isso é que as classes aprendidas pelo classificador são derivadas das classes de *Consolidados*. Por outro lado, as classes de *Registros* só podem ser comparadas com as 463 classes modeladas mediante comparações de strings, um método automático que induz esses resultados a erros[4]. Desta forma, pode ser especulado que um mapeamento entre todas as possibilidades de registros possíveis para cada classe de consolidado aprendido pode melhorar os efeitos destas comparações, e consequentemente aumentar a qualidade da separação entre os boletins confirmados pelo classificados e aqueles que são passados para leitura humana.

A implementação do classificador trouxe não apenas melhorias qualitativas dos dados processados na secretaria, ao realocar a carga de trabalho dos analistas para os relatos mais complexos de serem rotulados, mas também reduziu o tempo de geração dos relatórios diários que a administração pública demanda. Antes do uso da ferramenta, os relatórios eram lançados com até um dia de atraso após a coleta dos boletins, em comparação com o horário de lançamento atual, com uma diferença média de 5 horas após o início dos processos manuais de consolidação diários. A figura 15 mostra a distribuição dos horários de dois tipos de atualizações dos consolidados na base de dados diária, registrados pela equipe de analistas em um arquivo de registros à parte:

- Atualização de predição: é o momento em que é gerada a base de dados quantizada, contendo as predições do classificador devidamente verificadas pelos analistas em conjunto com a base não lida. Possui uma mediana de horário de 09:43.
- Atualização de fechamento: é o momento em que é gerada a base de dados consolidada e qualificada, sinalizando o fim do processo de análise diário e o envio da base tratada para a SEGUP. Possui uma mediana de horário de 14:49.

[4] Por exemplo, o consolidado *Homícidio* pode ser encontrado dentro do registro *Homicídio Culposo*, *Furto* dentro de *Tentativa de Furto*, *Estupro* dentro de *Estupro de Vulnerável*, e assim por diante.

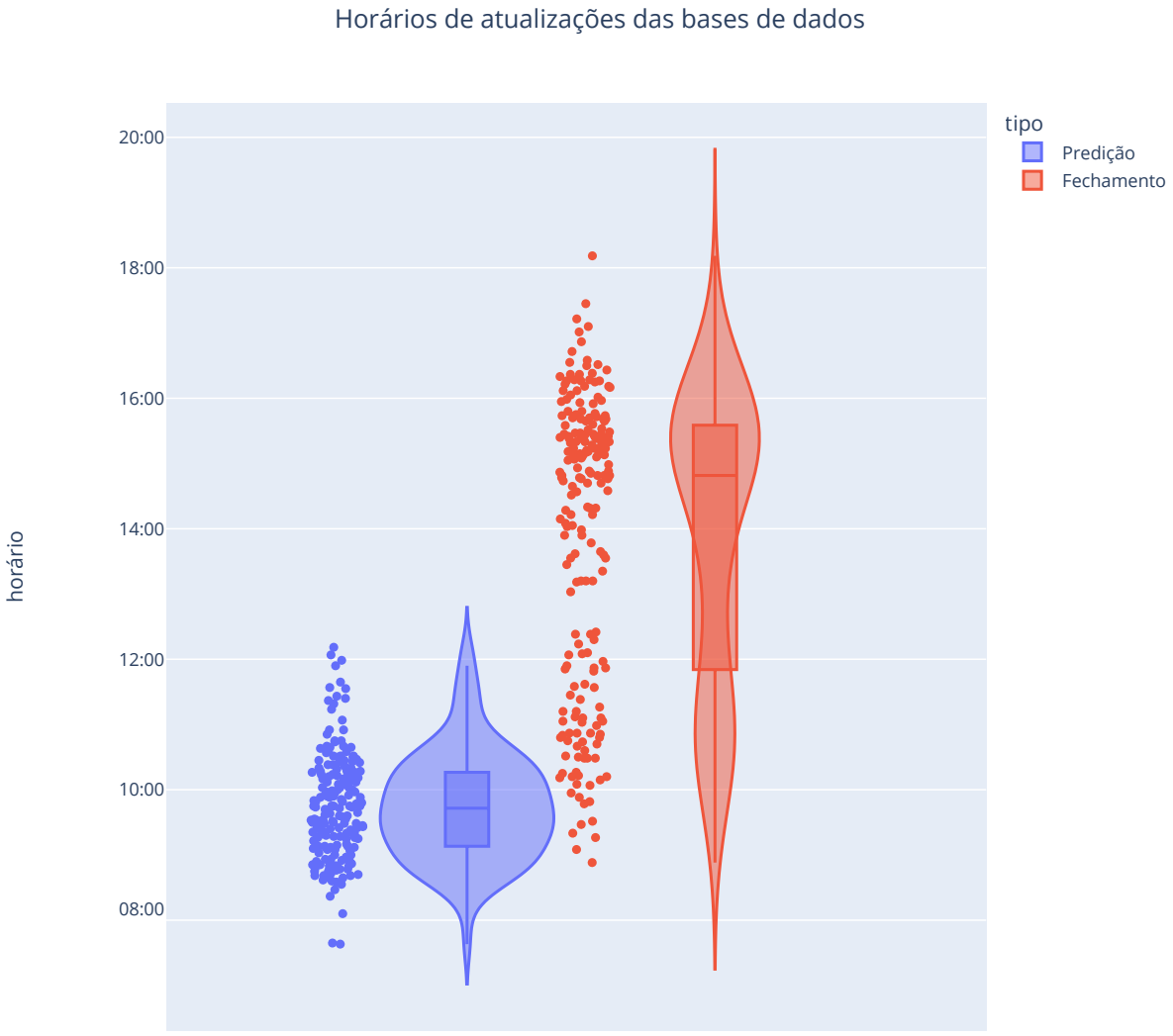


Figura 15 – Horários de atualizações da base de dados.

6 CONCLUSÃO

Este trabalho apresentou a construção de uma ferramenta de mineração de dados usada para extrair conhecimento de uma base de dados de relatos policiais através da aplicação de um modelo de classificação supervisionada. A arquitetura proposta foi baseada no CRISP-DM, um modelo padronizado usualmente utilizado em tarefas de mineração de dados. O modelo de classificação alcançou uma acurácia geral de 78%, um resultado diretamente impactado pela alta quantidade de classes no problema de classificação e pelo fato de que a contagem de exemplos de aprendizado para cada classe é desbalanceado, limitando a performance para classes raras. O modelo foi usado por agências estatísticas de segurança pública no estado do Pará, como uma forma de automatizar o processo de confirmação de determinados eventos dentro da descrição textual dos relatos, reduzindo a diferença entre o horário do início do processamento e a consolidação do boletins na base em até 5 horas. O impacto da aplicação do classificador em termos de melhoria de processos ainda há de ser medido e pode servir de motivação para futuras pesquisas. Outras agências de segurança pública também podem se beneficiar do uso do modelo proposto, já que o conhecimento coletado é baseado apenas no conjunto de padrões extraídos da descrição textual dos eventos rotulados, portanto a ferramenta gerada pode ser usada sob diferentes perspectivas, com as modificações necessárias, já que é completamente independente do ambiente legislativo onde será aplicada.

Possíveis melhorias para o classificador incluem a adoção de mais etapas do modelo CRISP-DM, uma decisão precisa de quais classes devem ser aprendidas, o balanceamento dos exemplos de aprendizado para todas as classes, a aplicação de diferentes algoritmos de aprendizado de máquina, e a aplicação de algoritmos de otimização de hiper-parâmetros. Em relação às etapas de pré-processamento, a aplicação da técnica de *Named-Entity Recognition* (NER) para o destaque de atributos socio-econômicos de impacto dentro dos relatórios, engenharia de atributos, e significância estatística das amostras podem ser exploradas a fim de melhorar a qualidade dos dados a serem processados. Um esforço de integração entre as áreas do direito e da inteligência artificial pode ser decisivo na aplicação de conhecimento específico para a legislação brasileira na descoberta de padrões determinísticos para o problema de predição das classes de eventos consolidados.

REFERÊNCIAS

ABDI. Sistemas aplicados à segurança pública. **Cadernos temáticos TICs - ABDI**, v. 3, p. 226, 2010. Disponível em: <http://livroaberto.ibict.br/handle/1/536>.

ABSP. Anuário Brasileiro de Segurança Pública. **Fórum Brasileiro de Segurança Pública**, v. 16, p. 516, 2022. ISSN 1983-7364. Disponível em: <https://forumseguranca.org.br/anuario-brasileiro-seguranca-publica/>.

AGGARWAL, C. C. **Data Mining**. New York: Springer, 2015.

ALDOSSARI, B. S. et al. A comparative study of decision tree and naive bayes machine learning model for crime category prediction in chicago. In: **Proceedings of 2020 the 6th International Conference on Computing and Data Engineering**. New York, NY, USA: Association for Computing Machinery, 2020. (ICCDE 2020), p. 34–38. ISBN 9781450376730. Disponível em: <https://doi.org/10.1145/3379247.3379279>.

AMORIM, M. da S.; PEREIRA, J. R. S. **Tipificação de ocorrências policiais utilizando machine learning**. Salvador: [s.n.], 2019. P. 50.

ASSAD, F. J. P.; CHAGAS, J. F. C. **Análise Preditiva de Manchas Criminais no Estado de São Paulo**. Niterói: [s.n.], 2019. P. 93.

BLUM, A.; HOPCROFT, J.; KANNAN, R. **Foundations of Data Science**. [s.n.], 2015. Disponível em: <https://books.google.com.br/books?id=YTmpswEACAAJ>.

BRASIL. **O Sinesp**. 2019. Acesso em: 16 de março de 2022. Disponível em: <https://www.gov.br/mj/pt-br/assuntos/sua-seguranca/seguranca-publica/sinesp-1/>.

CASTELLA, E. M. **Investigação criminal na era do governo eletrônico: inteligência artificial x boletim de ocorrência - BO, soluções em K.M.A.I.** Santa Catarina: [s.n.], 2002. P. 150.

CASTRO Úrsula Rosa Monteiro de. **Explorando aprendizagem supervisionada em dados heterogêneos para predição de crimes**. Belo Horizonte: [s.n.], 2020. P. 85.

CHEN, P.; KURLAND, J. Time, place, and modus operandi: A simple apriori algorithm experiment for crime pattern detection. In: **2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA)**. [S.l.: s.n.], 2018. p. 1–3.

CHIEW, L. S.; AMERUDIN, S.; YUSOF, Z. M. A spatial analysis of the relationship between socio-demographic characteristics with burglar behaviours on burglary crime. **IOP Conference Series: Earth and Environmental Science**, IOP Publishing, v. 540, n. 1, p. 012050, jul 2020. Disponível em: <https://doi.org/10.1088/1755-1315/540/1/012050>.

CHOLLET, F. **Deep Learning with Python**. 1st. ed. [S.l.]: Manning Shelter Island, 2018.

COSTA, C. F. P. d. S. et al. Equal criminal investigation for all? An analysis based on the profile of victims of intentional homicides in Belém/Pará. **Research, Society and Development**, v. 9, n. 12, p. e45491211439, Dec. 2020. Disponível em: <https://rsdjournal.org/index.php/rsd/article/view/11439>.

COSTA, J. C. de O. R. **Identificação de municípios pernambucanos para recomendação de políticas de segurança pública utilizando uma técnica de clusterização**. Caruaru: [s.n.], 2020. P. 70.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 14, n. 3, p. 121–134, 1996. ISSN 2595-1521. Disponível em: <https://www.theacademicsociety.net/tasjonline-v3-pg-121-134>.

FURTADO, L.; SOUZA, A. Uso de dados provenientes de rede social e técnica de mineração de dados para classificar crimes em belém-pa. **The Academic Society Journal**, v. 3, n. 2, p. 121–134, 2019. ISSN 2595-1521. Disponível em: <https://www.theacademicsociety.net/tasjonline-v3-pg-121-134>.

HAYKIN, S. **Redes Neurais: Princípios e Prática**. Porto Alegre: Bookman, 2001.

IBM Corp. **IBM SPSS Modeler CRISP-DM Guide**. 2011. Acesso em: 30 de novembro de 2022. Disponível em: <https://www.ibm.com/docs/en/spss-modeler/18.1.1?topic=spss-modeler-crisp-dm-guide>.

KIM, Y. Convolutional Neural Networks for Sentence Classification. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1746–1751. Disponível em: <https://aclanthology.org/D14-1181>.

KSHATRI, S. S. et al. An empirical analysis of machine learning algorithms for crime prediction using stacked generalization: An ensemble approach. **IEEE Access**, v. 9, p. 67488–67500, 2021.

MIRANDA, Z. Política de Ciência, Tecnologia e Inovação para a Segurança Pública. **Revista Brasileira de Segurança Pública**, v. 6, n. 2, set. 2012. Disponível em: <https://revista.forumseguranca.org.br/index.php/rbsp/article/view/129>.

PRADO, K. H. de J. **Data science aplicada à análise criminal baseada nos dados abertos governamentais do Brasil**. São Cristóvão: [s.n.], 2020. P. 146.

QAZI, N.; WONG, B. W. An interactive human centered data science approach towards crime pattern analysis. **Information Processing & Management**, v. 56, n. 6, p. 102066, 2019. ISSN 0306-4573. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0306457318302942>.

RATUL, M. A. R. A comparative study on crime in denver city based on machine learning and data mining. **CoRR**, abs/2001.02802, 2020. Disponível em: <http://arxiv.org/abs/2001.02802>.

REGATEIRO, H. A. S. et al. Assessment of crime in the state of pará. **Research, Society and Development**, v. 10, n. 3, p. e10010313088, Mar. 2021. Disponível em: <https://rsdjournal.org/index.php/rsd/article/view/13088>.

ROSENBLATT, F. **The perceptron - A perceiving and recognizing automaton**. Ithaca, New York, 1957.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence**. New Jersey: Prentice Hall, 2010.

SOUZA, D. C. de. **Introdução à Ciência do Direito**. [S.l.]: FGV, 1972.

SOUZA, J. R. M. de. **Utilização de aprendizagem de máquina na predição de crimes**. Niterói: [s.n.], 2018. P. 54.

SOUZA, S. L. de et al. Data Mining in Public Security Databases in Belém, Pará, Brazil. 2022.

SOUZA, S. L. de et al. Data Mining and Analysis Applied to Public Security Data in Belém of Pará, Brazil. 2022.

TRINDADE, E. A. R. de A. **Homicídios na Região Metropolitana de Belém: práticas para contenção e vulnerabilidades**. Belém: [s.n.], 2019. P. 155.

VARGAS, W. A. L. de. **Data science & segurança pública: padrões estatísticos sobre as ocorrências de flagrantes em roubo de celular na cidade de São Paulo**. São Paulo: [s.n.], 2019. P. 52.

VINHOLES, T. V. M. e L. M. F. Descoberta de conhecimento em banco de dados relacionados à violência contra a mulher. **Anais do Computer on the Beach**, v. 0, n. 0, p. 815–816, 2019. ISSN 2358-0852. Disponível em: <https://siaiap32.univali.br/seer/index.php/acotb/article/view/14428>.

Apêndices

APÊNDICE A – PUBLICAÇÕES

A.1 Trabalhos publicados

- **Helder Matos**, Samara Souza, Reginaldo Santos, João C. W. A. Costa, Cleyton Costa. *"A Supervised Classifier for Police Reports at the State of Pará, Brazil"*. II Escola Regional de Aprendizado de Máquina e Inteligência Artificial Norte 2. 2022.
- Samara Souza, **Helder Matos**, Cleyton Costa, Reginaldo Santos, João C. W. A. Costa. *"Data Mining in Public Security Databases in Belém, Pará, Brazil"*. II Escola Regional de Aprendizado de Máquina e Inteligência Artificial Norte 2. 2022.
- **Helder Matos**, Samara Souza, Reginaldo Santos, João C. W. A. Costa. *"A Supervised Classifier for Police Reports at the State of Pará, Brazil"*. 19th CONTECSI - International Conference on Information Systems and Technology Management. 2022.
- Samara Souza, **Helder Matos**, Reginaldo Santos, João C. W. A. Costa. *"Data Mining and Analysis Applied to Public Security Data in Belém of Pará, Brazil"*. 19th CONTECSI - International Conference on Information Systems and Technology Management. 2022.

A.2 Participações em eventos

- Luis J. L. Gonçalves, **Helder Matos**. *"Inteligência Artificial como facilitadora de processamento de grande massa de dados relacionados à segurança pública"*. 16º Encontro Anual do Fórum Brasileiro de Segurança Pública. 2022. São Paulo.
- **Helder Matos**, Samara Souza, Reginaldo Santos, João C. W. A. Costa, Cleyton Costa. *"A Supervised Classifier for Police Reports at the State of Pará, Brazil"*. II Escola Regional de Aprendizado de Máquina e Inteligência Artificial Norte 2. 2022. Belém.
- Samara Souza, **Helder Matos**, Cleyton Costa, Reginaldo Santos, João C. W. A. Costa. *"Data Mining in Public Security Databases in Belém, Pará, Brazil"*. II Escola Regional de Aprendizado de Máquina e Inteligência Artificial Norte 2. 2022. Belém.
- **Helder Matos**, Samara Souza, Reginaldo Santos, João C. W. A. Costa. *"A Supervised Classifier for Police Reports at the State of Pará, Brazil"*. 19th CONTECSI - International Conference on Information Systems and Technology Management. 2022. São Paulo.
- Samara Souza, **Helder Matos**, Reginaldo Santos, João C. W. A. Costa. *"Data Mining and Analysis Applied to Public Security Data in Belém of Pará, Brazil"*. 19th CONTECSI - International Conference on Information Systems and Technology Management. 2022. São Paulo.

APÊNDICE B – DICIONÁRIO DE DADOS

nome coluna	ordem	data type	length	Precision	DESCRIÇÃO
servidor	1	varchar	50	50	NOME DO SERVIDOR QUE PROCEDE A LEITURA DO BOP
nro_bop	2	varchar	500	500	NUMERO DO BOP
nro_bop_aditado	3	varchar	500	500	NUMERO DO BOP ADITADO
nro_tombo	4	varchar	500	500	NUMERO DO TOMBO
tipo_tombo	5	varchar	500	500	TIPO DO TOMBO
unidade_origem	6	varchar	500	500	UNIDADE DE REGISTRO DO BOP
unidade_responsavel	7	varchar	500	500	UNIDADE RESPONSAVEL PELO BOP
data_registro	8	date		13	DATA DE REGISTRO DO BOP
hora_registro	9	time		15	HORA DE REGISTRO DO BOP
data_fato	10	date		13	DATA DO FATO DO BOP
dia_semana	11	varchar	500	500	DIA DA SEMANA DO BOP
hora_fato	12	time		15	HORA DO FATO DO BOP
fx_4_hor*	13	varchar	500	500	FAIXA DE 06 HORAS DO BOP
fx_12_hr*	14	varchar	500	500	FAIXA DE 02 HORAS DO BOP
data_inst_proc	15	date		13	DATA DE INSTAURAÇÃO DO TOMBO
data_concl_proc	16	date		13	DATA DE CONCLUSAO DO TOMBO
sit_proc*	17	varchar	500	500	SITUAÇÃO DO TOMBO
classe_motivo	18	varchar	500	500	ENQUADRAMENTO DO CRIME - AUTOMATICO SISTEMA
mes_registro	19	varchar	500	500	MÊS REGISTRO DO BOP
mes_fato	20	varchar	500	500	MÊS FATO DO BOP
ano_registro	21	int4		10	ANO REGISTRO BOP
ano_fato	22	int4		10	ANO FATO BOP
registros	23	varchar	500	500	CRIMES REGISTRADOS
consolidado	24	varchar	500	500	RETRATA O CRIME POS LEITURA
fato_real	25	varchar	500	500	TODA LINHA PREENCHIDA REFLETE QUE O BOP PASSOU PELA LEITURA
especificacao_crime	26	varchar	500	500	ESPECIFICA O DELITO DO CONSOLIDADO
meio_emp_deac	27	varchar	500	500	MEIO EMPREGADO DETECTADO NA LEITURA
latitude	28	varchar	500	500	AUTO EXPLICATIVO
longitude	29	varchar	500	500	AUTO EXPLICATIVO
causa_presumivel	30	varchar	500	500	ENQUADRAMENTO DO CRIME - AUTOMATICO SISTEMA
especializacao_fato	31	varchar	500	500	ENQUADRAMENTO DO CRIME - AUTOMATICO SISTEMA
grupo_ocorrencia	32	varchar	500	500	ENQUADRAMENTO DO CRIME - AUTOMATICO SISTEMA
sub_grupo	33	varchar	500	500	ENQUADRAMENTO DO CRIME - AUTOMATICO SISTEMA
meio_empregado_sisp	34	varchar	500	500	MEIO EMPREGADO CADASTRADO NO REGISTRO DO BOP
distrito	35	varchar	500	500	AUTO EXPLICATIVO
municipios	36	varchar	500	500	AUTO EXPLICATIVO
regionais	37	varchar	500	500	DISTRIBUIÇÃO DOS MUNICIPIOS EM REGIONAIS
bairros	38	varchar	500	500	AUTO EXPLICATIVO
reg_integracao	39	varchar	500	500	DISTRIBUIÇÃO DOS MUNICIPIOS EM REGIONAIS PARA ATENDER FAPESPA
risp	40	varchar	500	500	RETRATA O CRIME POS LEITURA
aisp	41	varchar	500	500	DISTRIBUIÇÃO DOS MUNICIPIOS EM RISPS
rua_fato	42	varchar	500	500	ENDEREÇO DO FATO
empresa	43	varchar	500	500	PARA PREENCHIMENTO EM CASOS DE ROUBO A COLETIVO
linha	44	varchar	500	500	PARA PREENCHIMENTO EM CASOS DE ROUBO A COLETIVO
tipo_transporte	45	varchar	500	500	PARA PREENCHIMENTO EM CASOS DE ROUBO A COLETIVO
complemento	46	text			COMPLEMENTO DE ENDEREÇO DO FATO
local_ocorrencia	47	text			LOCALIZAÇÃO DO CRIME, SE BAR, VIA PUBLICA
identificacao_fato	48	text			RESUMO EM FORMA DE TITULO DO REGISTRO DO CRIME
relator	49	text			PESSOA QUE REGISTROU A OCORRENCIA
relato	50	text			RELATO DA OCORRENCIA

Tabela 3 – Dicionário de dados da base coletada (cont.)

nome coluna	ordem	data type	length	Precision	DESCRIÇÃO
atuacao	51	varchar	500	500	ATUAÇÃO DA PESSOA NO BOP, SE VITIMA, AUTOR, ETC...
vit_nome	52	text			NOME
vit_alcunha	53	varchar	500	500	ALCUNHA
vit_dt_nasc	54	date		13	DATA NASCIMENTO
vit_idade	55	int4		10	IDADE
vit_fx_etaria	56	varchar	500	500	FAIXA ETARIA
vit_nro_doc	57	varchar	500	500	NRO DOCUMENTO IDENTIFICAÇÃO
vit_tipo_doc	58	varchar	500	500	TIPO DOCUMENTO IDENTIFICAÇÃO
vit_pai	59	varchar	500	500	PAI
vit_mae	60	varchar	500	500	MAE
vit_tipo	61	varchar	500	500	PARA CASOS ESPECIFICOS QUE FOGEM DA NORMALIDADE COMO: POLICIAIS, USUARIOS DE DROGAS, PRESIDIARIOS, ETC...
vit_sexo	62	varchar	500	500	SEXO
vit_cor_pele	63	varchar	500	500	COR DA PELE
vit_grau_inst	64	varchar	500	500	INSTRUÇÃO
vit_profissao	65	varchar	500	500	PROFISSAO
vit_situacao_emprego	66	varchar	500	500	SE EMPREGADO, FOLGA, SERVIÇO, DESEMPREGADO, APOSENTADO, ETC...
vit_estado_civil	67	varchar	500	500	ESTADO CIVIL
aut_nome	68	varchar	500	500	NOME
aut_alcunha	69	varchar	500	500	ALCUNHA
aut_data_nasc	70	date		13	DATA NASCIMENTO
aut_idade	71	int4		10	IDADE
aut_fx_etaria	72	varchar	500	500	FAIXA ETARIA
aut_nro_doc	73	varchar	500	500	NRO DOCUMENTO IDENTIFICAÇÃO
aut_tipo_doc	74	varchar	500	500	TIPO DOCUMENTO IDENTIFICAÇÃO
aut_pai	75	varchar	500	500	PAI
aut_mae	76	varchar	500	500	MAE
aut_tipo	77	varchar	500	500	PARA CASOS ESPECIFICOS QUE FOGEM DA NORMALIDADE COMO: POLICIAIS, USUARIOS DE DROGAS, PRESIDIARIOS
aut_sexo	78	varchar	500	500	SEXO
grau_de_relacionamento	79	varchar	500	500	SE HÁ ALGUM TIPO DE RELAÇÃO ENTRE VITIMA E AUTOR
aut_cor_pele	80	varchar	500	500	COR DA PELE
aut_grau_inst	81	varchar	500	500	INSTRUÇÃO
aut_profissao	82	varchar	500	500	PROFISSAO
aut_sit_emprego	83	varchar	500	500	SE EMPREGADO, FOLGA, SERVIÇO, DESEMPREGADO, APOSENTADO, ETC...
aut_est_civil	84	varchar	500	500	ESTADO CIVIL
meio_locomocao	85	varchar	500	500	VEICULO UTILIZADO NO COMETIMENTO DA AÇÃO CRIMINOSA
cor_veiculo	86	varchar	500	500	COR DO VEICULO
marca_veic_fuga	87	varchar	500	500	MARCA DO VEICULO
modelo_do_veic_fuga	88	varchar	500	500	MODELO DO VEICULO
qtd_autor	89	int4		10	QUANTIDADE DE AUTORES
relatorio	90	varchar	500	500	SE NO TOMBAMENTO, HÁ RELATORIO DE CONCLUSAO DO PROCEDIMENTO
ident_autoria	91	varchar	500	500	SE AUTOR FOI IDENTIFICADO
pk	95	int4		10	CHAVE PRIMARIA

Tabela 4 – Dicionário de dados da base coletada.

Anexos

ANEXO A – LISTA DE CONSOLIDADOS

Tabela 5 – Lista de consolidados (cont.)

1	A ARREGIMENTACAO DE ELEITOR OU A PROPAGANDA DE BOCA DE URNA	31	ADQUIRIR OU RECEBER MADEIRA, LENHA SEM EXIGIR A EXIBICAO DE LICENCA DO VENDEDOR	61	APRESENTACAO SUSPEITO DE ROUBO
2	A DIVULGACAO DE QUALQUER ESPECIE DE PROPAGANDA DE PARTIDOS POLITICOS OU CANDIDATOS	32	ADQUIRIR, ARMAZENAR, POR QUALQUER MEIO, FOTOGRAFIA, VIDEO OU OUTRA FORMA DE REGISTRO QUE CONTENHA CENA DE SEXO EXPLICITO ENVOLVENDO CRIANCA OU ADOLESCENTE	62	APRESENTACAO SUSPEITO DE TRAFICO DE DROGAS
3	A INCOLUMIDADE FISICA DO INDIVIDUO	33	ADULTERACAO DE SINAL IDENTIFICADOR DE VEICULO AUTOMOTOR	63	APRESENTACAO TENTATIVA DE ROUBO
4	A INVIOLABILIDADE DO DOMICILIO	34	AFASTAR-SE O CONDUTOR DO LOCAL DO ACIDENTE	64	APROPRIA-SE DE BENS OU RENDIMENTO DO IDOSO
5	A VIOLENCIA PATRIMONIAL, ENTENDIDA COMO QUALQUER CONDUTA QUE CONFIGURE RETENCAO, SUBTRACAO, DESTRUICAO PARCIAL OU TOTAL DE SEUS OBJETOS, INSTRUMENTOS DE TRABALHO, DOCUMENTOS PESSOAIS, BENS, VALORES E DIREITOS OU RECURSOS ECONOMICOS, INCLUINDO OS DEST	35	AFIRMACAO FALSA OU ENGANOSA SOBRE A NATUREZA DE PRODUTOS OU SERVICOS	65	APROPRIACAO DE COISA ACHADA
6	ABANDONAR IDOSO EM HOSPITAIS	36	AFOGAMENTO COM MORTE	66	APROPRIACAO INDEBITA
7	ABANDONAR SUBSTANCIA TOXICA, PERIGOSA AO MEIO AMBIENTE, EM DESACORDO COM AS EXIGENCIAS LEGAIS	37	ALICIAR, ASSEDIAR, INSTIGAR POR QUALQUER MEIO DE COMUNICACAO, CRIANCA, COM O FIM DE COM ELA PRATICAR ATO LIBIDINOSO	67	APROPRIAR-SE DE BENS OU RENDIMENTO DO IDOSO
8	ABANDONO DE FUNCAO	38	ALIENACAO OU ONERACAO FRAUDULENTA DE COISA PROPRIA	68	ASSEDIO MORAL
9	ABANDONO DE INCAPAZ	39	ALTERACAO DE LIMITES	69	ASSEDIO SEXUAL
10	ABANDONO DE LAR	40	AMEACA	70	ASSEDIO SEXUAL POR SUPERIOR HIERARQUICO OU ASCENDENTE
11	ABANDONO DE TRABALHO	41	AOS DIREITOS E GARANTIAS LEGAIS ASSEGURADOS AO EXERCICIO PROFISSIONAL	71	ASSISTENCIA SOCIAL
12	ABANDONO INTELECTUAL	42	APOLOGIA AO CRIME OU CRIMINOSO	72	ASSOCIACAO CRIMINOSA
13	ABANDONO MATERIAL	43	APREENSAO DE ARMA BRANCA	73	ATENTADO CONTRA A LIBERDADE DE TRABALHO
14	ABORTO	44	APREENSAO DE ARMA DE FOGO	74	ATENTADO CONTRA A SEGURANCA DE SERVICO DE UTILIDADE PUBLICA
15	ABUSAR DO PODER ECONOMICO	45	APREENSAO DE ARMA DE FOGO DE USO PERMITIDO	75	ATENTADO CONTRA A SEGURANCA DE TRANSPORTE MARITIMO, FLUVIAL OU AEREO
16	ABUSO DE AUTORIDADE	46	APREENSAO DE BICICLETA	76	ATITUDE SUSPEITA E/OU APRESENTACAO
17	ACAO PENAL	47	APREENSAO DE CELULAR	77	ATIVIDADES CLANDESTINAS DE TELECOMUNICACOES
18	ACIDENTE AEREO COM MORTE	48	APREENSAO DE CELULAR EM PRESIDIO	78	ATO OBSCENO
19	ACIDENTE DE TRABALHO COM DANO	49	APREENSAO DE DROGAS	79	ATROPELAMENTO DE ANIMAL
20	ACIDENTE DE TRABALHO COM LESAO	50	APREENSAO DE MERCADORIAS	80	AUTO-ACUSACAO FALSA
21	ACIDENTE DE TRABALHO COM MORTE	51	APREENSAO DE OBJETOS	81	BAIXA DE ALIENACAO FIDUCIARIA DE VEICULO
22	ACIDENTE DE TRANSITO	52	APREENSAO DE SIMULACRO	82	BUSCA E APREENSAO
23	ACIDENTE DE TRANSITO SEM VITIMA	53	APREENSAO DE VEICULOS	83	CADASTRO DE ARMAS
24	ACIDENTE FORTUITO	54	APRESENTACAO - ATITUDE SUSPEITA	84	CALUNIA
25	ACIDENTE FORTUITO COM DANO	55	APRESENTACAO - MANDADO DE BUSCA E APREENSAO	85	CARCERE PRIVADO
26	ACIDENTE FORTUITO COM LESAO	56	APRESENTACAO - MANDADO DE PRISAO	86	CARTAO CLONADO
27	ACIDENTE FORTUITO COM MORTE	57	APRESENTACAO - PRESO FORAGIDO	87	CAUSAR POLUICAO DE QUALQUER NATUREZA
28	ACIDENTE MARITIMO COM DANO	58	APRESENTACAO - SUSPEITO DE FURTO	88	CAUSAR POLUICAO DE QUALQUER NATUREZA, RESULTANTE EM DANOS A SAUDE HUMANA
29	ACIDENTE MARITIMO COM LESAO	59	APRESENTACAO DE MENOR INFRATOR	89	CERTIDAO OU ATESTADO IDEOLOGICAMENTE FALSO
30	ACIDENTE MARITIMO COM MORTE	60	APRESENTACAO SUSPEITO DE HOMICIDIO	90	COACAO NO CURSO DO PROCESSO

Tabela 6 – Lista de consolidados (cont.)

91	COAGIR O IDOSO	121	CORTAR OU TRANSFORMAR EM CARVAO MADEIRA DE LEI	151	DESCUMPRIMENTO DE MEDIDAS PROTETIVAS DE URGENCIA
92	COMERCIALIZAR DERIVADOS DE PETROLEO EM DESACORDO COM AS NORMAS LEGAIS	122	CRIME AMBIENTAL	152	DESCUMPRIMENTO DOS DEVERES INERENTES AO PODER FAMILIAR, DE TUTELA OU DE GUARDA
93	COMERCIALIZAR MOTOSSERRA OU UTILIZA-LA EM FLORESTAS	123	CRIMES CONTRA A ADMINISTRACAO AMBIENTAL	153	DESCUMPRIR DEVERES INERENTES AO PODER FAMILIAR
94	COMERCIO ILEGAL DE ARMA DE FOGO	124	CRIMES CONTRA A FAUNA	154	DESCUMPRIR PRAZO FIXADO EM LEI, BENEFICIO DE ADOLESCENTE PRIVADO DE LIBERDADE
95	COMERCIO ILEGAL DE MADEIRA	125	CRIMES CONTRA A FLORA	155	DESMATAR FLORESTA, PLANTADA OU NATIVA, EM TERRAS DE DOMINIO PUBLICO OU DEVOLUTAS
96	COMUNICACAO DE FUGA DE PRESO	126	CRIMES CONTRA A ORDEM ECONOMICA E AS RELACOES DE CONSUMO	156	DESOBEDIENCIA
97	COMUNICACAO DE OBITO	127	CRIMES CONTRA A ORDEM TRIBUTARIA	157	DESOBEDIENCIA A DECISAO JUDICIAL
98	COMUNICACAO FALSA DE CRIME OU DE CONTRAVENCAO	128	CRIMES CONTRA ORDENAMENTO URBANO E CULTURAL	158	DESOBEDIENCIA A DECISAO JUDICIAL SOBRE PERDA OU SUSPENSAO DE DIREITO
99	COMUNICACAO PARA FINS DE DIREITO	129	CRUELDADE CONTRA ANIMAIS	159	DESTRUIR BEM PROTEGIDO POR LEI
100	CONCENTRACAO DE ELEITORES	130	CULTIVO DE DROGAS	160	DESTRUIR FLORESTAS NATIVAS DE VEGETACAO FIXADORA DE DUNAS, MANGUES, OBJETO DE PRESERVACAO
101	CONCUSSAO	131	CUMPRIMENTO - MANDADO DE PRISAO	161	DESTRUIR OU DANIFICAR FLORESTA PRESERVADA
102	CONDICIONAMENTO DE ATENDIMENTO MEDICO-HOSPITALAR EMERGENCIAL	132	CUMPRIMENTO DE MANDADO DE BUSCA E APREENSAO	162	DESTRUIR OU DANIFICAR VEGETACAO PRIMARIA
103	CONDUCAO SOB INFLUENCIA DE SUBSTANCIA QUE ALTERA A CONDICAO PSICOMOTORA	133	DANO	163	DESTRUIR PLANTAS DE ORNAMENTACAO DE LOGRADOUROS PUBLICOS OU EM PROPRIEDADE PRIVADA ALHEIA
104	CONDUZIR VEICULO AUTOMOTOR EM VIA PUBLICA, DE CORRIDA, DISPUTA OU COMPETICAO AUTOMOBILISTICA	134	DANO AMBIENTAL	164	DIFAMACAO
105	CONDUZIR VEICULO SOB INFLUENCIA DE ALCOOL	135	DANO AO PATRIMONIO PUBLICO	165	DIFICULTAR A ACAO
106	CONFIGURACAO DO CRIME INDEPENDENTEMENTE DA COMPETENCIA CIVIL OU CRIMINAL PARA DECRETACAO DAS MEDIDAS	136	DANO CIVIL	166	DILIGENCIA DE MANDADO
107	CONFLITO DE GUARDA	137	DANO NO TRANSITO	167	DIRECAO PERIGOSA DE VEICULO EM VIA PUBLICA
108	CONFLITOS CONJUGAIS	138	DEFICIENCIA	168	DIRIGIR SEM HABILITACAO
109	CONFLITOS FAMILIARES	139	DEFRAUDACAO DE PENHOR	169	DISCRIMINAR, DESDENHAR, HUMILHAR, MENOSPREZAR PESSOA IDOSA
110	CONFLITOS VICINAIS	140	DEIXAR DE CUMPRIR OBRIGACAO DE RELEVANTE INTERESSE AMBIENTAL	170	DISPARO ACIDENTAL DE ARMA DE FOGO
111	CONSTRANGIMENTO ILEGAL	141	DEIXAR DE FORNECER NOTA FISCAL DE VENDA OU DE SERVICO	171	DISPARO ACIDENTAL DE ARMA DE FOGO COM LESAO
112	CONSTRUCAO DE ESTABELECIMENTOS, OBRAS OU SERVICOS POTENCIALMENTE POLUIDORES, SEM LICENCA OU AUTORIZACAO LEGAL	142	DEIXAR DE PRESTAR ASSISTENCIA AO IDOSO	172	DISPARO ACIDENTAL DE ARMA DE FOGO COM RESULTADO MORTE
113	CONSUMO PESSOAL DE DROGAS	143	DEIXAR DE RECOLHER TRIBUTOS	173	DISPARO DE ARMA DE FOGO
114	CONTRABANDO	144	DENUNCIACAO CALUNIOSA	174	DISPARO DE ARMA DE FOGO COM DANO
115	CONTRABANDO OU DESCAMINHO	145	DESACATO	175	DISPOSICAO DE COISAS ALHEIAS COMO PROPRIAS
116	CONTROLE DE MUNICAO	146	DESAPARECIMENTO DE ANIMAIS	176	DISPOSICOES GERAIS
117	CORRUPCAO ATIVA	147	DESAPARECIMENTO DE PESSOA	177	DIVULGACAO DE CENA DE ESTUPRO OU DE CENA DE ESTUPRO DE VULNERAVEL, DE CENA DE SEXO OU DE PORNOGRAFIA
118	CORRUPCAO DE MENORES	148	DESCAMINHO	178	DIVULGACAO DE SEGREDO - BANCO DE DADOS DA ADMINISTRACAO PUBLICA
119	CORRUPCAO PASSIVA	149	DESCUMP.DEVERES INERENTE ... OU DECORRENT.DE TUTELA	179	DIVULGACAO DE SEGREDO - DOCUMENTO PARTICULAR
120	CORTAR ARVORE CONSIDERADA PRESERVADA	150	DESCUMPRIMENTO DE MEDIDAS PROTETIVAS	180	DIVULGAR, PRODUZIR, VENDER, FORNECER, BEBIDA ALCOOLICA A MENOR DE IDADE

Tabela 7 – Lista de consolidados (cont.)

181	ELABORAR RELATORIO AMBIENTAL FALSO OU ENGANOSO, INCLUSIVE POR OMISSAO	211	FACILITAR FUGA DE PRESO	241	FRAUDES EM AVALIACAO OU EXAMES PUBLICOS
182	ELEVAR O VALOR DE VENDA A PRAZO	212	FALSA IDENTIDADE	242	FRAUDES EM CONCURSO PUBLICO
183	EMBRIAGUEZ	213	FALSIDADE DE ATESTADO MEDICO	243	FRUSTRACAO DE DIREITO ASSEGURADO POR LEI TRABALHISTA
184	EMISSAO IRREGULAR DE CONHECIMENTO DE DEPOSITO OU "WARRANT"	214	FALSIDADE IDEOLOGICA	244	FRUSTRAR OU FRAUDAR O CARATER COMPETITIVO DO PROCEDIMENTO LICITATORIO
185	ENTREGA DE FILHO MENOR A PESSOA INIDONEA	215	FALSIDADE MATERIAL DE ATESTADO OU CERTIDAO	245	FUGA DE ABRIGO
186	ENTREGAR DIRECAO DE VEICULO A PESSOA NAO HABILITADA	216	FALSIFICACAO DE CARTAO DE CREDITO OU DEBITO	246	FUGA DE INSTITUICAO
187	ESBULHO POSSESSORIO	217	FALSIFICACAO DE DOCUMENTO	247	FUGA DE LOCAL DE CRIME
188	ESCRITO OU OBJETO OBSCENO	218	FALSIFICACAO DE DOCUMENTO PARTICULAR	248	FUGA DE MENOR INFRATOR
189	ESTELIONATO	219	FALSIFICACAO DE DOCUMENTO PUBLICO	249	FUGA DE PRESO
190	ESTUPRO	220	FALSIFICACAO DO SELO OU SINAL PUBLICO	250	FUGA DO LAR
191	ESTUPRO COM RESULTADO MORTE	221	FALSIFICACAO, CORRUPCAO, ADULTERACAO OU ALTERACAO DE PRODUTO DESTINADO A FINS TERAPEUTICOS OU MEDICINAIS	251	FURTO
192	ESTUPRO DE VULNERAVEL	222	FALSIFICACAO, CORRUPCAO, ADULTERACAO OU ALTERACAO DE SUBSTANCIA OU PRODUTOS ALIMENTICIOS	252	GERIR FRAUDULENTAMENTE INSTITUICAO FINANCEIRA
193	ESTUPRO DE VULNERAVEL COM RESULTADO MORTE	223	FALSO ALARME	253	HOMICIDIO
194	EVASAO HOSPITALAR	224	FALSO TESTEMUNHO OU FALSA PERICIA	254	HOMICIDIO CULPOSO
195	EXCESSO DE EXACAO - TRIBUTO OU CONTRIBUICAO SOCIAL INDEVIDA	225	FALTA DE COMUNICACAO AOS CONSUMIDORES SOBRE PERICULOSIDADE DE PRODUTO	255	HOMICIDIO NO TRANSITO
196	EXECUTAR LAVRA DE RECURSOS MINERAIS SEM AUTORIZACAO	226	FALTA DE TRANSFERENCIA DE PROPRIEDADE DE VEICULO	256	IMPEDIR ACAO DE AUTORIDADE JUDICIARIA
197	EXERCICIO ARBITRARIO DAS PROPRIAS RAZOES	227	FATO ATIPICO	257	IMPORTUNACAO SEXUAL
198	EXERCICIO ARBITRARIO OU ABUSO DE PODER	228	FAVORECER PROSTITUICAO OU EXPLORACAO SEXUAL DE CRIANCA, ADOLESCENTE OU VULNERAVEL	258	INCENDIO
199	EXERCICIO ILEGAL DA MEDICINA	229	FAVORECIMENTO A PROSTITUICAO	259	INCENDIO COM MORTE
200	EXERCICIO ILEGAL DA MEDICINA, ARTE DENTARIA OU FARMACEUTICA	230	FAVORECIMENTO PESSOAL	260	INCENDIO CRIMINOSO
201	EXERCICIO ILEGAL DE PROFISSAO	231	FAVORECIMENTO REAL	261	INCENDIO EM LAVOURA, PASTAGEM, MATA OU FLORESTA
202	EXERCICIO ILEGAL DE PROFISSAO OU ATIVIDADE	232	FRAUDAR LICITACAO INSTAURADA PARA AQUISICAO OU VENDA DE BENS OU MERCADORIAS	262	INCITACAO AO CRIME
203	EXPLORAR RECURSOS MINERAIS SEM AUTORIZACAO	233	FRAUDAR PRECOS	263	INDUZIMENTO A FUGA
204	EXPOR A PERIGO A INTEGRIDADE E A SAUDE DO IDOSO	234	FRAUDE A CREDORES	264	INDUZIMENTO A FUGA, ENTREGA ARBITRARIA OU SONEGACAO DE INCAPAZES
205	EXTORSAO	235	FRAUDE NA ENTREGA DE COISA	265	INDUZIMENTO DO CONSUMIDOR A ERRO
206	EXTORSAO MEDIANTE SEQUESTRO	236	FRAUDE NO COMERCIO	266	INDUZIMENTO, INSTIGACAO OU AUXILIO A SUICIDIO
207	EXTRACAO DE AREIA, CAL, OU QUALQUER OUTRO MINERAL	237	FRAUDE NO PAGAMENTO POR MEIO DE CHEQUE	267	INFORMACAO FALSA A AUTORIDADE FAZENDARIA
208	EXTRACAO ILEGAL DE MADEIRA	238	FRAUDE PARA RECEBIMENTO DE INDENIZACAO OU VALOR DE SEGURO	268	INFRACAO DE MEDIDA SANITARIA PREVENTIVA
209	EXTRAVIO	239	FRAUDE PROCESSUAL	269	INJURIA
210	FABRICO, COMERCIO OU DETENCAO DE ARMAS OU MUNICAO	240	FRAUDES E ABUSOS NA FUNDACAO OU ADMINISTRACAO DE SOCIEDADE POR ACOES	270	INJURIA RACIAL

Tabela 8 – Lista de consolidados (cont.)

271	INSERCAO DE DADOS FALSOS EM SISTEMA DE INFORMACOES	301	MANIPULAR SUBSTANCIA TOXICA, PERIGOSA AO MEIO AMBIENTE, EM DESACORDO COM AS EXIGENCIAS LEGAIS	331	OMISSAO NA GUARDA DE ANIMAIS
272	INSTRUMENTO DE EMPREGO USUAL NA PRATICA DE FURTO	302	MANTER CASA DE PROSTITUICAO	332	OUTRAS FRAUDES
273	INTERCEPTACAO TELEFONICA	303	MAUS TRATOS	333	OUTRAS FRAUDES - FRAUDE A EXECUCAO
274	INTERRUPCAO DE FORNECIMENTO DE UTILIDADE PUBLICA	304	MAUS TRATOS A ANIMAIS	334	PARALIZACAO DE TRABALHO
275	INTRODUCAO OU ABANDONO DE ANIMAIS EM PROPRIEDADE ALHEIA	305	MAUS TRATOS COM RESULTADO MORTE	335	PARTICIPACAO EM CORRIDA, DISPUTA OU COMPETICAO AUTOMOBILISTICA NAO AUTORIZADA
276	INUTILIZAR MATERIA PRIMA OU MERCADORIA	306	MEDIACAO PARA SERVIR A LASCIVIA DE OUTREM	336	PECULATO
277	INVASAO DE DISPOSITIVO INFORMATICO	307	MERCADORIA EM DESACORDO COM DESCRICAO LEGAL	337	PERDA OU EXTRAVIO DE ARMA DE FOGO
278	INVASAO DE DOMICILIO	308	MISTURAR GENEROS E MERCADORIAS	338	PERIGO DE CONTAGIO VENEREO
279	INVASAO DE ESTABELECIMENTO INDUSTRIAL, COMERCIAL OU AGRICOLA	309	MOEDA FALSA	339	PERIGO PARA VIDA E SAUDE DE OUTREM
280	INVASAO DE FAZENDA	310	MORTE A ESCLARECER	340	PERSEGUICAO
281	INVASAO DE TERRAS DA UNIAO, DOS ESTADOS E DOS MUNICIPIOS	311	MORTE A ESCLARECER COM INDICIO DE CRIME	341	PERTURBACAO DO SOSSEGO ALHEIO
282	INVASAO DE TERRENO CONSTRUIDO	312	MORTE A ESCLARECER SEM INDICIO DE CRIME	342	PESCA EM PERIODO DEFESO
283	INVASAO DE TERRENO CONSTRUIDO OBJETO DE FINANCIAMENTO DO SFHABITACAO	313	MORTE ACIDENTAL	343	PESSOA LOCALIZADA
284	JOGO DE AZAR	314	MORTE COM EXCLUDENTE DE ILICITUDE	344	PETRECHOS PARA FALSIFICACAO DE MOEDA
285	JOGO DO BICHO	315	MORTE NATURAL	345	PICHAR OU CONSPURCAR EDIFICACAO OU MONUMENTO URBANO
286	LATROCINIO	316	MORTE NO TRANSITO	346	PLACA CLONADA
287	LAVAGEM DE DINHEIRO	317	MORTE POR AFOGAMENTO	347	POLUICAO A NATUREZA
288	LESAO CORPORAL	318	MORTE POR INTERVENCAO DE AGENTE DO ESTADO	348	POLUICAO ATMOSFERICA
289	LESAO CORPORAL CULPOSA	319	MOTIM DE PRESOS	349	POLUICAO SONORA
290	LESAO CORPORAL SEGUIDA DE MORTE	320	MULTA INDEVIDA	350	PORTE DE ARMA BRANCA
291	LESAO DECORRENTE DE INTERVENCAO DE AGENTE DO ESTADO	321	NAO OBSERVAR A ORDEM EM QUE OS ELEITORES DEVEM SER CHAMADOS A VOTAR	351	PORTE ILEGAL DE ARMA DE FOGO
292	LESAO NO TRANSITO	322	NAO PROVER AS NECESSIDADES BASICAS DO IDOSO	352	PORTE ILEGAL DE ARMA DE FOGO DE USO PERMITIDO
293	LGBTFOBIA	323	NEGAR EMISSAO DE NOTA FISCAL	353	PORTE ILEGAL DE ARMA DE FOGO DE USO RESTRITO
294	LOCALIZACAO DE PESSOA	324	OBRAS/SERVICOS QUE CAUSEM POLUICAO	354	POSSE IRREGULAR DE ARMA DE FOGO DE USO PERMITIDO
295	LOCALIZACAO DE VEICULO	325	OBTER OU TENTAR OBTER GANHOS ILICITOS EM DETRIMEN.	355	POSSE OU PORTE DE ARMA DE FOGO DE USO RESTRITO
296	LOTEAMENTO SEM AUTORIZACAO LEGAL	326	OBTER, MEDIANTE FRAUDE, FINANCIAMENTO EM INSTITUICAO FINANCEIRA	356	POSSE OU PORTE ILEGAL DE ARMA DE FOGO DE USO RESTRITO
297	MANDADO DE BUSCA E APREENSAO	327	OFERECER, DISPONIBILIZAR, TRANSMITIR, PUBLICAR POR QUALQUER MEIO, INCLUSIVE SISTEMA DE INFORMATICA OU TELEMATICO, FOTOGRAFIA, VIDEO OU OUTRO REGISTRO QUE CONTENHA CENA DE SEXO EXPLICITO ENVOLVENDO CRIANCA OU ADOLESCENTE	357	POSSE OU PORTE ILEGAL DE ARMA FOGO DE USO RESTRITO
298	MANDADO DE BUSCA E APREENSAO DE ADOLESCENTE	328	OMISSAO DE CAUTELA	358	POSSIBILIDADE DE APLICACAO CUMULATIVA DE OUTRAS SANCOES
299	MANDADO DE PRISAO	329	OMISSAO DE COMUNICACAO DE CRIME DE ACAO PUBLICA EXERCICIO DE MEDICINA OU VIGILANCIA SANITARIA	359	PRATICAR A DISCRIMINACAO OU PRECONCEITO DE RACA, COR, ETNIA, RELIGIAO OU PROCEDENCIA NACIONAL
300	MANIPULACAO DE PRODUTOS PERIGOSOS SEM AUTORIZACAO	330	OMISSAO DE SOCORRO	360	PRESO FORAGIDO

Tabela 9 – Lista de consolidados.

361	PREVARICACAO	396	SEM PERMISSAO AUTORIDADE COMPETENTE - VENDA DE ANIMAIS	431	ULTRAJE A CULTO E IMPEDIMENTO OU PERTURBACAO DE ATO A ELE RELATIVO
362	PRODUTO ASSINALADO COM MARCA ILICITAMENTE REPRODUZIDA	397	SEQUESTRO E CARCERE PRIVADO	432	USO DE DOCUMENTO FALSO
363	PRODUZIR, FOTOGRAFAR, FILMAR POR QUALQUER MEIO, CENA DE SEXO EXPLICITO ENVOLVENDO CRIANCA OU ADOLESCENTE	398	SONEGACAO DE PAPEL OU OBJETO DE VALOR PROBATORIO	433	USO INDEVIDO DE NOME COMERCIAL
364	PRODUZIR,REPRODUZIR,DIRIGIR, FOTOGRAFAR,FILMAR	399	SONEGAR INSUMOS OU BENS	434	USURA PECUNIARIA OU REAL - COBERANCA DE JUROS SUPERIORES A TAXA PERMITIDA POR LEI
365	PROMOV.NA VIA COMPET. EVENT.ORGANIZ.EXIB...	400	SONEGAR MERCADORIAS	435	USURPACAO
366	PROMOVER TUMULTO, PRATICAR OU INCITAR A VIOLENCIA, OU INVADIR LOCAL RESTRITO AOS COMPETIDORES EM EVENTOS ESPORTIVOS	401	SUBMETER CRIANCA OU ADOLESCENTE A EXPLORACAO SEXUAL	436	USURPACAO DE FUNCAO PUBLICA
367	PROMOVER, CONSTITUIR, FINANCIAR OU INTEGRAR, PESSOALMENTE OU POR INTERPOSTA PESSOA, ORGANIZACAO CRIMINOSA	402	SUBMETER CRIANCA OU ADOLESCENTE A VEXAME OU A CONSTRANGIMENTO	437	VANDALISMO
368	PROPAGANDA ENGANOSA	403	SUBMETER PESSOA SOB SUA GUARDA A VEXAME OU A CONSTRANGIMENTO NAO AUTORIZADO EM LEI	438	VENDA DE VEICULO SEM TRANSFERENCIA DE PROPRIEDADE
369	PROPRIEDADE	404	SUBTRACAO DE INCAPAZ	439	VENDER BEBIDA ALCOOLICA A MENOR DE IDADE
370	PROVOCACAO DE TUMULTO	405	SUICIDIO	440	VENDER PRODUTOS IMPROPRIOS PARA CONSUMO
371	PROVOCAR A ALTA OU BAIXA DE PRECOS DE MERCADORIAS, TITULOS PUBLICOS, OU SALARIOS POR MEIO DE NOTICIAS FALSAS	406	SUPRESSAO DE DOCUMENTOS	441	VENDER, FORNECER AINDA QUE GRATUITAMENTE, DE QUALQUER FORMA, A CRIANCA OU ADOLESCENTE ARMA, MUNICAO OU EXPLOSIVO
372	PROVOCAR INCENDIO EM FLORESTA	407	SUSPEITO DE ESTUPRO	442	VENDER, FORNECER AINDA QUE GRATUITAMENTE, ENTREGAR, DE QUALQUER FORMA, A CRIANCA OU ADOLESCENTE, PRODUTOS CUJOS COMPONENTES POSSAM CAUSAR DEPENDENCIA FISICA OU PSIQUICA
373	QUEBRA DE CONTRATO	408	SUSPEITO DE ESTUPRO DE VULNERAVEL	443	VENDER, FORNECER, BEBIDA ALCOOLICA A MENOR DE IDADE
374	RACISMO	409	TENTATIVA DE ESTELIONATO	444	VENDER, FORNECER, BEBIDA ALCOOLICA A MENOR DE IDADE
375	RECEPTACAO	410	TENTATIVA DE ESTUPRO	445	VIAS DE FATO
376	RECUPERACAO DE VEICULOS	411	TENTATIVA DE ESTUPRO DE VULNERAVEL	446	VILIPENDIO A CADAVER
377	RECUSA DE DADOS SOBRE A PROPRIA IDENTIDADE	412	TENTATIVA DE EXTORSAO	447	VIOLACA DE DOMICILIO
378	REDUCAO A CONDICAO DE ESCRAVO	413	TENTATIVA DE FUGA	448	VIOLACAO DA SUSPENSAO OU PROIBICAO DE SE OBTER A PERMISSAO OU HABILITACAO PARA DIRIGIR VEICULO AUTOMOTOR
379	REGISTRO NAO AUTORIZADO DA INTIMIDADE SEXUAL	414	TENTATIVA DE FURTO	449	VIOLACAO DE COMUNICACAO COM ABUSO DE FUNCAO
380	REPARACAO COM PECAS USADAS	415	TENTATIVA DE HOMICIDIO	450	VIOLACAO DE CORRESPONDENCIA
381	RESISTENCIA	416	TENTATIVA DE LATROCINIO	451	VIOLACAO DE DIREITO AUTORAL
382	RESISTENCIA A PRISAO	417	TENTATIVA DE LESAO CORPORAL	452	VIOLACAO DE DOMICILIO
383	RETENCAO DE DOCUMENTO	418	TENTATIVA DE MOTIM	453	VIOLACAO DE LACRE DE INSTITUICOES PUBLICAS
384	RETER CARTAO MAGNETICO DE CONTA BANCARIA	419	TENTATIVA DE ROUBO	454	VIOLACAO DE SEGREDO PROFISSIONAL
385	RETER CARTAO MAGNETICO DE CONTA BANCARIA COM OBJETIVO DE ASSEGURAR RECEBIMENTO DE DIVIDA	420	TENTATIVA DE SEQUESTRO	455	VIOLACAO DE SIGILO FUNCIONAL
386	RIXA	421	TENTATIVA DE SUICIDIO	456	VIOLACAO SEXUAL MEDIANTE FRAUDE
387	ROUBO	422	TENTATIVA DE VIOLACAO DE DOMICILIO	457	VIOLAR OU TENTAR VIOLAR O SIGILO DO VOTO
388	RUFIANISMO	423	TERMO DE GARANTIA NAO PERMITIDO	458	VIOLENCIA DOMESTICA E FAMILIAR CONTRA A MULHER - DEFINICAO
389	SATISFACAO DE LASCIVIA MEDIANTE PRESENCA DE CRIANCA OU ADOLESCENTE	424	TORTURA	459	VIOLENCIA FISICA, QUALQUER CONDUTA QUE OFENDA SUA INTEGRIDADE OU SAUDE CORPORAL
390	SEM PERMISSAO AUTORIDADE COMPETENTE - CACA DURANTE A NOITE	425	TRAFEGAR EM VELOCIDADE INCOMPATIVEL C/ A SEGURANCA	460	VIOLENCIA MORAL, QUALQUER CONDUTA QUE CONFIGURE CALUNIA, DIFAMACAO OU INJURIA
391	SEM PERMISSAO AUTORIDADE COMPETENTE - CACA EM UNIDADE DE CONSERVACAO	426	TRAFEGAR EM VELOCIDADE INCOMPATIVEL COM A SEGURANCA	461	VIOLENCIA PATRIMONIAL
392	SEM PERMISSAO AUTORIDADE COMPETENTE - CACA PROFISSIONAL	427	TRAFICO DE DROGAS	462	VIOLENCIA PSICOLOGICA
393	SEM PERMISSAO AUTORIDADE COMPETENTE - CRIAR ANIMAIS	428	TRANSACOES BANCARIAS	463	VIOLENCIA PSICOLOGICA, QUALQUER CONDUTA QUE LHE CAUSE DANO EMOCIONAL, DIMINUICAO DA AUTO-ESTIMA, LHE PREJUDIQUE E PERTURBE O PLENO DESENVOLVIMENTO, VISE DEGRADAR OU CONTROLAR SUAS ACOES, COMPORTAMENTOS, CRENCAS E DECISOES, MEDIANTE AMEACA, CONSTRANGIMENTO
394	SEM PERMISSAO AUTORIDADE COMPETENTE - ESPECIE AMEACADA DE EXTINCAO	429	TRANSFERENCIA IRREGULAR DE GUARDA		
395	SEM PERMISSAO AUTORIDADE COMPETENTE - MATAR ANIMAIS	430	TURBACAO		