UNIVERSIDADE VIRTUAL DO ESTADO DE SÃO PAULO

Damaris dos Santos Neves
Felippe Martiniano de Oliveira
Gabriel Eberle Carvalho Schmidt
Johnny Matos
Márcio Lima Brunelli
Renan do Prado Bernardes de Moura
Talyson Almeida de Araújo Santos

Ciência de Dados Aplicada à Segurança Pública: Modelagem Preditiva de Ocorrências Criminais a partir de dados da Secretaria de Segurança Pública do Estado de São Paulo

São Paulo
2025

UNIVERSIDADE VIRTUAL DO ESTADO DE SÃO PAULO

Ciência de Dados Aplicada à Segurança Pública: Modelagem Preditiva de Ocorrências Criminais a partir de dados da Secretaria de Segurança Pública do Estado de São Paulo

Monografia apresentada como requisito parcial para obtenção de título de bacharel de Ciência da Dados da Universidade Virtual do Estado de São Paulo (UNIVESP).

São Paulo
2025

BRUNELLI, Márcio; MOURA, Renan; NEVES, Damaris; OLIVEIRA, Felippe; MATOS, Johnny; SANTOS, Talyson; SCHMIDT, Gabriel; **Ciência de Dados Aplicada à Segurança Pública: Modelagem Preditiva de Ocorrências Criminais a partir de dados da Secretaria de Segurança Pública do Estado de São Paulo**. 45f. Trabalho de Conclusão de Curso. Ciência de Dados – Universidade Virtual do Estado de São Paulo. Tutor: Maria Eduarda Guillen Diomasio. 2025.

RESUMO

Diante do advento da tecnologia, a aplicação da ciência de dados na segurança pública tem se mostrado uma abordagem promissora para a otimização da alocação de recursos e a prevenção de crimes. Este estudo propõe a utilização de modelos de aprendizado de máquina para a predição de ocorrências criminais, analisando padrões históricos a partir de dados públicos da Secretaria de Segurança Pública do Estado de São Paulo (SSP-SP). Para isso, foram testados os algoritmos Naive Bayes e Random Forest, com o objetivo de estimar a probabilidade de crimes com base em variáveis como localização, horário e tipo de ocorrência. A metodologia incluiu a coleta e tratamento de dados, implementação dos modelos preditivos e avaliação de desempenho com diversas métricas. Os resultados indicam a possibilidade de uso de técnicas supervisionadas de aprendizado de máquina para a predição de crimes.. No entanto, o poder dessas aplicações é limitado. Apenas com a junção de outras técnicas computacionais e de outras áreas do conhecimento, é possível que se tenha uma abordagem sinérgica e integrada para o combate ao crime. Possíveis sinergias, como a integração de tecnologias avançadas, incluindo câmeras inteligentes e drones, para reforçar o monitoramento e a segurança pública, são discutidas ao final do texto.

PALAVRAS-CHAVE: Segurança pública, aprendizado de máquina, modelagem preditiva, crime, ciência de dados.

BRUNELLI, Márcio; MOURA, Renan; NEVES, Damaris; OLIVEIRA, Felippe; MATOS, Johnny; SANTOS, Talyson; SCHMIDT, Gabriel; **Data Science Applied to Public Security: Predictive Modeling of Criminal Incidents at the São Paulo State Public Security Department.** 45p. Final Thesis. Data Science – Universidade Virtual do Estado de São Paulo. Tutor: Maria Eduarda Guillen Diomasio. 2025.

ABSTRACT

In the face of technological advancements, the application of data science in public security has proven to be a promising approach for optimizing resource allocation and preventing crimes. This study proposes the use of machine learning models for predicting criminal incidents, analyzing historical patterns from public data provided by the São Paulo State Public Security Department (SSP-SP). The Naive Bayes and Random Forest algorithms were tested to estimate the probability of crimes based on variables such as location, time, and crime type. The methodology included data collection and preprocessing, implementation of predictive models, and performance evaluation using accuracy metrics. The results indicate that supervised machine learning techniques may be a viable way to anticipate crimes. However, these techniques could not solely tackle the entire problem. Integrated approaches, which would include other computational techniques as well as other knowledge realms, could bring more synergetic and effective results. Therefore, we also discuss the limitations of the approach and possible improvements, such as the integration of advanced technologies, including smart cameras and drones, to strengthen monitoring and public security.

PALAVRAS-CHAVE: Public security, machine learning, predictive modeling, crime, data science.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE TRECHOS DO CÓDIGO-FONTE

LISTAS DE ABREVIATURAS E SIGLAS

AM……... Aprendizado de Máquina
CD……... Ciência de Dados
IA……. Inteligência Artificial
LGPD……... Lei Geral de Proteção de Dados
SSP-SP……Secretaria de Segurança Pública de São Paulo

SUMÁRIO

1. INTRODUÇÃO

O Painel Estatístico da Secretaria de Segurança Pública de São Paulo (SSP-SP, 2025) apresenta dados detalhados sobre a criminalidade no estado, abrangendo ocorrências de diversos tipos de crimes, como homicídios, furtos e roubos. Esses dados são fundamentais para compreender o cenário da segurança pública e embasar a formulação de políticas de prevenção e controle da criminalidade.

A SSP-SP é responsável pela coordenação das políticas públicas voltadas à segurança, incluindo a prevenção, controle e combate à criminalidade, além da proteção da sociedade. Para isso, emprega uma abordagem estratégica que integra recursos humanos, sistemas judiciais e tecnologias emergentes, buscando adaptar-se continuamente às dinâmicas sociais em evolução.

EGBERT e LEESE (2021) informam que aplicar Ciência de Dados (CD) não se trata de uma completa novidade, porém algo necessário para acompanhar os avanços tecnológicos e da sociedade. No contexto da segurança pública, a predição de crimes se mostra como ferramenta alternativa, e possivelmente crucial, para otimizar o uso de recursos e direcionar ações preventivas a áreas e momentos de maior risco.

Diante desse cenário, este trabalho propõe o uso de técnicas de Aprendizado de Máquina (AM) para prever ocorrências criminais, utilizando dados históricos de crimes em áreas urbanas. O objetivo principal é avaliar a eficácia de diferentes algoritmos de AM na previsão de crimes, proporcionando *insights* que possam contribuir para o aprimoramento das estratégias de segurança pública.

Por fim, este estudo também discute direções futuras para a segurança pública baseada em inteligência artificial. Dessa forma, busca-se não apenas demonstrar a viabilidade da ciência de dados na segurança pública, mas também indicar caminhos para o avanço dessa tecnologia no combate à criminalidade e contribuir com a literatura sobre as aplicações de dados em diversos setores.

2. OBJETIVOS

Neste capítulo serão apresentados os objetivos do presente trabalho acadêmico.

2.1. Objetivo Geral

Investigar a aplicação de técnicas de AM na predição de crimes, analisando a viabilidade e as limitações do uso de dados públicos para antecipação de ocorrências criminais e otimização da alocação de recursos da segurança pública.

2.2. Objetivos Específicos

Visando alcançar o objetivo geral, foram levantados objetivos específicos, que visam tirar o melhor proveito da participação dos diferentes profissionais que realizaram esta pesquisa:

- Coletar e analisar bases de dados públicas relacionadas a ocorrências criminais em São Paulo, identificando padrões relevantes;
- Implementar modelos de aprendizado de máquina baseados em classificação para analisar a probabilidade de crimes com base em características específicas.

- Comparar o desempenho dos algoritmos de classificação para avaliar sua precisão e aplicabilidade no contexto da segurança pública;

- Discutir as limitações e desafios da modelagem preditiva em algoritmos de classificação, incluindo questões de viés nos dados e interpretabilidade dos modelos;

- Apresentar possíveis direções futuras, como a integração de monitoramento por câmeras inteligentes e drones na segurança pública.

3. JUSTIFICATIVA E DELIMITAÇÃO DO PROBLEMA

A SSP-SP, por meio do seu Grupo de Tecnologia da Informação, é responsável por coordenar, monitorar e operar a infraestrutura de tecnologia da informação, garantindo a continuidade e a evolução dos serviços tecnológicos essenciais à administração da segurança pública. Entre suas atribuições, destacam-se (SSP-PS, 2025):

- Gerenciar contratos para assegurar a execução das obrigações contratuais;
- Executar a manutenção dos serviços contínuos que viabilizam o funcionamento da SSP-SP;
- Elaborar estudos e propor ações para a evolução do Parque Tecnológico da SSP-SP.

Neste contexto, a crescente demanda por soluções inovadoras no campo da segurança pública impulsiona a necessidade de integrar novas tecnologias para a prevenção e o combate à criminalidade. A utilização de AM para a predição de crimes surge como uma proposta para otimizar as operações policiais, direcionando os recursos de maneira mais eficiente e adaptada à dinâmica criminosa das grandes cidades. Mediante essa investigação, procura-se contribuir para o avanço do uso da CD na segurança pública, cujo fim seria o melhor uso de recursos públicos e maior bem estar social, devido a um possível aumento nessa área da administração pública.

4. FUNDAMENTAÇÃO TEÓRICA

A crescente utilização de técnicas de AM para a predição de crimes tem atraído o interesse de pesquisadores e órgãos governamentais devido ao seu potencial para otimizar a alocação de recursos e antecipar a ocorrência de crimes. Diversos estudos abordam o impacto e os desafios dessa abordagem, oferecendo uma compreensão abrangente dos métodos utilizados e suas aplicações práticas (CHOULDECHOVA, 2017; EGBER E LIESSE, 2019; RIBEIRO, 2020; SANTOS JÚNIOR, 2024). Nesta seção, serão discutidos os principais conceitos relacionados à aplicação da CD e AM à temática desta monografia.

4.1. Legislações

Antes da aplicação dos estudos em como a utilização de CD e tecnologias informatizadas contribuem na prevenção de crimes e eventual garantia da manutenção da segurança, faz-se necessário entender a história dos esforços públicos para a oferta de segurança pública para toda a população do estado de São Paulo.

A história da SSP-SP está diretamente relacionada à formação da Secretaria da Segurança Pública do Governo do Estado, cujas origens datam de mais de um século, mais precisamente a data de 17 de setembro de 1906, quando da promulgação da Lei nº 1006/1906, pelo então presidente do Estado de São Paulo, Jorge Tibiriçá. Essa lei instituiu a Secretaria de Estado dos Negócios da Justiça, instituindo o cargo de Secretário para esta pasta ao extinguir o então vigente cargo de Chefe de Polícia do Estado. Com esta lei, o policiamento fora distribuído pelo Estado de São Paulo, com delegados sendo direcionados aos municípios e comarcas para a execução do trabalho de policiamento e manutenção da paz.

Em 1930, com a publicação do Decreto 4.789 de 05 de dezembro de 1930, a Secretaria foi desmembrada em duas entidades, a Secretaria de Estado dos Negócios da Justiça e a Secretaria de Estado dos Negócios de Segurança Pública. Devido a turbulências políticas no período, apesar de dados históricos serem escassos para justificar os motivos, o órgão então conhecido como Secretaria da Segurança Pública foi extinto em 1931, um ano antes da eclosão da Revolução Constitucionalista de 1932. Somente em 1934 a Secretaria foi

novamente restituída. Porém, em 1939, durante a ditadura do Estado Novo, a Secretaria foi novamente extinta. E somente no ano de 1941 ela foi final e definitivamente estabelecida graças ao decreto-lei nº 12.163 de 10 de setembro de 1941, durante o governo de Fernando Costa, que ocupou o cargo de Interventor Federal. Graças ao decreto-lei, a Secretaria de Estado dos Negócios da Segurança Pública foi definitivamente constituída, tornando-se o órgão atualmente conhecido como a Secretaria de Segurança Pública do Estado de São Paulo.

A estrutura atual da SSP-SP comporta todos os órgãos policiais em ação no Estado de São Paulo, sendo estes a Polícia Científica, Polícia Civil, Polícia Militar, Corpo de Bombeiros, CONSEG (Conselhos Comunitários de Segurança) e a Ouvidoria das Polícias. Cada um desses órgãos é regido através de legislações específicas, que estabelecem normas de conduta e regramento para seus corretos funcionamentos em acordo com a legislação estadual e federal. A Lei Federal 13.675 de 11 de junho de 2018, que estabelece a Política Nacional de Segurança Pública e Defesa Social (PNSPDS) e institui o Sistema Único de Segurança Pública (Susp), fornece as bases legais para as legislações estatuais quanto à Segurança Pública, em especial na cooperatividade entre os diferentes órgãos nas esferas estaduais, de forma a promover a Segurança Pública em âmbito nacional, com seus princípios descritos em seu Artigo 4º:

"São princípios da PNSPDS:

I - respeito ao ordenamento jurídico e aos direitos e garantias individuais e coletivos;

II - proteção, valorização e reconhecimento dos profissionais de segurança pública;

III - proteção dos direitos humanos, respeito aos direitos fundamentais e promoção da cidadania e da dignidade da pessoa humana;

IV - eficiência na prevenção e no controle das infrações penais;

V - eficiência na repressão e na apuração das infrações penais;

VI - eficiência na prevenção e na redução de riscos em situações de emergência e desastres que afetam a vida, o patrimônio e o meio ambiente;

VII - participação e controle social;

VIII - resolução pacífica de conflitos;

IX - uso comedido e proporcional da força;

X - proteção da vida, do patrimônio e do meio ambiente;

XI - publicidade das informações não sigilosas;

XII - promoção da produção de conhecimento sobre segurança pública;

XIII - otimização dos recursos materiais, humanos e financeiros das instituições;

XIV - simplicidade, informalidade, economia procedimental e celeridade no serviço prestado à sociedade;

XV - relação harmônica e colaborativa entre os Poderes;

XVI - transparência, responsabilização e prestação de contas."

Dada a promulgação desta lei, o Governo do Estado de São Paulo elaborou o Plano Estadual de Segurança Pública, aprovado pelo Decreto Nº 65.657, de 27 de abril de 2021, que visa executar a Política Estadual de Segurança Pública, que estabelece a Missão, Visão, Valores e as Diretrizes para a SSP-SP, sendo essas diretrizes descritas conforme o Artigo 5º do referido Decreto:

"As Diretrizes para a Secretaria de Segurança Pública do Estado de São Paulo:

I - Buscar a redução da criminalidade e a melhoria da percepção de segurança com efetividade, respeitando os direitos e a dignidade da pessoa humana;

II - Melhorar a gestão pública com indicadores, transparência, simplicidade e integração;

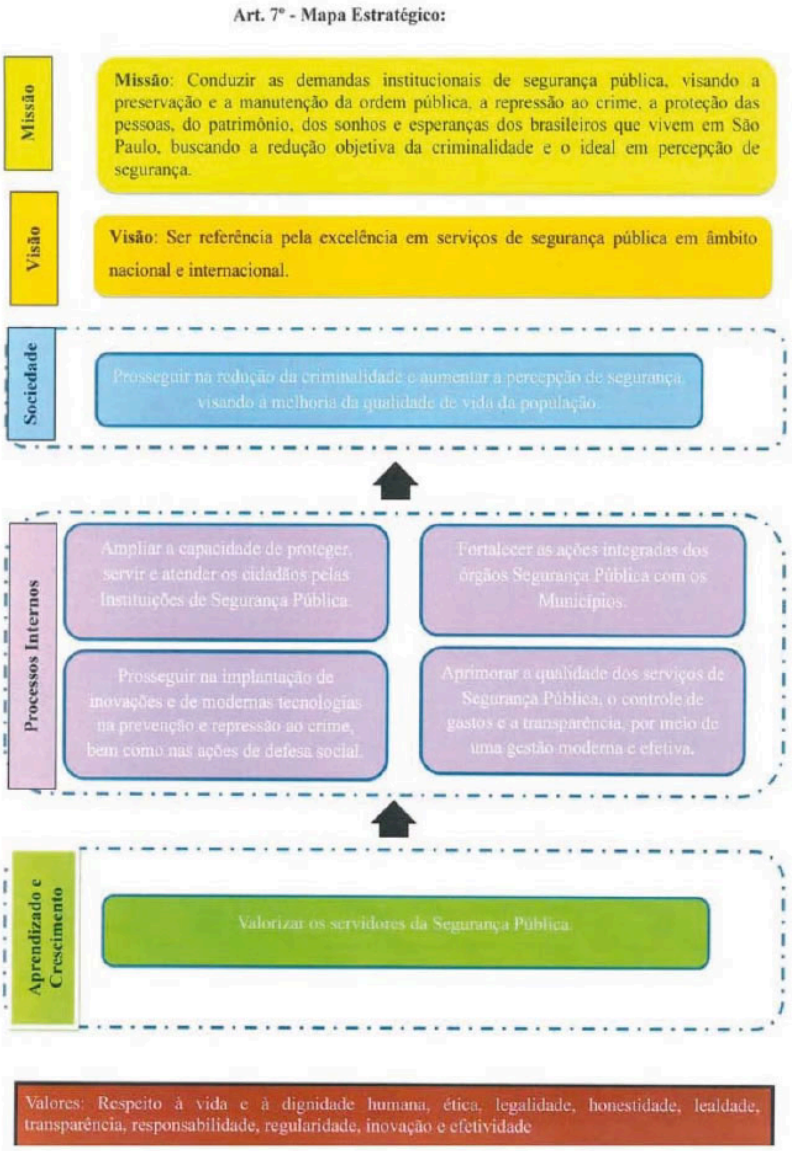
III - Ampliar o sentimento de orgulho e de comprometimento dos servidores em relação às Instituições da Segurança Pública;

IV - Estruturar os Objetivos e Estratégias no trinômio: Inteligência, Tecnologia e Valorização das pessoas que trabalham nas Instituições de segurança Pública.

V - Integrar planejamentos e ações, internamente, com outros órgãos e com as estruturas sociais."

O Decreto, em conjunto com a Resolução SSP-99 de 25 de novembro de 2019 também divulgou o mapa estratégico da Política Estadual de Segurança Pública, conforme segue:

Figura 1: Mapa Estratégico da Política Estadual de Segurança Pública



De acordo com o artigo nº 144 da Constituição Federal de 1988, a garantia da Segurança Pública é um dever do Estado, com o parágrafo 6º deste artigo clarificando que os órgãos responsáveis pela manutenção da segurança pública subordinam-se aos Governadores Estaduais:

"§ 6º As polícias militares e os corpos de bombeiros militares, forças auxiliares e reserva do Exército subordinam-se, juntamente com as polícias civis e as polícias penais estaduais e distrital, aos Governadores dos Estados, do Distrito Federal e dos Territórios. (Redação dada pela Emenda Constitucional nº 104, de 2019)"

Tanto os textos da Constituição como da Legislação Federal e Estadual determinam o papel e o dever do Estado na manutenção da Segurança no território nacional, através da aplicação da legislação vigente e do disposto no Decreto-Lei Nº 2.848, de 7 de dezembro de 1940, que estabelece o Código Penal Brasileiro, documento que define todas as formas de crime reconhecidas pela República Federativa do Brasil e os instrumentos de aplicação de pena para cada crime definido.

A Segurança Pública, garantida e exercida pela esfera Estadual mediante determinações e regramento estabelecido pela esfera Federal não é, entretanto, responsabilidade exclusiva dessas esferas. As esferas municipais possuem papel fundamental, ainda que legislativamente insuficiente, na garantia da Segurança Pública. KHAN e ZANETIC (2006) discorrem sobre como os municípios da Região Metropolitana de São Paulo agiam dentro de suas esferas no combate à criminalidade. Entre as ações tomadas, diversos municípios criaram Secretarias de Segurança Municipais e estabeleceram Guardas Civis Municipais, cujas atuações, além de suas funções principais de vigilância patrimonial, incluíam a aplicação da legislação municipal, a citar como exemplo a Lei Seca instituída pelos 16 municípios da Região Metropolitana de São Paulo em 2002, que definia o regramento para venda e consumo de bebidas alcoólicas por estabelecimentos como bares e adegas.

Cada município dispõe de sua própria legislação para o combate à criminalidade em âmbitos administrativos, como o estabelecimento de "leis do silêncio" próprias, instituição de código tributário que define regras e punições para atividades comerciais irregulares, além da criação de programas sociais para combate ao consumo de álcool, drogas, fomento ao esporte e a educação, por exemplo. Essas ações visam o combate à criminalidade, não através do uso de força policial, sendo essa uma competência do Estado, mas no trabalho de educação e conscientização da população em prol da boa vivência e manutenção da paz dentro da esfera municipal.

A garantia de segurança para a população vai além da atuação direta da Polícia e de ações legislativas Municipais. Graças a expansão contínua e rápida dos sistemas informatizados, em especial com o advento da CD, Big Data e *Business Inteligence*, a integração dos sistemas informatizados no campo da segurança pública passou de ser um objetivo a ser alcançado, tornando-se uma necessidade crescente. SALES e LUI (2023) apontam em seu estudo que o uso da tecnologia, como *Big Data* e AM, é de grande relevância na formulação de políticas de segurança pública, especialmente nas chamadas "Smart Cities". Pinheiro (2024) levanta a questão das problemáticas surgidas do uso dessas tecnologias para a Segurança Pública, levantando a questão sobre a privacidade individual e o limite de onde a segurança individual termina para a aplicação da segurança pública, especialmente no contexto de preconceito racial e regional e uso de videomonitoramento com reconhecimento facial. SALES e LUI (2023) apontam que os benefícios para a Segurança Pública na utilização de tecnologias de informação dependem diretamente do arcabouço institucional e jurídico de um país. Dessa forma, a realidade do uso de tais tecnologias têm impacto diferente em países como Brasil, Estados Unidos e China.

De fato, apesar de vivermos na chamada "Era dos Dados", em constante e exponencial crescimento, a maturidade no uso de tais tecnologias, tanto no mercado de trabalho, quanto no contexto desta pesquisa, no seu uso para a Segurança Pública, dependem da legislação e do amadurecimento legislativo do país para o uso, tratamento e gerenciamento adequado dos dados produzidos por tais tecnologias. O Brasil promulgou em 2018 a Lei Geral da Proteção de Dados, a LGPD, Lei 13.079 de 14 de agosto de 2018, que determina todo o tratamento para dados pessoais dos cidadãos brasileiros e entidades nacionais e internacionais que atuam de acordo com a legislação do país. Esta Lei foi criada para fornecer segurança e garantir a privacidade das pessoas no meio digital, instituindo que dados não podem ser manipulados, compartilhados e utilizados sem consentimento. A Lei, no entanto, abre exceção para o uso de dados pessoas no âmbito da segurança, conforme disposto em seu artigo 4º:

"Art. 4º Esta Lei não se aplica ao tratamento de dados pessoais:

(...)

III - realizado para fins exclusivos de:

a) segurança pública;

b) defesa nacional;

c) segurança do Estado; ou

d) atividades de investigação e repressão de infrações penais;"

A LGPD corrobora com os apontamentos de Sales e Lui (2023), exemplificando como a estrutura legislativa do país influencia no uso de tecnologias e tratamento de dados na aplicação da Segurança Pública. A Legislação vigente no país garante o direito à privacidade de todos os seus cidadãos, mas estipula que em questões de Segurança Pública, a manipulação de dados sem o consentimento do cidadão é autorizada.

4.2. Coleta de Dados

Conforme descrito no item 4.1, o Brasil instaurou em 2018 a Lei Geral da Proteção de Dados – LGPD – que define o regramento para uso e manipulação de dados sensíveis e pessoais em âmbito nacional. Conforme estipulado por essa lei, via seu artigo 7º, dados pessoais sensíveis só podem ser utilizados com o consentimento individual de cada pessoa aos quais eles pertencem, por tempo determinado, podendo o cidadão exigir que seus dados não sejam mais utilizados e/ou excluídos.

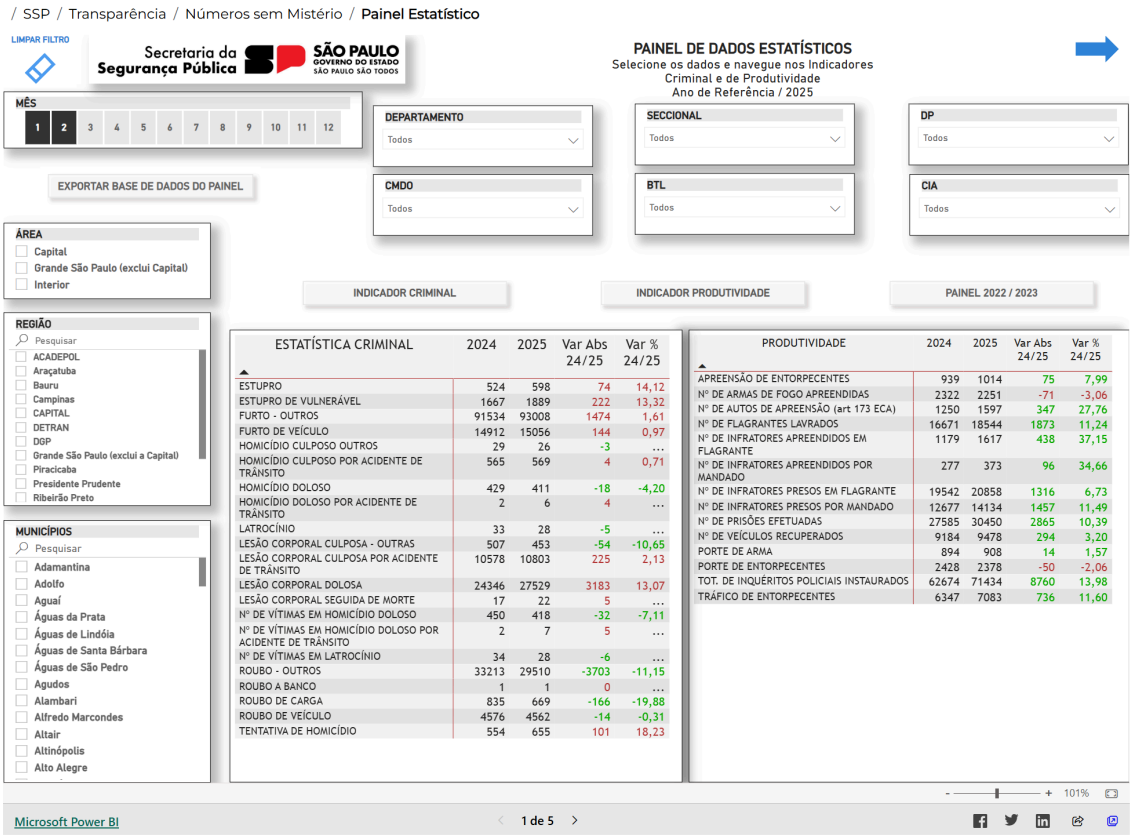
No entanto, conforme o Artigo 4º da LGPD estipula, em questões de Segurança Pública a utilização de dados pessoais não requer consentimento. Os artigos 23 e 25 da lei especificam que tais dados devem ser usados pelo Poder Público de forma a atender uma finalidade pública, com esses dados devendo estar devidamente estruturados e disponíveis à prestação de serviços públicos. A SSP-SP utiliza dados para registrar todas as ocorrências de crimes no Estado, com tabulação de dados sobre quantidades de crimes, agrupadas por períodos de 1 mês, 3 meses e 12 meses, estratificados por tipo de crime ocorrido e por município.

Os dados são publicados pela SSP-SP tanto de forma bruta, quanto com aplicação de filtros para visualização, além de fornecer um *dashboard* interativo em seu sítio na internet para

visualização e interação por parte dos usuários, sem a necessidade de cadastros para acessar os mesmos. Os dados brutos são acessíveis por qualquer pessoa, podendo ser lidos e gravados por qualquer cidadão. A seguir, detalharemos as formas como a SSP-SP garante a publicidade desses dados para a população.

Uma das principais ferramentas fornecidas pela SSP-SP é seu painel estatístico, acessível através do site oficial da entidade. O painel, criado com a utilização da ferramenta Power Bi da Microsoft, apresenta uma Dashboard interativa para visualização de dados estatísticos de crimes ocorridos dentro do Estado de São Paulo. A ferramenta, acessível gratuitamente, permite que usuários naveguem por diferentes paineis com informações apresentadas de forma intuitiva e interativa, permitindo que sejam aplicados filtros espaciais, temporais e até mesmo por tipo de crime, conforme ilustrado na imagem a seguir:

Figura 2: Painel Estatístico da SSP-SP.



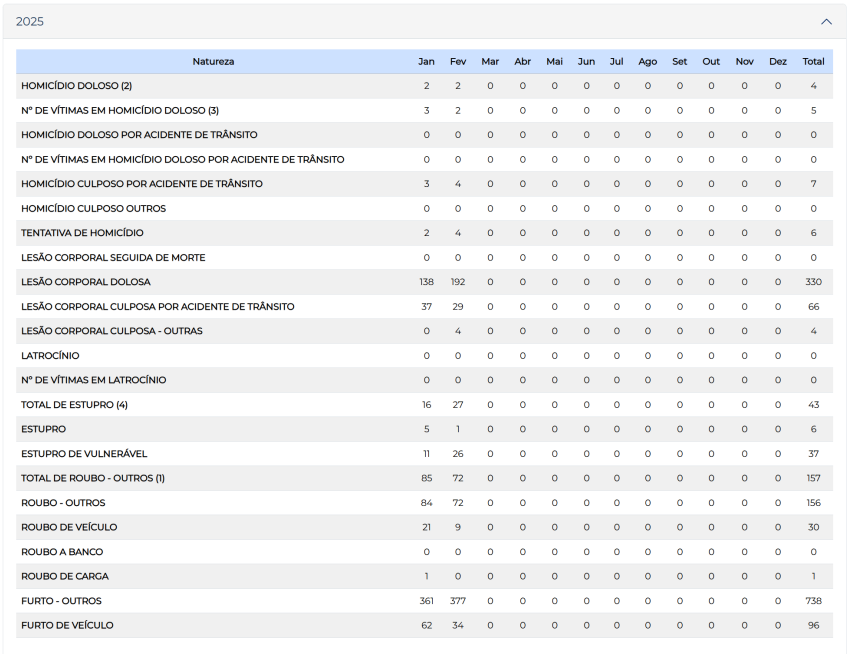
Fonte: https://www.ssp.sp.gov.br/estatistica/painel-estatistico

O próprio painel permite a exportação de todos os seus dados com um simples botão, que direciona para uma pasta armazenada no serviço em nuvem do Google Drive, oferecendo acesso a dados de ocorrências de crimes nos anos de 2023, 2024 e 2025. Assim, a SSP-SP garante a publicidade de dados referentes à segurança pública no estado.

Felizmente, graças a essa publicidade garantida pela SSP-SP, a coleta de dados para uso na construção do modelo proposto para este trabalho torna-se uma tarefa simples, pois a quantidade de dados disponibilizados é de alta escala e em diferentes níveis de tratamento, o que permite ao grupo flexibilidade em sua utilização, dentro do escopo proposto para esta monografia. Os dados obtidos através do Painel Estatístico, ainda que extensos, possuem limitações, em especial quanto à localização geográfica das ocorrências de crimes listadas, estratificando tais localidades através das áreas de atuação de delegacias e postos policiais espalhados por todo o estado.

Esta é, porém, apenas parte da forma como a SSP-SP fornece esses dados para a população. Todos os dados estatísticos do órgão, incluindo o acesso ao painel, estão agrupados em uma página específica de informações estatísticas, oferecendo a opção acessar dados em agrupamentos mensais, trimestrais ou realizar consultas específicas, o que permite a filtragem dos dados de ocorrências de crimes por ano, localidade, e infração cometida. E todos esses dados podem ser exportados em arquivos de extensão ”.xlsx”.

Figura 3: Crimes no ano de 2025 para a cidade de Mogi das Cruzes-SP.



2025

Natureza	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez	Total
HOMICÍDIO DOLOSO (2)	2	2	0	0	0	0	0	0	0	0	0	0	4
Nº DE VÍTIMAS EM HOMICÍDIO DOLOSO (3)	3	2	0	0	0	0	0	0	0	0	0	0	5
HOMICÍDIO DOLOSO POR ACIDENTE DE TRÂNSITO	0	0	0	0	0	0	0	0	0	0	0	0	0
Nº DE VÍTIMAS EM HOMICÍDIO DOLOSO POR ACIDENTE DE TRÂNSITO	0	0	0	0	0	0	0	0	0	0	0	0	0
HOMICÍDIO CULPOSO POR ACIDENTE DE TRÂNSITO	3	4	0	0	0	0	0	0	0	0	0	0	7
HOMICÍDIO CULPOSO OUTROS	0	0	0	0	0	0	0	0	0	0	0	0	0
TENTATIVA DE HOMICÍDIO	2	4	0	0	0	0	0	0	0	0	0	0	6
LESÃO CORPORAL SEGUIDA DE MORTE	0	0	0	0	0	0	0	0	0	0	0	0	0
LESÃO CORPORAL DOLOSA	138	192	0	0	0	0	0	0	0	0	0	0	330
LESÃO CORPORAL CULPOSA POR ACIDENTE DE TRÂNSITO	37	29	0	0	0	0	0	0	0	0	0	0	66
LESÃO CORPORAL CULPOSA - OUTRAS	0	4	0	0	0	0	0	0	0	0	0	0	4
LATROCÍNIO	0	0	0	0	0	0	0	0	0	0	0	0	0
Nº DE VÍTIMAS EM LATROCÍNIO	0	0	0	0	0	0	0	0	0	0	0	0	0
TOTAL DE ESTUPRO (4)	16	27	0	0	0	0	0	0	0	0	0	0	43
ESTUPRO	5	1	0	0	0	0	0	0	0	0	0	0	6
ESTUPRO DE VULNERÁVEL	11	26	0	0	0	0	0	0	0	0	0	0	37
TOTAL DE ROUBO - OUTROS (1)	85	72	0	0	0	0	0	0	0	0	0	0	157
ROUBO - OUTROS	84	72	0	0	0	0	0	0	0	0	0	0	156
ROUBO DE VEÍCULO	21	9	0	0	0	0	0	0	0	0	0	0	30
ROUBO A BANCO	0	0	0	0	0	0	0	0	0	0	0	0	0
ROUBO DE CARGA	1	0	0	0	0	0	0	0	0	0	0	0	1
FURTO - OUTROS	361	377	0	0	0	0	0	0	0	0	0	0	738
FURTO DE VEÍCULO	62	34	0	0	0	0	0	0	0	0	0	0	96

Fonte: https://www.ssp.sp.gov.br/estatistica/dados-mensais

Respeitando a legislação de Proteção de Uso de Dados, a LGPD, todos os dados divulgados publicamente respeitam a privacidade dos cidadãos, não expondo informações como nomes, números de documentos de registro ou endereços particulares. Esses dados, ainda que possam ser utilizados sem restrição, tendo em vista enquadrarem-se no disposto no Artigo 4º da Lei 13.079 de 14 de agosto de 2018, respeitam a privacidade individual dos cidadãos.

4.3. Fundamentos de AM

Uma definição canônica de AM reza: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E" (MITCHELL, 1997). Ou seja, AM é uma técnica computacional dita da área de IA que permite a um algoritmo melhorar sua performance em uma tarefa quanto maior sua experiência naquela mesma tarefa.

Além disso, AM é uma área da Ciência da Computação voltada à previsão de padrões e valores estatísticos com base em dados históricos, empregando métodos matemáticos e métricas específicas para aprimorar sua acurácia e desempenho. Para garantir que os resultados produzidos sejam representativos da realidade e isentos de vieses, é imprescindível a adoção de fundamentos teóricos sólidos e boas práticas metodológicas, tais como a curadoria criteriosa dos dados, a aplicação de técnicas robustas de validação e o monitoramento contínuo da performance dos modelos (GAMA, 2012).

Historicamente, o campo evoluiu a partir de fundamentos estatísticos e da IA, ganhando força com o avanço computacional e o crescimento dos dados digitais. O desenvolvimento de um modelo típico envolve etapas como coleta e tratamento de dados, divisão em conjuntos de treino e teste, escolha de algoritmos, avaliação por métricas específicas e ajustes de desempenho. Entre os principais desafios enfrentados estão o sobreajuste, a presença de viés nos dados, a interpretabilidade dos modelos e a necessidade de equilíbrio entre desempenho e generalização.

Esses princípios são transversais aos diferentes paradigmas do aprendizado de máquina, cada um com características e finalidades distintas, sendo eles:

4.3.1. Aprendizado supervisionado

Nesse paradigma, os algoritmos são treinados com base em um conjunto de dados rotulados, ou seja, cada entrada é associada a uma saída conhecida. O objetivo é aprender uma função que, a partir de novas entradas, consiga prever as saídas corretas. Esse tipo de aprendizado é amplamente utilizado em tarefas de classificação (como o reconhecimento de imagens) e regressão (como a previsão de preços). A eficácia do modelo depende fortemente da qualidade e da quantidade dos dados de treinamento.

4.3.2. Aprendizado não supervisionado

Ao contrário do supervisionado, esse tipo de aprendizado lida com dados sem rótulos. O algoritmo busca identificar padrões ocultos ou estruturas nos dados, como agrupamentos ou

correlações. Técnicas como clustering (agrupamento) e dimensionality reduction (redução de dimensionalidade) são comuns nesse contexto. Esse paradigma é útil para explorar dados, segmentar públicos ou detectar anomalias sem conhecimento prévio das categorias envolvidas.

4.3.3. Aprendizado por reforço

Nesse paradigma, o modelo aprende por meio da interação com um ambiente dinâmico, recebendo recompensas ou punições conforme suas ações. O objetivo é desenvolver uma política de ação que maximize a recompensa acumulada ao longo do tempo. Essa abordagem é inspirada no comportamento de aprendizado de seres vivos e é aplicada em áreas como robótica móvel, jogos eletrônicos e sistemas de recomendação em tempo real.

Esses fundamentos se refletem em aplicações diversas na sociedade atual, como diagnósticos médicos, manutenção preditiva na indústria, recomendação de produtos e detecção de fraudes financeiras.

4.4. Métricas para validação de aprendizado de máquinas

Após aplicação de modelos algoritmos em aprendizado de máquinas, torna-se necessário validar aquilo que está em análise. Segundo Provost e Fawcett (2013), o objetivo desta etapa é garantir que os resultados sejam válidos e confiáveis, antes de prosseguir. Os autores afirmam que testar os modelos em ambientes controlados apresenta-se mais vantajoso do que atuar e tomar decisões com dados imprecisos, o que tornaria a atividade economicamente inviável. Para validar modelos de aprendizado de máquinas utiliza-se dos conceitos:

- **Verdadeiro Positivo (true positive – TP)**: Trata-se dos casos em que o modelo prevê corretamente a classe positiva. A previsão do modelo é positiva e o valor real também é positivo. Ou seja, ele confirma a realidade do dado;
- **Verdadeiro Negativo (true negative – TN)**: Refere-se aos casos em que o modelo prevê corretamente a classe negativa durante a análise. Neste caso, tanto a previsão do modelo quanto a realidade do fato são negativas;

- **Falso Positivo (*false positive* – FP)**: Indica quando o modelo prevê erroneamente uma classe positiva para um caso que, na realidade, é negativo. Sendo também chamado do erro Tipo I ou alarme falso;

- **Falso Negativo (*false negative* – FN)**: Ocorre quando o modelo prevê erroneamente uma classe negativa, para um caso que, na realidade, é positivo, também conhecido como erro do Tipo II.

SICSÚ, SAMARTINI e BARTH (2025) apresentam a matriz de confusão, ou classificação, onde cada um dos eixos apresenta os valores de previsão e os observados na realidade. A tabela 1 exemplifica um recurso inicial das avaliações de métricas.

Tabela 1 – Matriz de Confusão

MATRIZ DE CONFUSÃO		REALIDADE	
		POSITIVO	NEGATIVO
PREVISÃO	POSITIVO	TP	FP
	NEGATIVO	FN	TN

Fonte: dos autores

A partir da obtenção dos valores da matriz de confusão, torna-se possível calcular as principais métricas utilizadas em estudos iniciais de aprendizado de máquinas.

4.4.1. Acurácia

A acurácia (*accuracy, ACC*) está relacionada a proporção de previsões corretas em relação ao total de previsões realizadas. Contudo, em modelos que estão desbalanceados, ou seja, modelos em que as classes não estão igualmente representadas, essa métrica pode comprometer a análise. Provost e Fawcett (2013) informam que à medida que a distribuição

das classes se torna mais desbalanceada, a avaliação baseada em acurácia deixa de ser confiável.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

4.4.2. Precisão

A precisão (*precision*), também é conhecida como valor preditivo positivo, ou seja, foca nas instâncias corretamente previstas, indicando a proporção de verdadeiros positivos entre todas as previsões positivas feitas pelo modelo. É especialmente útil quando o custo de falsos positivos é alto, por exemplo em diagnósticos médicos. Tal métrica indica a capacidade do modelo de evitar previsões falsas positivas.

$$precision = \frac{TP}{TP + FP} \times 100\%$$

4.4.3. Revogação

Geralmente conhecida pela nomenclatura em inglês (*Recall*), a Revogação apresenta a sensibilidade ou taxa de verdadeiros positivos, medindo a proporção de instâncias corretamente prevista como positivas em relação ao total de instâncias realmente positivas. Também uma métrica aplicada nos cenários onde o custo de falsos negativos torna-se alto. Por exemplo, monitoramento de atividades criminais.

$$Recall = \frac{TP}{TP + FN} \times 100\%$$

4.4.4. F1-Score

O F1-Score é a média harmônica entre precisão (*precision*) e revogação (*recall*), proporcionando um equilíbrio entre as duas métricas. Seu valor varia entre 0 e 1, sendo utilizado para avaliar o desempenho geral do modelo, pois considera os FP e FN.

$$F1 - Score = \frac{precision \times Recall}{precision + Recall} \times 100\%$$

As métricas mencionadas anteriormente são comumente utilizadas em estudos de classificação por meio de Aprendizado de Máquina, assim como foi apresentado por SANTOS JUNIOR (2024). No campo da avaliação de modelos, há outras métricas possíveis,

como a Área Sob a Curva ROC (AUC-ROC), Log-Loss (Perda Logarítmica), Kolmogorov-Smirnov (KS) e outros.

4.5. Técnicas de visualização

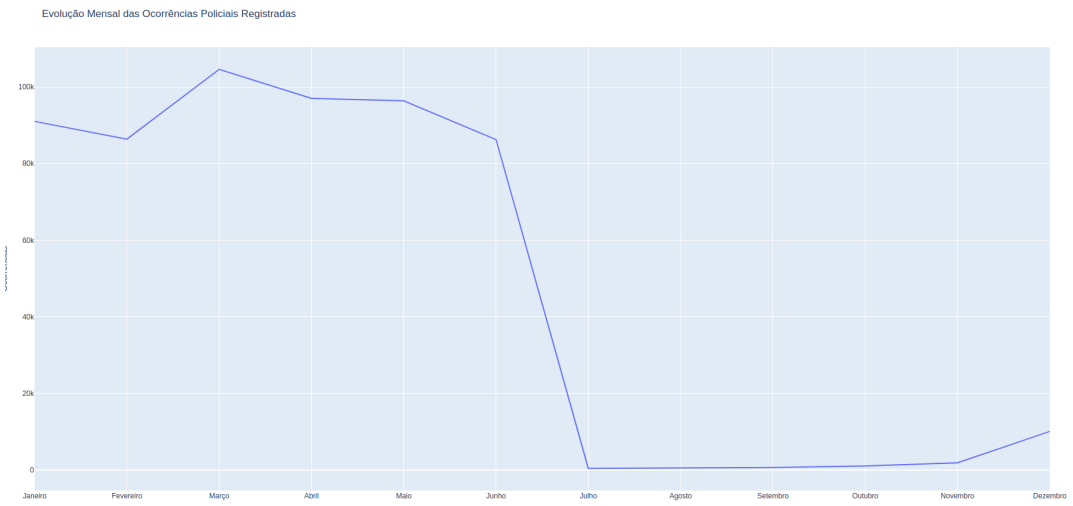
A visualização de dados constitui uma ferramenta essencial para a análise, interpretação e comunicação de informações em projetos de CD. No contexto deste trabalho, que aborda a análise de registros de ocorrências criminais, a utilização de técnicas visuais adequadas é fundamental para evidenciar padrões espaciais, temporais e categóricos presentes no conjunto de dados.

As técnicas de visualização permitem sintetizar grandes volumes de dados em formatos intuitivos, favorecendo tanto a compreensão rápida por parte de gestores públicos quanto a elaboração de *insights* para formulação de políticas mais eficazes. A escolha criteriosa do tipo de gráfico ou representação visual impacta diretamente a qualidade da interpretação e, consequentemente, a tomada de decisão.

Dentre as técnicas empregadas neste trabalho, destacam-se:

- **Gráficos de Linhas**: Utilizados para representar a evolução temporal de ocorrências criminais, possibilitando a identificação de tendências de aumento ou redução de delitos ao longo dos meses ou anos.

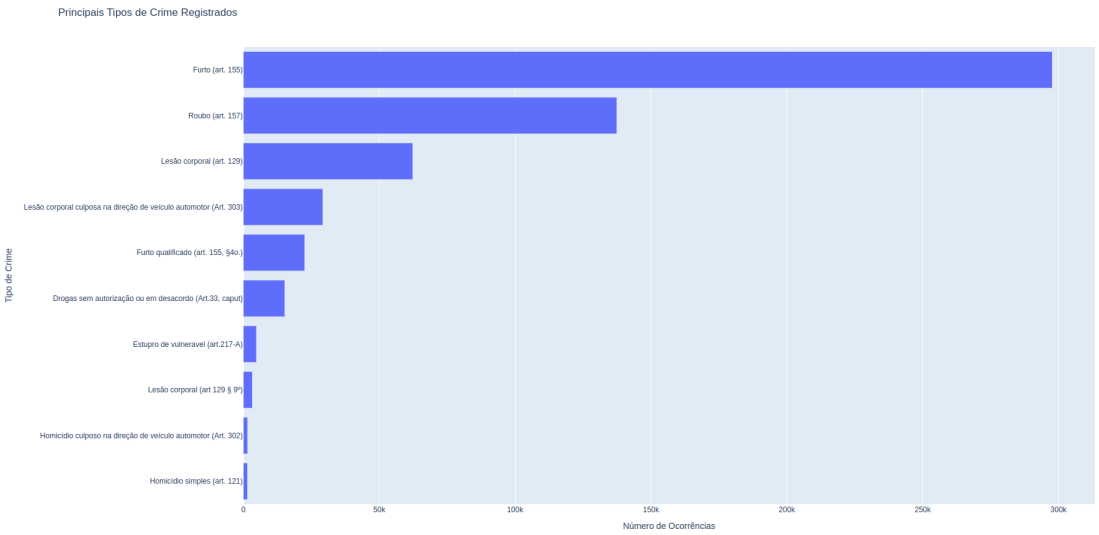
Figura 4: Gráfico de Linhas



Fonte: dos Autores

- **Gráficos de Barras e Colunas**: Aplicados para comparar diferentes categorias de crimes entre si ou entre localidades, permitindo análises comparativas claras

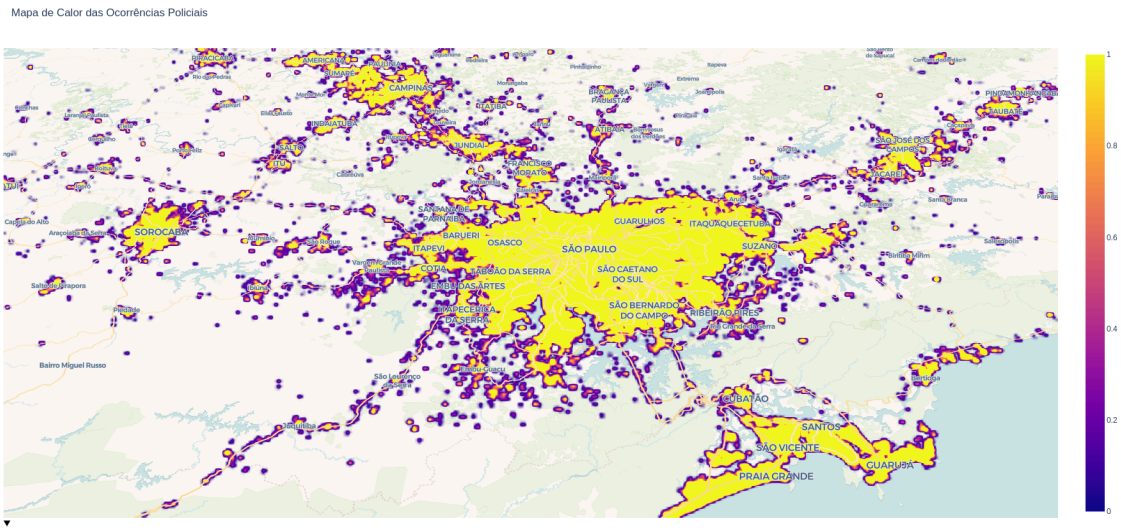
Figura 5: Gráfico de Barras e Colunas



Fonte: dos Autores

- **Mapas de Calor (*Heatmaps*)**: Desenvolvidos para visualizar a distribuição espacial dos registros criminais, evidenciando regiões com maior concentração de incidentes, o que auxilia no direcionamento de políticas de segurança pública.

Figura 6: Mapa de Calor



Fonte: dos Autores

- **Gráficos de Pizza ou Donut**: Utilizados com cautela para ilustrar a proporção de tipos de crimes dentro de um conjunto total de registros, em casos em que a comparação de partes em relação ao todo é relevante.

- **Dashboards Interativos**: Construídos para integrar múltiplas visualizações em um ambiente navegável, permitindo a análise dinâmica por filtros como tipo de crime, período e localização geográfica.

A adoção dessas técnicas, sempre alinhada aos princípios de clareza, precisão e objetividade, potencializa a comunicação dos resultados para públicos diversos, desde analistas técnicos até gestores políticos.

Além disso, a visualização desempenha um papel estratégico neste projeto ao conectar a análise estatística com a percepção espacial e temporal dos dados, tornando os resultados mais acessíveis e impactantes.

5. MATERIAIS E MÉTODOS

5.1. Fonte e coleta de dados

Os dados utilizados neste trabalho foram obtidos no portal da SSP-SP.[1] Há uma variedade de dados disponíveis nesse portal, quais sejam: dados referentes a crimes envolvendo celulares, veículos e outros objetos; dados de mortes decorrentes de intervenção policial; dados de produtividade da polícia; dados gerais de crimes ocorridos no estado de São Paulo. Todos estes conjuntos de dados estão granularizados, quanto ao seu período, por ano de ocorrência.

O objetivo estabelecido inicialmente para este trabalho foi a construção de um classificador capaz de prever crimes, dada uma localização. Sendo assim, defronte às possibilidades oferecidas pelos dados disponíveis no portal da SSP-SP, escolheu-se os dados gerais de crimes ocorridos, pois estes dariam uma gama maior de possibilidades ao classificador a ser construído. Quanto ao ano de ocorrências, decidiu-se que o conjunto de dados referentes ao ano de 2023 seria suficiente para a construção dos classificadores no contexto deste trabalho, já que, de antemão, era possível prever que a utilização de dados em escala muito grande poderia trazer dificuldade no processamento para a criação dos modelos. Sabe-se que ao se agregar mais dados ao treinamento de classificadores, geralmente há melhoria na acurácia destes. No entanto, como o poder computacional disponível aos membros deste trabalho é modesto, acordou-se que seria necessário o uso do menor conjunto de dados disponível, o qual oferecesse um balanceamento: uma boa perspectiva de acurácia do modelo construído com mais baixo uso de memória primária possível.

5.2. Ferramentas de análise e elaboração de modelos

As ferramentas utilizadas neste trabalho foram a suíte Anaconda e as bibliotecas da linguagem de programação Python denominadas NumPy, Pandas, Scikitlearn e Matplotlib. Da suíte de ferramentas Anaconda, utilizou-se especialmente o JupyterLab[2], uma interface que permite a

[1] Dados obtidos no portal da SSP-SP, no link https://www.ssp.sp.gov.br/estatistica/consultas. Acesso em 28 de Abril de 2025.
[2] O JupyterLab pode ser baixado no endereço https://jupyter.org/

edição em *notebooks* do código do projeto. Editar o código em blocos é uma grande vantagem nessa seara, pois, diferentemente de outros projetos de software, a mineração de dados pressupõe de um avanço progressivo na construção da solução, parecendo-se bastante com a resolução de problemas matemáticos. Além disso, é bastante intuitivo ler o código neste ambiente, já que anotações, ou comentários, são parte integral dos procedimentos realizados sobre os dados, ao contrário de programas escritos para outras finalidades.

Sobre as bibliotecas de Python citadas, Pandas[3] é uma biblioteca que permite a análise inicial dos dados, assim como o entendimento de estatísticas descritivas básicas do conjunto analisado. Ainda, permite a edição do arquivo perscrutado de acordo à necessidade do classificador a ser construído. Por exemplo, toda a filtragem e eliminação de dados não necessários ou que foram por motivos específicos desconsiderados, foi realizada mediante o uso dessa biblioteca. Já o Scikitlearn é um pacote que permite a aplicação de uma grande diversidade de algoritmos de aprendizado de máquina, sem a necessidade de codificação completa destes. Isso dá uma rapidez enorme aos projetos de CD, e permite o teste de vários algoritmos sobre determinado dataset, além da comparação das suas métricas de desempenho, algo que será realizado no presente trabalho. Finalmente, a biblioteca Matplotlib foi usada para obter gráficos e visuais referentes aos classificadores construídos. Por fim, usou-se a biblioteca NumPy, a qual permite a realização de operações de Álgebra Linear de maneira intuitiva sobre os dados trabalhados. As quatro bibliotecas citadas, assim como o JupyterLab, são padrões largamente empregados nas mais variadas aplicações de CD, desde aplicações da indústria de software até pesquisas avançadas em Computação. Além disso, tem uso gratuito e tem código fonte aberto à comunidade.

Trecho de código-fonte 1 – Importação de Bibliotecas e Pacotes

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import CategoricalNB
from sklearn import metrics
```

[3] Os softwares citados neste apartado podem ser encontrados, respectivamente, nos seguintes endereços: https://pandas.pydata.org/, https://scikit-learn.org/stable/, https://matplotlib.org/ .

Fonte: dos autores

5.3. Preparação dos dados

A preparação dos dados foi iniciada a partir do carregamento dos dados no arquivo com extensão ipybn criado, nomeado “TCC”.

Trecho de código-fonte 2 – Carregamento de dados

```
df = pd.read_csv("SPDadosCriminais_2023.csv")
```

Fonte: dos autores

A partir desse momento foi possível notar as características básicas dos dados disponibilizados pela SSP-SP.

Trecho de código-fonte 3 – Características iniciais

```
#   Column                       Non-Null Count   Dtype
---  ------                      --------------   -----
 0   NOME_DEPARTAMENTO           618395 non-null  object
 1   NOME_SECCIONAL              618395 non-null  object
 2   NOME_DELEGACIA              618395 non-null  object
 3   NOME_MUNICIPIO              618395 non-null  object
 4   NUM_BO                      618397 non-null  object
 5   ANO_BO                      618397 non-null  int64
 6   DATA_REGISTRO               618396 non-null  object
 7   DATA_OCORRENCIA_BO          616258 non-null  object
 8   HORA_OCORRENCIA_BO          432730 non-null  object
 9   DESC_PERIODO                185667 non-null  object
 10  DESCR_SUBTIPOLOCAL          618397 non-null  object
 11  BAIRRO                      608728 non-null  object
 12  LOGRADOURO                  618397 non-null  object
 13  NUMERO_LOGRADOURO           516929 non-null  object
 14  LATITUDE                    556799 non-null  float64
 15  LONGITUDE                   556799 non-null  float64
 16  NOME_DELEGACIA_CIRCUNSCRIÇÃO    618397 non-null  object
 17  NOME_DEPARTAMENTO_CIRCUNSCRIÇÃO 618397 non-null  object
 18  NOME_SECCIONAL_CIRCUNSCRIÇÃO    618397 non-null  object
 19  NOME_MUNICIPIO_CIRCUNSCRIÇÃO    618397 non-null  object
 20  RUBRICA                     618383 non-null  object
 21  DESCR_CONDUTA               469999 non-null  object
 22  NATUREZA_APURADA            618397 non-null  object
 23  MES_ESTATISTICA             618397 non-null  int64
 24  ANO_ESTATISTICA             618397 non-null  int64
dtypes: float64(2), int64(3), object(20)
memory usage: 117.9+ MB
```

Fonte: dos autores

O conjunto de dados original possuía 618.397 exemplos e 25 atributos, características retornadas através do atributo da classe DataFrame denominada shape. No trecho de código-fonte 3, temos o retorno do método info. Percebe-se que a maioria das colunas do conjunto de dados foi denominada como object, ou seja, o pacote Pandas não conseguiu identificar um tipo de dados específico para essas colunas. Apenas as colunas “ANO_BO”, “MES_ESTATISTICA”, “ANO_ESTATISTICA”, “LATITUDE” e “LONGITUDE” foram devidamente reconhecidas como valores inteiros e decimais, sendo as duas últimas pertencentes a este tipo de dados. Outra informação que se pode obter mediante análise do exposto no trecho de código-fonte 3 é a ausência de dados em muitas colunas. Em caráter de exemplo, percebe-se que grande parte das colunas tem 618.397 valores não-nulos e algumas colunas possuem um valor inferior a esse. Essa diferença impactou a decisão sobre a escolha de atributos a serem usados como recursos (*features*) para o classificador.

No entanto, mesmo notando essas discrepâncias nos dados logo em primeira análise, decidiu-se discutir sobre os atributos que pudessem ser mais relevantes ao classificador em questão, qual seja, aquele que pudesse retornar um rótulo para um crime a partir de uma localidade e tempo de ocorrência. Sendo assim, descartaram-se primeiramente os atributos relativos ao local de registro do boletim de ocorrência, sendo eles as colunas “NOME_DEPARTAMENTO”, “NOME_SECCIONAL”, “NOME_DELEGACIA”. Igualmente, notou-se que o atributo “NUM_BO” se referia ao número do boletim de ocorrência em questão. Por ser um atributo único, essa coluna também foi descartada, pois serviria apenas para gerar ruído na classificação. Seguidamente, chamou a atenção o conteúdo das colunas “NOME_DELEGACIA_CIRCUNSCRIÇÃO”, “NOME_DEPARTAMENTO_CIRCUNSCRIÇÃO”, “NOME_SECCIONAL_CIRCUNSCRIÇÃO”, “NOME_MUNICIPIO_CIRCUNSCRIÇÃO”, que, na grande maioria dos casos, apenas repetiam valores de outros atributos, como “NOME_MUNICIPIO”. Dessa forma, excluíram-se essas colunas. Por fim, nessa primeira análise, as colunas “ANO_ESTATISTICA” e “MES_ESTATISTICA” foram tidas como inúteis a construção do modelo, pois se repetiam (por se tratar de um conjunto de dados referentes apenas a 2023) ou poderiam ser obtidas pelo desmembramento de outras colunas (como “DATA_OCORRENCIA_BO”).

Abaixo, o comando utilizado para a realização da exclusão de colunas.

Trecho de código-fonte 4 – Exclusão de colunas

```
df.drop(['NOME_DEPARTAMENTO', 'NOME_SECCIONAL', 'NOME_DELEGACIA',
'NUM_BO', 'ANO_BO', 'DESC_PERIODO', 'DESCR_SUBTIPOLOCAL',
'NOME_DELEGACIA_CIRCUNSCRIÇÃO', 'NOME_DEPARTAMENTO_CIRCUNSCRIÇÃO',
'NOME_SECCIONAL_CIRCUNSCRIÇÃO', 'DESCR_CONDUTA', 'MES_ESTATISTICA',
'ANO_ESTATISTICA'], axis=1, inplace=True)
```

Fonte: dos autores

Após as exclusões houve a redução de metade dos atributos, ou seja, de 24 para 12, e o conjunto de dados passou de um tamanho de aproximadamente 118 MB para 56 MB. Essa redução foi fundamental para o processamento do modelo, dadas as limitações de hardware impostas à realização deste trabalho.

Trecho de código-fonte 5 – Atributos restantes pós-exclusão de colunas

```
#   Column                    Non-Null Count   Dtype
---  ------                   --------------   -----
 0   NOME_MUNICIPIO           618395 non-null  object
 1   DATA_REGISTRO            618396 non-null  object
 2   DATA_OCORRENCIA_BO       616258 non-null  object
 3   HORA_OCORRENCIA_BO       432730 non-null  object
 4   BAIRRO                   608728 non-null  object
 5   LOGRADOURO               618397 non-null  object
 6   NUMERO_LOGRADOURO        516929 non-null  object
 7   LATITUDE                 556799 non-null  float64
 8   LONGITUDE                556799 non-null  float64
 9   NOME_MUNICIPIO_CIRCUNSCRIÇÃO  618397 non-null  object
 10  RUBRICA                  618383 non-null  object
 11  NATUREZA_APURADA         618397 non-null  object
dtypes: float64(2), object(10)
memory usage: 56.6+ MB
```

Fonte: dos autores

Em seguida, percebeu-se que não havia registro de localidade de muitos exemplos constantes no conjunto de dados. Os valores eram preenchidos com a cadeia "VEDAÇÃO DA DIVULGAÇÃO DOS DADOS RELATIVOS"[4]. Como o classificador a ser treinado deveria

[4] Segundo o que foi possível apurar, as restrições a divulgações de dados relativos a crimes se devem principalmente às Leis 11.340/2006 e 12.527/2011, conhecidas popularmente como Lei Maria da Penha e Lei de

partir de uma localização, decidiu-se que os exemplos onde esse valor constava deveriam ser retirados. Assim sendo, utilizou-se o seguinte comando.

Trecho de código-fonte 6 – Filtragem de valores

```
df = df[df["LOGRADOURO"] != "VEDAÇÃO DA DIVULGAÇÃO DOS DADOS RELATIVOS"]
```

Fonte: dos autores

Além disso, e ainda na seara de tratamento e limpeza de dados, foi possível encontrar uma diversidade de rubricas no conjunto de dados analisados, as quais foram tomadas como as etiquetas de classificação pretendidas para o classificador em questão.

Trecho de código-fonte 7 – Tipos de Delitos recuperados da coluna "RUBRICA"

```
'Furto (art. 155)',
    'Homicídio culposo na direção de veículo automotor (Art. 302)',
    'Homicídio (art. 121)', 'Lesão corporal (art. 129)',
    'Lesão corporal culposa na direção de veículo automotor (Art. 303)',
    'Porte ilegal de arma de fogo de uso permitido (Art. 14)',
    'Posse ou porte ilegal de arma de fogo de uso restrito (Art. 16)',
    'Drogas para consumo pessoal sem autorização ou em desacordo (Art.28,caput)',
    'Roubo (art. 157)',
    'Drogas sem autorização ou em desacordo (Art.33, caput)',
    'Localização/Apreensão de objeto', 'Outros não criminal',
    'Associarem-se duas ou mais pessoas - arts. 33, caput e § 1o, e 34 (Art.35,caput)',
    'Comunicação de óbito', 'Entrega de objeto localizado/apreendido',
    'Porte de arma (art. 19)', 'Atropelamento', 'Morte suspeita',
    'Capotamento', 'Desobediência (art. 330)',
    'Adulteração de sinal identificador de veículo automotor (art. 311)',
    'Furto de coisa comum (art. 156)',
    'Extorsão mediante seqüestro (art. 159)', 'Colisão', 'Choque',
    'Engavetamento', 'Favorecimento pessoal (art. 348)',
    'Ameaça (art. 147)', 'Fuga de local de acidente (Art. 305)', nan,
    'Embriaguez ao volante (Art. 306)',
    'Dirigir sem Permissão ou Habilitação (Art. 309)',
    'Localização/Apreensão e Entrega de objeto',
    'Entrega de veículo localizado/apreendido',
    'Receptação (art. 180)',
    'Cumprimento de mandado de prisão temporária',
    'Localização/Apreensão de veículo',
    'Destruir ou danificar vegetação primária ou secundária(Art.38-A)',
    'Contrabando (Art 334A )',
    'Caput Corromper ou facilitar a corrupção de menor de 18 anos (244B)',
    'Auto lesão', 'Localização/Apreensão e Entrega de veículo',
```

Acesso à Informação. A Lei Geral de Proteção de Dados, discutida anteriormente, também pode ter influência sobre essa divulgação.

```
'Qq. objeto destinado a fabr., prep., prod. ou transformação de drogas (Art.34)',
'Direção perigosa de veículo na via pública (art. 34)',
'Captura de procurado',
'Cumprimento de mandado de busca e apreensão',
'Entrada ilegal de aparelho movél de comunicação em estabelecimento prisional',
'Resistência (art. 329)', 'Dano (art. 163)', 'Abalroamento',
'Suicídio consumado', 'Associação Criminosa (art. 288)',
'Posse irregular de arma de fogo de uso permitido (Art.12)',
'homicídio culposo',
'Trafegar em velocidade incompatível (Art. 311)',
'Disparo de arma de fogo (Art. 15)',
'Destruição, subtração ou ocultação de cadáver (art. 211)',
'Apreensão de Adolescente', 'Art. 213 - Estupro',
'Desacato (art. 331)',
'HomicIdio culposo na direção de veIculo automotor (Art. 302)',
'Jogo de azar (art. 50)'
```

Fonte: dos autores

Nota-se que muitos desses valores não têm uma classificação enquadrada no Código Penal Brasileiro adicionada na cadeia de valores que descreve o crime. Vide, por exemplo, a cadeia "Captura de procurado", onde não há referência a um artigo do legislação brasileira referente a crimes. Já que o classificador a ser treinado pertence a um projeto de pretensões modestas, que visa a máxima correção das classificações em detrimento da abrangência dos objetos a serem classificados, e, além disso, para a correta classificação de um crime é, por excelência, indispensável a participação de um profissional do Direito, indisponível no escopo deste trabalho, decidiu-se excluir crimes que não dispõem de enquadramento na sua descrição. Para tanto, utilizou-se o comando do trecho de código-fonte 8.

Trecho de código-fonte 8 – Cadeias excluídas da coluna "RUBRICA"

```
excluir = ['Localização/Apreensão de objeto', 'Outros não criminal', 'Comunicação de óbito',
'Entrega de objeto localizado/apreendido', 'Atropelamento', 'Morte suspeita',
'Capotamento','Colisão', 'Choque','Engavetamento','Localização/Apreensão e Entrega de
objeto','Entrega de veículo localizado/apreendido', 'Cumprimento de mandado de prisão
temporária',
    'Localização/Apreensão de veículo','Auto lesão', 'Localização/Apreensão e Entrega de
veículo','Cumprimento de mandado de busca e apreensão', 'Entrada ilegal de aparelho movél de
comunicação em estabelecimento prisional', 'Abalroamento', 'Suicídio consumado','homicídio
culposo','Apreensão de Adolescente']
df = df[~df["RUBRICA"].isin(excluir)]
```

Fonte: dos autores

Os crimes corretamente enquadrados segundo artigos do Código Penal Brasileiro resultantes da filtragem foram:

Trecho de código-fonte 9 – Tipos de delitos pós-filtragem

```
'Furto (art. 155)',
    'Homicídio culposo na direção de veículo automotor (Art. 302)',
    'Homicídio (art. 121)', 'Lesão corporal (art. 129)',
    'Lesão corporal culposa na direção de veículo automotor (Art. 303)',
    'Porte ilegal de arma de fogo de uso permitido (Art. 14)',
    'Posse ou porte ilegal de arma de fogo de uso restrito (Art. 16)',
    'Drogas para consumo pessoal sem autorização ou em desacordo (Art.28,caput)',
    'Roubo (art. 157)',
    'Drogas sem autorização ou em desacordo (Art.33, caput)',
    'Associarem-se duas ou mais pessoas - arts. 33, caput e § 1o, e 34 (Art.35,caput)',
    'Porte de arma (art. 19)', 'Desobediência (art. 330)',
    'Adulteração de sinal identificador de veículo automotor (art. 311)',
    'Furto de coisa comum (art. 156)',
    'Extorsão mediante seqüestro (art. 159)',
    'Favorecimento pessoal (art. 348)', 'Ameaça (art. 147)',
    'Fuga de local de acidente (Art. 305)',
    'Embriaguez ao volante (Art. 306)', nan,
    'Dirigir sem Permissão ou Habilitação (Art. 309)',
    'Receptação (art. 180)',
    'Destruir ou danificar vegetação primária ou secundária(Art.38-A)',
    'Contrabando (Art 334A )',
    'Caput Corromper ou facilitar a corrupção de menor de 18 anos (244B)',
    'Qq. objeto destinado a fabr., prep., prod. ou transformação de drogas (Art.34)',
    'Direção perigosa de veículo na via pública (art. 34)',
    'Captura de procurado', 'Resistência (art. 329)',
    'Dano (art. 163)', 'Associação Criminosa (art. 288)',
    'Posse irregular de arma de fogo de uso permitido (Art.12)',
    'Trafegar em velocidade incompatível (Art. 311)',
    'Disparo de arma de fogo (Art. 15)',
    'Destruição, subtração ou ocultação de cadáver (art. 211)',
    'Art. 213 - Estupro', 'Desacato (art. 331)',
    'HomicIdio culposo na direção de veIculo automotor (Art. 302)',
    'Jogo de azar (art. 50)'
```

Fonte: dos autores

Depois de realizada essa etapa, verificou-se que, mesmo com a exclusão de muitas colunas e valores, ainda havia muitos valores nulos em várias colunas. A decisão de desconsiderar os valores nulos foi natural, pois o dataset continha muito mais exemplos do que o necessário para o treinamento de um classificador do porte necessário ao escopo deste trabalho. Dessa maneira, usou-se o seguinte comando:

Trecho de código-fonte 10 – Exclusão de valores nulos

```
df.dropna(inplace=True)
```

Fonte: dos autores

Observou-se que, para a construção de um modelo eficaz, uma informação não diretamente constante dos dados poderia contribuir, qual seja a diferença entre a data de registro e a data de ocorrência do crime. Ademais, foram transformadas as datas constantes no conjunto de dados em atributos numéricos. Todos os procedimentos acima descritos foram realizados da maneira que se descreve no próximo bloco de códigos.

Trecho de código-fonte 11 – Extração de atributos numéricos de atributos de período

```
df["DATA_REGISTRO"] = pd.to_datetime(df["DATA_REGISTRO"])
df["DATA_OCORRENCIA_BO"] = pd.to_datetime(df["DATA_OCORRENCIA_BO"])
df["DIAS_REGISTRO"] = (df["DATA_REGISTRO"] -
df["DATA_OCORRENCIA_BO"]).astype(str)
df["DIAS_REGISTRO"] = df["DIAS_REGISTRO"].str[0]
df["MES_REGISTRO"] = df["DATA_REGISTRO"].dt.month df["DIA_REGISTRO"] =
df["DATA_REGISTRO"].dt.day
df["MES_OCORRENCIA"] = df["DATA_OCORRENCIA_BO"].dt.month
df["DIA_OCORRENCIA"] = df["DATA_OCORRENCIA_BO"].dt.day
```

Fonte: dos autores

Destaca-se nesse trecho de código a facilidade oferecida pela biblioteca Pandas, a qual permitiu a extração de atributos numéricos de mês e dia a partir de datas com apenas uma linha de código. Pensando em outras abordagens de programação, a recuperação dessa informação poderia ser bem mais complicada.

O conjunto de dados resultante de todos esses processos de limpeza e extração de dados continha 327.218 registros e 14 colunas, as quais foram reorganizadas com o seguinte comando:

Trecho de código-fonte 12 – Reorganização dos dados

```
df = df[['NOME_MUNICIPIO', 'DIAS_REGISTRO', 'MES_REGISTRO', 'DIA_REGISTRO',
'MES_OCORRENCIA', 'DIA_OCORRENCIA', "HORA_OCORRENCIA_BO", 'BAIRRO',
```

```
'LOGRADOURO', 'NUMERO_LOGRADOURO', 'LATITUDE', 'LONGITUDE',
'NOME_MUNICIPIO_CIRCUNSCRIÇÃO', 'NATUREZA_APURADA', 'RUBRICA']]
```

Fonte: dos autores

A reorganização se deve a uma tradição dentro da Ciência de Dados, que preza pela manutenção da etiqueta ou atributo alvo como último atributo dentro do conjunto de dados trabalhado. Por fim, antes da aplicação de *encoders*[5] a atributos não numéricos do *dataset*, realizaram-se algumas conversões e ajustes de dados numéricos.

Trecho de código-fonte 13 – Transformação de tipo de dados e modularização

```
df["NUMERO_LOGRADOURO"] = df["NUMERO_LOGRADOURO"].astype(np.int64)
df["LATITUDE"] = df["LATITUDE"].abs()
df["LONGITUDE"] = df["LONGITUDE"].abs()
```

Fonte: dos autores

Os números de logradouro foram convertidos para int64, devido a sua extensão em bits. Outrossim, a latitude e longitudes presente no conjunto de dados tiveram que ser modularizadas, já que os algoritmos utilizados no trabalho não admitiam a introdução de valores negativos[6]. Os valores modularizados correspondem a outras localidades geográficas. No entanto, como os dados no contexto deste trabalho são sempre provenientes da SSP-SP, não se viu problema em utilizar tal procedimento.

Posteriormente, foi necessário aplicar codificações aos atributos não numéricos do conjunto de dados, quais sejam “NOME_MUNICIPIO”, “BAIRRO”, “LOGRADOURO” e “NATUREZA_APURADA”.

[5] *Enconders* são mapeadores de atributos que, geralmente, atribuem unicamente um número natural a uma categoria representada por uma cadeia. A título de exemplo, ao crime Embriaguez ao volante (Art. 306), pode ter sido atribuído o valor numérico “1”, e ao crime Roubo (art. 157), o valor “2”. Assim, o algoritmo atribuiria um valor numérico a todos os crimes considerados no escopo do conjunto de dados em questão, de acordo a distribuição estatística dos dados.

[6] Considerando a geografia do estado de São Paulo e o sistema de coordenadas geodésicas utilizado no conjunto de dados, os valores estariam entre –19,29 e –25,16 (Norte-Sul) e –44,18 e –53,13 (Leste-Oeste), ou seja, nas coordenadas correspondentes ao estado de São Paulo

Trecho de código-fonte 14 – Aplicação de codificação em atributos classificadores

```
labelenconder = LabelEncoder()

df["NOME_MUNICIPIO"] = labelenconder.fit_transform(df["NOME_MUNICIPIO"])
df["BAIRRO"] = labelenconder.fit_transform(df["BAIRRO"]) df["LOGRADOURO"] =
labelenconder.fit_transform(df["LOGRADOURO"]) df["NATUREZA_APURADA"] =
labelenconder.fit_transform(df["NATUREZA_APURADA"])
```

Fonte: dos autores

A função *LabelEncoder*, do pacote preprocessing, facilitou enormemente o trabalho, já que permite a realização da tarefa de atribuir um valor numérico a uma cadeia sem a necessidade da escrita de uma função. Entretanto, após os primeiros experimentos com todas as colunas resultantes de todo o pré-processamento realizado, foi necessário abandonar as colunas "BAIRRO", "LOGRADOURO" e "NATUREZA_APURADA", pois essas trouxeram uma quantidade de valores muito alta ao conjunto de dados. Isso gerou estouro da quantidade de memória RAM disponível ao treinamento do modelo e fez com que reconsiderações sobre os atributos fossem necessárias. Devido a repetição causada pelos atributos anteriores citados (a informação da localidade do crime já estava imbuída nas colunas "LATITUDE" e "LONGITUDE"), decidiu-se excluir as três colunas mencionadas acima.

Esse procedimento trouxe um grande alívio na quantidade de informação fornecida ao modelo, mas ainda não possibilitou a aplicação dos modelos de aprendizado de máquina escolhidos. Dessa forma, foi necessária mais uma filtragem, desta vez realizada na coluna de atributos alvo "RUBRICA".

Trecho de código-fonte 15 – Filtragem de delitos pouco representativos

```
df= df[df["RUBRICA"].isin(df["RUBRICA"].value_counts()[df["RUBRICA"].value_counts() >
200].index)]
```

Fonte: dos autores

Este comando filtra os crimes que tiveram mais de 200 registros no conjunto de dados trabalhado, possibilitando um efeito duplamente benéfico ao modelo: por um lado, permitiu o carregamento dos dados em memória principal e, por outro, permitiu a remoção de ruídos

estatísticos nos resultados, conhecidos como *outliers*. A quantidade de ocorrências para o rol final de crimes considerados pode ser visto no trecho de código-fonte 16.

Trecho de código-fonte 16 – Exemplos para cada tipo de delito

```
Furto (art. 155)                                                    137108
Roubo (art. 157)                                                    106804
Lesão corporal culposa na direção de veículo automotor (Art. 303)    29101
Lesão corporal (art. 129)                                           28187
Drogas sem autorização ou em desacordo (Art.33, caput)              13639
Drogas para consumo pessoal sem autorização ou em desacordo (Art.28,caput)   6000
Homicídio (art. 121)                                                1962
Associarem-se duas ou mais pessoas - arts. 33, caput e § 1o, e 34 (Art.35,caput)   1115
Homicídio culposo na direção de veículo automotor (Art. 302)        1035
Porte ilegal de arma de fogo de uso permitido (Art. 14)             854
Posse ou porte ilegal de arma de fogo de uso restrito (Art. 16)     645
Porte de arma (art. 19)                                             291
HomicIdio culposo na direção de veIculo automotor (Art. 302)        240
Name: RUBRICA, dtype: int64
```

Fonte: dos autores

5.4. Aplicação dos modelos

A aplicação dos modelos começou com a definição de quais colunas fariam parte dos atributos seriam os rótulos e aquele que faria o papel do atributo alvo (*label*).

Trecho de código-fonte 17 – Escolha dos atributos classificadores e alvo

```
X = df.iloc[:, : -2]
y = df.loc[:, "CLASSIFICACAO"]
```

Fonte: dos autores

Logo em seguida, usou-se a função *train_test_split*, do pacote *model_prediction*, para a divisão das bateladas de treino e teste. Optou-se, heuristicamente, por uma divisão 80/20, recomendação dada em várias fontes para treinamento de modelos supervisionados (GRUS, 2016; GOODFELLOW, 2016).

Trecho de código-fonte 18 – Divisão das bateladas de treinamento e teste

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, train_size=0.8,
random_state=1)
```

Fonte: dos autores

Por fim, e depois de muitos experimentos realizados, notou-se que o tamanho máximo de arquivo passível de ser processado na máquina disponível era de aproximadamente 50 MB, o que corresponde a 300.000 linhas.

O último passo foi o treinamento dos modelos através dos comandos seguintes:

Trecho de código-fonte 19 – Instanciação de modelo e treinamento - Naive Bayes

```
cnb = CategoricalNB()
cnb.fit(X_train, y_train)
```

Fonte: dos autores

Trecho de código-fonte 20 – Instanciação de modelo e treinamento - Random Forest

```
rf = RandomForestClassifier(n_estimators=100, random_state=42)
rf.fit(X_train,y_train)
```

Fonte: dos autores

5.5. Validação dos modelos

Os modelos treinados tiveram desempenho semelhante àqueles descritos por Ribeiro et al (2020), em trabalho que cria um classificados para crimes usando dados da Secretaria de Segurança Pública do Maranhão. A medição da adequação dos modelos aos dados se deu mediante a obtenção da acurácia dos dois modelos treinados, como no trabalho anteriormente citado. Essas métricas foram obtidas através do uso das seguintes função do módulo *metrics,* da biblioteca *sklearn*.

Trecho de código-fonte 21 – Obtenção de acurácia - Naive Bayes

```
cnb.score(X_train, y_train)

cnb.score(X_test, y_test)
```

Fonte: dos autores

Trecho de código-fonte 22 – Obtenção de acurácia - Random Forest

```
rf.score(X_train, y_train)

rf.score(X_test, y_test)
```

Fonte: dos autores

Os resultados obtidos estão sumarizados na tabela 2 e serão discutidos em seção posterior a presente.

Tabela 2 – Acurácia para os modelos em ambos conjuntos

	Naive Bayes	**Random Forest**
Conjunto de Treinamento	45,12%	99,33%
Conjunto de Teste	44,68%	49,90%

Fonte: dos autores

6. DISCUSSÃO DOS RESULTADOS

Como mencionado anteriormente, os modelos treinados obtiveram resultados semelhantes à literatura consultada (RIBEIRO, 2020). No entanto, é notável que, no conjunto de dados de treinamento, o modelo *Random Forest* tenha obtido uma acurácia alta. Esse geralmente é um sinal de sobreajuste, conforme a literatura especializada (GRUS, 2016; GOODFELLOW, 2016). Também, é possível que a diferença entre a acurácia de teste e treino para esse algoritmo se deve a distribuição dos dados ou a randomização das bateladas de teste. Os experimentos realizados neste trabalho não foram suficientes para dar conta de todas essas variáveis. Porém, é possível prever que as questões discutidas anteriormente poderiam diminuir essa diferença sensivelmente.

Outro fato notável a partir dos resultados obtidos foi a baixa acurácia do modelo *Naive Bayes* usado. Esse fato já era esperado, já que esse algoritmo é usado muitas vezes apenas de forma comparativa com outros algoritmos, o que normalmente se denomina como algoritmo *baseline*. A vantagem do *Naive Bayes* é claramente um rápido processamento[7]. No entanto, os resultados indicam claramente que, entre os dois candidatos, o algoritmo *Random Forest* deveria ser o escolhido, se a eficácia de classificação for a característica pretendida por alguma aplicação.

Por fim, percebe-se que os classificadores não são próprios para o uso em sistemas reais. No entanto, os resultados indicam que, mesmo com poucos exemplos e baixo poder de processamento, é possível criar modelos que conseguem ter desempenho moderado para a classificação de crimes. Tendo em vista que autoridades e instituições governamentais possuem grandes recursos financeiros, essas barreiras técnicas poderiam ser facilmente quebradas, obtendo-se classificadores com desempenho muito superior àquele obtido neste trabalho.

[7] Nossos experimentos indicaram uma diferença de cerca de 400% no tempo de processamento do *Random Forest* em relação ao *Naive Bayes.*

7. CONSIDERAÇÕES FINAIS

O presente trabalho examinou a aplicação de um ramo específico da Ciência da Computação, qual seja, a CD em um intricado tema da vida cotidiana, a segurança pública. Primeiramente, teceu-se um panorama jurídico-legal da área de segurança pública, no qual buscou-se evidenciar a legislação federal e estadual concernente a essa seara. Logo, buscou-se os fundamentos das técnicas de Computação que pudessem ser úteis na realização deste trabalho. Frisou-se as possibilidades do aprendizado de máquina na descoberta de padrões em dados, característica identificada como pertinente e útil a aplicações dentro da temática escolhida. Juntamente com a visualização de dados, também discutida em seção teórica do presente texto, procurou-se demonstrar a utilidade desse conjunto de ferramentas à elaboração de melhores estratégias de segurança pública, visando mais a prevenção do que o combate ou repressão ao crime. Em consulta à literatura especializada, observou-se que essas ideias são conhecidas como policiamento preditivo, cuja operabilidade é mostrada por bons resultados em vários exemplos já implementados, também discutidos neste texto.

A contribuição específica deste trabalho é considerar a aplicação de modelos de aprendizado de máquina já existentes a dados da SSP-SP, algo ainda não realizado na literatura científica. Apesar dos resultados serem acanhados, percebe-se que o uso de poucos recursos computacionais permite a construção de classificadores para crimes, algo em si complexo, dada as diversas variáveis humanas - e portanto dificilmente mensuráveis e recuperáveis - envolvidas nessa seara.

Ademais, vistos os exemplos de aplicação de policiamento preventivo em outros países e do próprio estado de São Paulo, nota-se que esse trabalho contribui para a disseminação da discussão desses tópicos mediante a realização de experimentos autônomos à SSP-SP e ao governo estadual. Os experimentos aqui conduzidos validaram a eficiência das técnicas adotadas, mesmo com as diversas barreiras impostas pelos materiais utilizados. Assim sendo, pode-se inferir que a política de segurança paulista segue um caminho interessante da aplicação de tecnologia da informação em favor dos cidadãos paulistas, merecendo, como em toda construção tecno-sócio-política aprimoramentos e maiores investigações de maneira intermitente.

8. TRABALHOS FUTUROS

Dadas as restrições impostas à realização deste trabalho, seria desejável contar com maior poder computacional para a realização de testes com outros classificadores, como redes neurais artificiais, por exemplo.

Além disso, o enriquecimento dos dados obtidos da SSP-SP mediante a adição de dados não estruturados, como gravações de vídeo, técnica conhecida na literatura como multi modalidade, poderia ser de grande valia no aumento da acurácia dos modelos construídos. Certamente, haveria aumento significativo da necessidade de processamento para a construção desses modelos, o que acarretaria em necessidades financeiras elevadas para a realização da investigação.

Por fim, valeria a pena investigações e promoção de debates junto a grupos interdisciplinares de pesquisa, já que a aplicação de algoritmos de IA a um tema tão sensível quanto à segurança pública requer um grande esforço de entendimento de diversas áreas de conhecimento, dada às imensas possibilidades de maior eficiência na prevenção e combate ao crime dentro do estado, mas também aos enormes riscos que essas tecnologias podem trazer a temas críticos como esse.

REFERÊNCIAS BIBLIOGRÁFICAS

CHOULDECHOVA, A. In Pursuit of Interpretable, Fair and Accurate Machine Learning for Criminal Recidivism Prediction. 2017. Disponível em: https://arxiv.org/abs/1704.02317. Acesso em: 19 mar. 2025.

EGBERT, Simon. LEESE, Matthias. **Criminal Futures: Predictive policing and everyday police work.** Nova York: Routledge, 2019.

GAMA, J. A survey on learning from data streams: current and future trends. **Prog Artif Intell** 1, 45–55 (2012). https://doi.org/10.1007/s13748-011-0002-6

GRUS, Joel. **Data Science do zero**. São Paulo: Alta Books, 2016.

GOODFELLOW, Ian e BENGIO, Yoshua. **Deep Learning**. Cambridge: MIT Press, 2016.

MITCHELL, T. **Machine Learning**. Columbus: McGraw Hill, 1997.

MORETTIN, Pedro A. e SINGER, Julio M. **Introdução à Ciência de Dados Fundamentos e Aplicações**. [Preprint] Disponível em https://www.ime.usp.br/~jmsinger/MAE5755/cdados2019ago06.pdf. Acesso em 15 mai. 2025.

PROVOST, F.; FAWCETT, T. **Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking**. Sebastopol: O'Reilly Media, 2013.

RIBEIRO, F. L. et al. Policiamento preditivo e inteligência artificial: análise de desempenho do algoritmo de aprendizado de máquina supervisionado Random Forest na predição de ocorrências policiais de roubo nas zonas da região metropolitana de São Luís/MA. 2020. **Revista FT** Disponível em: https://revistaft.com.br/policiamento-preditivo-e-inteligencia-artificial-analise-de-desempenho-do-algoritmo-de-aprendizado-de-maquina-supervisionado-random-forest-na-predicao-de-ocorrencias-policiais-de-roubo-nas-zonas-da-r/. Acesso em: 19 mar. 2025.

SALES, E. R. de; LUI, L. Perspectivas sobre segurança pública em cidades inteligentes: uma revisão da literatura de 2002 a 2022. **Revista de Gestão dos Países de Língua Portuguesa**, Rio de Janeiro, v. 22, n. 2, p. 83–101, 2023. DOI: 10.12660/rgplp.v22n2.2023.88882. Disponível em: https://periodicos.fgv.br/rgplp/article/view/88882. Acesso em: 3 abr. 2025.

SANTOS JÚNIOR, Ramiro de Vasconcelos dos. **Using machine learning to classify criminal macrocauses in smart city contexts**. Orientador: Dr. Nélio Alessandro Azevedo Cacho. 2024. 108f. Tese (Doutorado em Ciência da Computação) - Centro de Ciências Exatas e da Terra, Universidade Federal do Rio Grande do Norte, Natal, 2024.

SECRETARIA DA SEGURANÇA PÚBLICA DO ESTADO DE SÃO PAULO. **Funções e competências**. Disponível em: https://www.ssp.sp.gov.br/institucional/funcoes-e-competencias. Acesso em: 19 mar. 2025.

SECRETARIA DA SEGURANÇA PÚBLICA DO ESTADO DE SÃO PAULO. **Painel Estatístico**. Disponível em: https://www.ssp.sp.gov.br/estatistica/painel-estatistico. Acesso em: 19 mar. 2025.

SECRETARIA DA SEGURANÇA PÚBLICA DO ESTADO DE SÃO PAULO. **Histórico.** Disponível em: https://www.ssp.sp.gov.br/institucional/historico. Acesso em 19 mar. 2025.

SICSÚ, Abraham L.; SAMARTINI, André; BARTH, Nelson L. **Técnicas de machine learning**. São Paulo: Blucher, 2023.

DIÁRIO OFICIAL. **Poder Executivo - Seção I. 8 – São Paulo, 129 (223).** Disponível em: https://www.imprensaoficial.com.br/DO/BuscaDO2001Documento_11_4.aspx?link=%2f2019%2fexecutivo%2520secao%2520i%2fnovembro%2f26%2fpag_0008_1dcb169e9892d4aaea4f52dddb201ce6.pdf&pagina=8&data=26/11/2019&caderno=Executivo%20I&paginaordenacao=100008. Acesso em: 20 mar. 2025.

BRASÍLIA. **Decreto nº 9.489, de 30 de agosto de 2018.** Regulamenta, no âmbito da União, a Lei nº 13.675, de 11 de junho de 2018, para estabelecer normas, estrutura e procedimentos para a execução da Política Nacional de Segurança Pública e Defesa Social. Brasília. Presidência da República. Secretaria-Geral. Subchefia para Assuntos Jurídicos. Disponível em: https://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2018/Decreto/D9489.htm. Acesso em 20 mar. 2025.

BRASÍLIA. **Lei nº 13.675, de 11 de junho de 2018.** Disciplina a organização e o funcionamento dos órgãos responsáveis pela segurança pública, nos termos do § 7º do art. 144 da Constituição Federal; cria a Política Nacional de Segurança Pública e Defesa Social (PNSPDS); institui o Sistema Único de Segurança Pública (Susp); altera a Lei Complementar nº 79, de 7 de janeiro de 1994, a Lei nº 10.201, de 14 de fevereiro de 2001, e a Lei nº 11.530, de 24 de outubro de 2007; e revoga dispositivos da Lei nº 12.681, de 4 de julho de 2012. Brasília. Diário Oficial da União. Disponível em: https://www.in.gov.br/materia/-/asset_publisher/Kujrw0TZC2Mb/content/id/25212052/do1-2018-06-12-lei-n-13-675-de-11-de-junho-de-2018-25211917. Acesso em 21 mar. 2025.

BRASÍLIA. **Lei nº 13.709, de 14 de agosto de 2018.** Lei Geral da Proteção de Dados (LGPD). Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em 04 abr. 2025.

KAHN, Tulio,. ZANETIC, André. **O Papel dos Municípios da Segurança Pública.** GOV.BR: 2006. Disponível em: https://www.gov.br/mj/pt-br/assuntos/sua-seguranca/seguranca-publica/analise-e-pesquisa/do

wnload/estudos/sjcvolume1/papel_municipios_seguranca_publica.pdf. Acesso em 04 abr. 2025.

APÊNDICES

Código-fonte

```
#!/usr/bin/env python
# coding: utf-8

# In[ ]:


# 1. Importando Bibliotecas


# In[2]:


import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import CategoricalNB
from sklearn import metrics
from sklearn.metrics import confusion_matrix
from sklearn.ensemble import RandomForestClassifier


# In[ ]:


# 2. Importando o conjunto de dados


# In[ ]:


df = pd.read_csv("SPDadosCriminais_2023.csv")
```

df.info()

In[ ]:

3. Processamento de Dados

In[ ]:

3.1 Excluindo atributos não relevantes ao modelo

In[5]:

```
df.drop(['NOME_DEPARTAMENTO', 'NOME_SECCIONAL', 'NOME_DELEGACIA',
'NUM_BO', 'ANO_BO', 'DESC_PERIODO', 'DESCR_SUBTIPOLOCAL',
'NOME_DELEGACIA_CIRCUNSCRIÇÃO',
'NOME_DEPARTAMENTO_CIRCUNSCRIÇÃO',
 'NOME_SECCIONAL_CIRCUNSCRIÇÃO', 'DESCR_CONDUTA', 'MES_ESTATISTICA',
'ANO_ESTATISTICA'], axis=1, inplace=True)
```

In[ ]:

```
df.info()
```

In[ ]:

3.2 Filtrando dados nulos ou inúteis

In[7]:

```
df = df[df["LOGRADOURO"] != "VEDAÇÃO DA DIVULGAÇÃO DOS DADOS
RELATIVOS"]
```

In[ ]:

```
from pprint import pprint
pprint(df["RUBRICA"].unique())


# ## 3.3 Filtrando crimes não devidamente enquadrados no Código Penal

# In[9]:


excluir = ['Localização/Apreensão de objeto', 'Outros não criminal', 'Comunicação de óbito',
'Entrega de objeto localizado/apreendido', 'Atropelamento', 'Morte suspeita',
'Capotamento','Colisão', 'Choque','Engavetamento','Localização/Apreensão e Entrega de
objeto','Entrega de veículo localizado/apreendido', 'Cumprimento de mandado de prisão
temporária',
      'Localização/Apreensão de veículo','Auto lesão', 'Localização/Apreensão e Entrega de
veículo','Cumprimento de mandado de busca e apreensão', 'Entrada ilegal de aparelho movél
de comunicação em estabelecimento prisional', 'Abalroamento', 'Suicídio
consumado','homicídio culposo','Apreensão de Adolescente']

df = df[~df["RUBRICA"].isin(excluir)]


# In[10]:


df["RUBRICA"].unique()


# ## 3.4 Excluindo valores nulos

# In[11]:


df.dropna(inplace=True)


# In[12]:


df. head(3)


# ## 3.5 Obtendo valores atomizados a partir de datas

# In[13]:
```

```
df["DATA_REGISTRO"] = pd.to_datetime(df["DATA_REGISTRO"])
df["DATA_OCORRENCIA_BO"] = pd.to_datetime(df["DATA_OCORRENCIA_BO"])


# In[14]:


df["DIAS_REGISTRO"] = (df["DATA_REGISTRO"] -
df["DATA_OCORRENCIA_BO"]).astype(str)


# In[15]:


df["DIAS_REGISTRO"] = df["DIAS_REGISTRO"].str[0]


# In[16]:


df["MES_REGISTRO"] = df["DATA_REGISTRO"].dt.month
df["DIA_REGISTRO"] = df["DATA_REGISTRO"].dt.day


# In[17]:


df["MES_OCORRENCIA"] = df["DATA_OCORRENCIA_BO"].dt.month
df["DIA_OCORRENCIA"] = df["DATA_OCORRENCIA_BO"].dt.day


# In[18]:


df["HORA_OCORRENCIA_BO"] = df["HORA_OCORRENCIA_BO"].str[:2]


# In[19]:


df.head(3)


# ## 3.6 Reorganizando o conjunto de dados

# In[20]:
```

```
df = df[['NOME_MUNICIPIO', 'DIAS_REGISTRO', 'MES_REGISTRO', 'DIA_REGISTRO',
'MES_OCORRENCIA', 'DIA_OCORRENCIA',
       "HORA_OCORRENCIA_BO", 'BAIRRO', 'LOGRADOURO',
      'NUMERO_LOGRADOURO', 'LATITUDE', 'LONGITUDE',
      'NOME_MUNICIPIO_CIRCUNSCRIÇÃO', 'NATUREZA_APURADA', 'RUBRICA']]


# In[21]:


df.head(3)


# ## 3.7 Transformando valores em formato de cadeia em valores numéricos

# In[22]:


df.info()


# In[23]:


df["DIAS_REGISTRO"] = df["DIAS_REGISTRO"].astype(np.int64)
df["HORA_OCORRENCIA_BO"] = df["HORA_OCORRENCIA_BO"].astype(np.int64)
df["NUMERO_LOGRADOURO"] = df["NUMERO_LOGRADOURO"].astype(np.int64)


# In[24]:


df.info()


# ## 3.8 Modularizando coordenadas geográficas

# In[25]:


df["LATITUDE"] = df["LATITUDE"].abs()
df["LONGITUDE"] = df["LONGITUDE"].abs()


# In[26]:
```

```
df.head(3)


# ## 3.9 Excluindo atributos sobressalentes após experimentos iniciais

# In[27]:


df.drop(["NOME_MUNICIPIO", "BAIRRO",
"LOGRADOURO","NOME_MUNICIPIO_CIRCUNSCRIÇÃO",
"NATUREZA_APURADA" ], axis=1, inplace=True)


# In[28]:


df.head(3)


# ## 3.10 Aplicando a função LabelEnconder ao atributo alvo

# In[29]:


labelenconder = LabelEncoder()

df["CLASSIFICACAO"] = labelenconder.fit_transform(df["RUBRICA"])


# In[30]:


df.head(3)


# ## 3.11 Organizando os dados

# In[31]:


df = df[['DIAS_REGISTRO', 'MES_REGISTRO', 'DIA_REGISTRO',
'MES_OCORRENCIA', 'DIA_OCORRENCIA',
      "HORA_OCORRENCIA_BO", 'LATITUDE', 'LONGITUDE',
      'CLASSIFICACAO','RUBRICA']]
```

In[32]:

```
df.head(3)
```

3.12 Diminuindo a quantidade de atributos alvo

In[33]:

```
contador = df["RUBRICA"].value_counts()
df.value_counts()
```

In[34]:

```
df = 
df[df["RUBRICA"].isin(df["RUBRICA"].value_counts()[df["RUBRICA"].value_counts() > 
200].index)]
```

In[35]:

```
df["RUBRICA"].value_counts()
```

3.13 Balanceado os dados

In[36]:

```
teste = df.sample(100000)

teste
```

In[37]:

```
teste["RUBRICA"].value_counts()
```

4. Treinando o modelo

4.1 Separando atributos base e atributos alvo

In[38]:

```
X = df.iloc[:, : -2]
y = df.loc[:, "CLASSIFICACAO"]
X
```

4.2 Separando conjunto de teste e treinamento

In[39]:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, train_size=0.8,
random_state=1)
```

In[40]:

```
len(X_train), len (X_test), len(y_train), len(y_test)
```

4.3 Instanciando o treinamento do modelo Categorical Naive Bayes

In[41]:

```
cnb = CategoricalNB()
cnb.fit(X_train, y_train)
```

4.4 Checando métricas

In[42]:

```
round(cnb.score(X_train, y_train),4)
```

In[43]:

```
cnb.score(X_test, y_test)
```

4.5 Instanciando o treinamento do modelo random forest

In[44]:

```
rf = RandomForestClassifier(n_estimators=100, random_state=42)
```

In[45]:

```
rf.fit(X_train,y_train)
```

4.6 Checando métricas

In[48]:

```
rf.score(X_train, y_train)
```

In[47]:

```
rf.score(X_test, y_test)
```