UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO ACADÊMICO DO AGRESTE
NÚCLEO DE TECNOLOGIA
CURSO DE BACHARELADO EM ENGENHARIA DE PRODUÇÃO

LUANN BRUNO VIDAL DE ANDRADE

APLICAÇÃO DE ALGORITMOS DE MACHINE LEARNING NA PREVISÃO DE MORTES VIOLENTAS INTENCIONAIS NO MUNICÍPIO DE CARUARU, INTERIOR DE PERNAMBUCO

Caruaru
2024

LUANN BRUNO VIDAL DE ANDRADE

APLICAÇÃO DE ALGORITMOS DE MACHINE LEARNING NA PREVISÃO DE MORTES VIOLENTAS INTENCIONAIS NO MUNICÍPIO DE CARUARU, INTERIOR DE PERNAMBUCO

Trabalho de Conclusão de Curso apresentado ao Curso de Engenharia de Produção da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Barachel em Engenharia de Produção.

Área de concentração: Pesquisa Operacional.

Professor: Lucio Camara e Silva

Caruaru
2024

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Andrade, Luann Bruno Vidal de.
Aplicação de algoritmos de machine learning na previsão de mortes violentas intencionais no município de Caruaru, interior de Pernambuco / Luann Bruno Vidal de Andrade. - Caruaru, 2024.
48 p. : il., tab.

Orientador(a): Lucio Camara e Silva
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro Acadêmico do Agreste, Engenharia de Produção, 2024.
Inclui referências.

1. Mortes Violentas Intencionais. 2. Segurança pública. 3. Machine learning. 4. Regressão estatística. I. Silva, Lucio Camara e. (Orientação). II. Título.

310 CDD (22.ed.)

LUANN BRUNO VIDAL DE ANDRADE

APLICAÇÃO DE ALGORITMOS DE MACHINE LEARNING NA PREVISÃO DE MORTES VIOLENTAS INTENCIONAIS NO MUNICÍPIO DE CARUARU, INTERIOR DE PERNAMBUCO

Trabalho de Conclusão de Curso apresentado à Coordenação do Curso de Engenharia de Produção do Campus Agreste da Universidade Federal de Pernambuco – UFPE, na modalidade de monografia, como requisito parcial para a obtenção do grau de bacharel em Engenharia de Produção.

Aprovada em: 17/10/2024 às 11:00hs

BANCA EXAMINADORA

_____

Prof. Dr. Lúcio Camara e Silva (Orientador)
Universidade Federal de Pernambuco

_____

Prof. Dr. Lucimário Gois de Oliveira Silva (Examinador Interno)
Universidade Federal de Pernambuco

_____

Prof. Dr. Thyago Celso Cavalcante Nepomuceno (Examinador Externo)
Universidade Federal de Pernambuco

RESUMO

Considerando o atual cenário da segurança pública no Brasil e a importância da compreensão e análise dos dados de indicadores criminais para a criação de políticas públicas eficazes no combate e repreensão dos crimes, este trabalho se propõe a construir um modelo de previsão de Mortes Violentas Intencionais (MVI), utilizando algoritmos de Machine Leaning (ML) para a identificar os principais fatores influenciadores dos MVI no município de Caruaru em Pernambuco, assim como prever os futuros números do indicador. Os dados utilizados neste estudo foram disponibilizados pela Secretaria de Ordem Pública da Prefeitura de Caruaru (SECOP) e compreendem o período de Janeiro de 2017 a Junho de 2024. No trabalho foram utilizados os algoritmos de machine learning XGBoost e Random Forest afim de determinar o método mais adequado para o problema proposto. A avaliação dos resultados indicou que o modelo de XGBoost apresentou o melhor desempenho, com um erro absoluto percentual médio (MAPE) de 5,7% e um coeficiente de determinação (R^2) de 0,6042. A análise dos fatores influentes apontou a predominância de variáveis geoespaciais e temporais como os principais determinantes dos números de MVI, embora variáveis relacionadas ao perfil do crime e da vítima também tenham mostrado impacto relevante.

Palavras-chave: mortes violentas intencionais. segurança pública. machine learning. regressão estatística.

ABSTRACT

Considering the current public security situation in Brazil and the importance of understanding analyzing criminal indicator data to create effective public policies for crime prevention and repression, this work proposes tobuild a predictive model of Intentional Violent Deaths (MVI) using Machine Learning (ML) algorithms to identify the main influencing factors of MVI in the city of Caruaru in Pernambuco, as well as to predict future numbers of this indicator. The data used in this study were provided by the Department of Public Order of the Caruaru City Hall (SECOP) and cover the period from January 2017 to June 2024. The work utilized the XGBoost and Random Forest based machine learning algorithms to determine the most suitable method for the proposed problem. The evaluation of the results indicated that the XGBoost model showed the best performance, with a mean absolute percentage error (MAPE) of 5.7% and a coefficient of determination (R^2) of 0.6042. The analysis of influential factors highlighted the predominance of geospatial and temporal variables as the main determinants of MVI numbers, although variables related to the crime and victim profile also showed relevant impact.

Keywords: intentional violent deaths. public safety. machine learning. statistical regression.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

A violência tem se destacado como uma das principais pautas no cenário nacional. De acordo com dados do Anuário Brasileiro de Segurança Pública de 2024 (ABSP), em 2023 o Brasil empenhou um total de R$ 137,9 bilhões em segurança pública, valor superior aos gastos da União com a educação, que em 2023 atingiram R$ 100,8 bilhões, marcando um aumento de 4,9% em relação ao ano de 2022, mesmo em um cenário de restrição orçamentária (FBSP, 2024). Dados do Instituto Brasileiro de Geografia e Estatística (IBGE) mostram que, em 2023, o país somou um produto interno bruto (PIB) de R$ 10,9 trilhões (IBGE, 2024), sendo assim, os gastos com segurança pública representam cerca de 1,26% do PIB. Entretanto, o custo real da violência pode ser significativamente maior do que os valores apresentados, segundo CERQUEIRA (2017) existem custos econômicos intangíveis que, embora não apareçam diretamente como despesas públicas, contribuem para diminuir a qualidade de vida e o bem-estar da população.

Apesar dos elevados investimentos públicos, o Brasil continua apresentando altos índices de violência. Ainda de acordo com dados do ABSP, em 2023 foram registrados 46.328 Mortes Violentas Intencionais (MVI) no território nacional, resultando em uma taxa de 22,8 MVIs a cada 100 mil habitantes. Já em 2021, o Brasil registrou 47.722 MVIs, o maior número registrado no mundo (WELLE, 2023). Estes resultados demonstram um cenário de crise na segurança pública do país, com uma preocupante banalização da vida, evidenciando falhas nas estratégias de alocação de recursos e na formulação de políticas públicas eficazes no combate ao crime.

Neste mesmo contexto, Jaitman et al. (2017) argumenta sobre a urgência na construção do conhecimento necessário para empregar políticas públicas baseadas em evidências. Em resposta à crise de segurança, o governo estadual de Pernambuco adotou, por muitos anos, o programa de segurança pública Pacto Pela Vida (PPV), o qual tinha como principal objetivo a redução nos números do indicador Mortes Violentas Letais Intencionais (CVLI), equivalente ao atual MVI. O PPV se baseava na gestão por resultados, implementando medidas de repressão qualificada, prevenção social e modernização das agências de segurança com base na qualificação dos casos de CVLI (SAURET, 2022). Estima-se que, em 14 anos de vigência, o PPV foi responsável por salvar 17.388 vidas (SDS, 2021).

Em Caruaru, a violência também alcança níveis alarmantes sendo, em 2023, a sexta cidade com maior número de MVIs em Pernambuco. Apesar disso, Caruaru vem apresentando reduções consecutivas na taxa de MVI em relação à média estadual. Em 2017 Caruaru adotou o programa municipal de segurança pública Juntos pela Segurança (JPS), saindo de uma taxa de MVI 27% maior que a taxa de Pernambuco em 2017 para uma taxa 16% menor, em 2023. Similar ao PPV, o JPS tem como atribuição "atuar com fundamentos no modelo de gestão por resultados, através do monitoramento das metas estabelecidas e acompanhamento das ações definidas para a diminuição dos indicadores de violência"trazendo para o contexto da segurança

pública a integração entre os diversos órgãos do âmbito municipal em consonância com o poder público federal e estatual (CARUARU, 2017).

Na literatura, é amplamente reconhecida a associação entre a ocorrência de crimes e uma variedade de fatores, incluindo traços psicológicos, condições do ambiente, padrões espaciais, e indicadores sociais e econômicos (ALVES; RIBEIRO; RODRIGUES, 2018). Diante disso, é razoável assumir que esses fatores influenciam diretamente os indicadores criminais. A identificação desses elementos possibilita o desenvolvimento de políticas públicas direcionadas, voltadas tanto à mitigação de crimes quanto à manutenção de ambientes mais seguros. À medida que vários fatores ambientais são identificados como influenciadores das oportunidades de crime, previsões detalhadas de crimes que incorporem esses fatores tornam-se possíveis (ZHANG et al., 2022). A maioria dos trabalhos no contexto de previsão de crimes aplicam modelos de regressão linear tradicionais, no entanto, como sugerido por Alves, Ribeiro e Rodrigues (2018), tais modelos podem apresentar dificuldades ao lidar com os padrões complexos nos dados criminais, levando a conclusões errôneas. Por outro lado, algoritmos de Machine Learning (ML) tendem a apresentar melhores resultados que modelos de regressão linear tradicionais no contexto de previsão de crimes (ZHANG et al., 2022).

Machine learning é a ciência de fazer computadores tomar decisões sem intervenção humana (KIM et al., 2018). Tais modelos são capazes de ingerir séries de dados complexos e multidimensionais incluindo dados de variáveis históricas e geográficas, dispensando a necessidade de especificar um algorítimo específico para a regressão dos dados, ficando a cargo do próprio computador decidir qual função é a mais adequada (ZHANG et al., 2022).

Dessa forma, este trabalho visa contribuir para a eficácia no direcionamento das ações de segurança pública, especificamente para o cenário do município de Caruaru, através da utilização de algoritmos de machine learning para identificação de fatores influentes nos números de MVI, além de fornecer uma métrica robusta para a avaliação dos resultados.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Aplicar algoritmos preditivos capazes de estimar os resultados e fornecer insights sobre os dados de Mortes Violentas Intencionais (MVI) no município de Caruaru, Pernambuco, a fim de avaliar os resultados e propor políticas públicas de intervenção.

1.1.2 Objetivos Específicos

Para atingir o objetivo geral, os seguintes objetivos específicos devem ser cumpridos:

- Coletar dados referentes a MVI e formular as técnicas de tratamento de dados necessárias;

- Aplicar algoritmos preditivos, utilizando a linguagem Python, para a previsão dos números de MVI para os próximos seis meses afim de, ao final de cada período, avaliar os resultados obtidos;

- Aplicar algoritmos preditivos para a identificação de variáveis significativas nos números de MVI, visando orientar a formulação de políticas públicas de intervenção;

1.2 JUSTIFICATIVA E RELEVÂNCIA DO TRABALHO

Diante de um cenário de crise em segurança pública, o bem-estar da população é afetado diretamente, tanto de forma financeira, forçando os indivíduos a empenhar recursos para se proteger e afastando investimentos, quanto de forma física e psicológica, causando dados muitas vezes irreparáveis às famílias e forçando a mudança de comportamento dos indivíduos para evitar o crime, (JAITMAN et al., 2017), (KUROKI, 2013). Portanto, é evidente a necessidade da criação de métodos que contribuam para a alocação de recursos públicos de forma eficiente na área de segurança pública, voltadas tanto à mitigação de crimes quanto à manutenção de ambientes mais seguros.

Através da análise de dados históricos sobre os casos de mortes violentas intencionais, este trabalho contribui para a identificação de padrões temporais relacionados à ocorrência desse tipo de crime, possibilitando um planejamento mais eficiente do emprego de forças de segurança. Além disso, este trabalho visa identificar fatores reais no contexto da cidade de Caruaru que contribuem para o aumento da criminalidade, possibilitando a adoção de medidas que favoreçam o bem-estar da população do município. Com a aplicação de algoritmos de machine learning para a previsão de crimes, o presente trabalho contribui com a consolidação da utilização destes modelos no contexto da segurança pública ao reproduzir e adaptar técnicas de análise de dados referentes aos registros criminais já aplicadas em cidades e contextos distintos.

2 REFERENCIAL TEÓRICO

2.1 A VIOLÊNCIA URBANA E SEU IMPACTO NA SOCIEDADE

A violência, seja no campo ou nas cidades, sempre ocorreu, assumindo formas específicas conforme o momento histórico, e atingindo, preferencialmente, as camadas subalternas da população (COSTA, 1999). A violência urbana é, ao mesmo tempo, um fenômeno social e um problema de ordem estrutural que pode ser observado em cidades de todo o mundo, sejam elas metrópoles globais, cidades médias ou cidades pequenas (GUITARRARA, 2024b).

Guitarrara (2024a) argumenta que um dos principais fatores causadores da violência no Brasil é a falta de punições rígidas o suficiente. Durante a história, sempre foi de interesse das sociedades prevenir os crimes, punindo indivíduos que cometem tais atos, de forma proporcional ao delito (GORDON et al., 2009). De acordo com os resultados obtidos por Gordon et al. (2009), punir crimes resulta em um impacto positivo no nível de honestidade de uma população. Adicionalmente, a ocorrência de crimes pode ser associada a diversos fatores de caráter socio-econômico, como nível de desemprego, riqueza, nível de educação, entre outros (BOGOMOLOV et al., 2014). Alves et al. (2013) investiga a relação entre o tamanho da população e métricas urbanas como PIB, renda, desemprego, entre outros, com a ocorrência de crimes. De acordo com o apresentado, essa relação tende a seguir leis de escala, indicando que a ocorrência de crimes, incluindo homicídios, crescem de forma não linear, conforme a população aumenta. Ainda, Becker (1974) argumenta que o nível de instrução formal de um indivíduo reduz a chance do mesmo se envolver em atividades ilícitas.

O crime afeta diretamente a qualidade de vida e o desenvolvimento econômico de uma população (BOGOMOLOV et al., 2014). Kuroki (2013) estuda a relação entre o crime e o bem estar da população, relacionando também a renda dos indivíduos estudados. Os resultados mostram o impacto negativo que uma ocorrência de crime pode causar na felicidade reportada pela vítima, e ainda que esse impacto é potencializado em indivíduos de baixa renda. Em estudo realizado em países da América Latina e Caribe, Parente (2023) mostra que a redução nos níveis de criminalidade aumentariam de forma significativa o crescimento econômico da região. Jaitman et al. (2017) aponta que, perante os altos índices de criminalidade, os custos do crime podem ser significativos e estima que o crime no Brasil custa 3,14% do PIB.

2.2 ANÁLISE DE DADOS EM SEGURANÇA PÚBLICA

No Brasil, as políticas públicas de segurança são comumente voltadas à *defesa social*, focando em estatísticas a fim de identificar onde o crime tende a se concentrar no tempo e no espaço (DURANTE et al., 2022). ALDADO (2021) destaca que, nos últimos tempos, o investimento na segurança pública focou no contingente policial, modernização de comunicação e

frotas de veículos. Conforme verificado por Araujo (2019), existem evidências da efetividade do patrulhamento de pontos quentes de crimes.

No entanto, conforme o avanço da criminologia, novos conceitos foram desenvolvidos colocando maior relevância no entendimento dos padrões relacionados a causas e consequências dos crimes (DURANTE et al., 2022), levando à necessidade de mapear extensivamente os fatores que influenciam os mesmos. Para Lira, Caballero e Nascimento (2022), toda decisão no âmbito da segurança pública deve ser baseada em análise das estatísticas criminais. No entanto, em seu trabalho Lira, Caballero e Nascimento (2022) apontam como dificuldade a obtenção de informações a partir dos dados e a qualidade e padronização dos dados acerca de eventos criminais. Nesse contexto, nos últimos anos o governo brasileiro direcionou esforços para criação de medidas que visam a unificação e padronização do registro e compartilhamento de dados de segurança pública aprimorando recursos da plataforma do Sistema Nacional de Informações de Segurança Pública, Prisionais e sobre Drogas (Sinesp) (MJSP, s.d.).

Na literatura é possível encontrar diferentes abordagens na utilização de dados de segurança pública. Cunha (2022) propõe uma metodologia baseada em estatística descritiva para explicar padrões e tendências entre a variáveis explicativas dos crimes de morte em Caruaru-PE. Alves et al. (2013) utiliza modelos de regressão linear simples relacionando números de homicídios com indicadores urbanos. Sauret (2022) utiliza métodos de projeções dos indicadores de criminalidade para verificar a efetividade ações de intervenção. Embora os métodos tradicionais apresentem vantagens, Alves, Ribeiro e Rodrigues (2018) argumenta que técnicas baseadas em aprendizado de máquina são mais adequadas para lidar com a complexidade dos dados e sua natureza não linear.

2.3 MACHINE LEARNING

Machine learning é a ciência de fazer computadores tomar decisões sem intervenção humana (KIM et al., 2018). Machine learning está associado ao conceito de Inteligência Artificial (AI), o termo "inteligência artificial"pode aplicado a qualquer dispositivo que executa funções que humanos associam à mente humana, como "aprender"e "resolver problemas". Aprender é um aspecto essencial das máquinas, portanto, machine learning é uma subcampo da IA (SHINDE; SHAH, 2018). Portanto, machine learning pode ser definido como a utilização de computadores para a execução de atividades tipicamente associadas aos humanos.

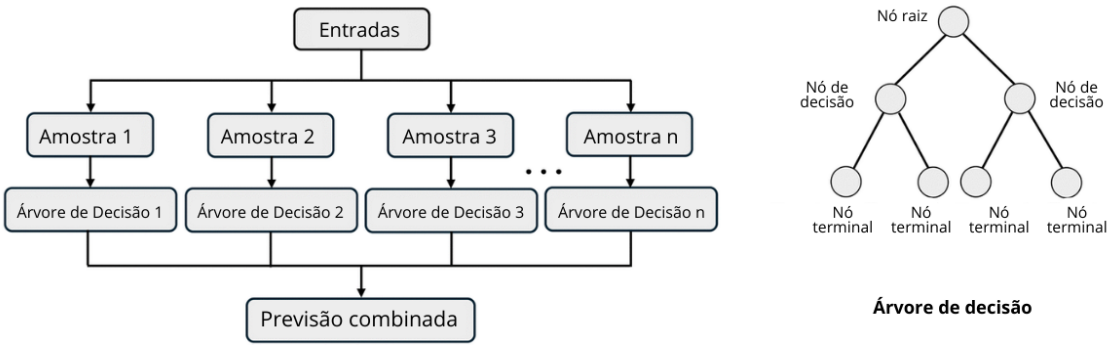
Os algoritmos de Machine learning podem ser divididos em várias subcategorias, sendo as principais o aprendizado supervisionado, aprendizado não-supervisionado e aprendizado reforçado (KIM et al., 2018). Como descrito por Rabbani et al. (2022), algoritmos de aprendizagem supervisionada são algoritmos que, ao contrário da aprendizagem não supervisionada, esperam receber dados já rotulados onde a variável objetivo é conhecida, auxiliando o modelo a gerar previsões mais precisas ao avaliar as relações entre as variáveis de inputs ao modelo. Ainda, conforme descrito por Hastie, Tibshirani e Friedman (2001), os métodos de aprendizado

supervisionado são conhecidos como *aprendizado através do exemplo*. Em um problema onde pode-se assumir o modelo $Y = f(x)+\varepsilon$, o método de aprendizado supervisionado tenta "aprender"a função f através de um *professor*, que se refere a um conjunto de dados de exemplos conhecidos (dados de treino), contendo dados tanto das variáveis explicatórias como da variável dependente. Após o aprendizado, espera-se que os valores estimados pela função estimada $\hat{f}$ sejam próximos o suficiente dos valores reais ao ponto de serem úteis na prática.

Seguindo com foco no aprendizado supervisionado, pode-se tomar como exemplo algoritmos como Random Forest e Gradient Boosting. Tais algoritmos são exemplos de técnicas de aprendizado supervisionado baseadas em árvores de decisão, usados tanto em problemas de classificação como em regressão (BELYADI; HAGHIGHAT, 2021). Árvores de decisão são algorítimos baseados em uma estrutura semelhante aos galhos de uma árvore, onde cada divisão nos galhos representam uma teste realizado em uma variável, enquanto os ramos indicam os resultados gerados por esse teste. Como resultado, tem-se uma estrutura correspondente às regras de decisão definidas pelo modelo baseado nos valores para cada variável (SAHA; MANICKAVASAGAN, 2021). Além disso, os algoritmos random forest e gradient boosting são considerados métodos de "ensemble", que significa que tais métodos avalia os resultados de diferentes árvores de decisão combinando-os em um modelo mais robusto, melhorando a precisão do modelo (SAHA; MANICKAVASAGAN, 2021).

Na Figura 1 uma ilustração da estrutura de um modelo de machine learning baseado em random forest além da estrutura de uma árvore de decisão. Inicialmente o algoritmo constrói múltiplas árvores de decisão usando diferente sub-conjuntos de dados dos dados de treino, e posteriormente, une as árvores criadas em um modelo mais robusto (WANG; CHAKRABORTY; CHAKRABORTY, 2020).

Figura 1 – Representação gráfica da arquitetura do Random Forest (esquerda) e um modelo de árvore de decisão (direita)

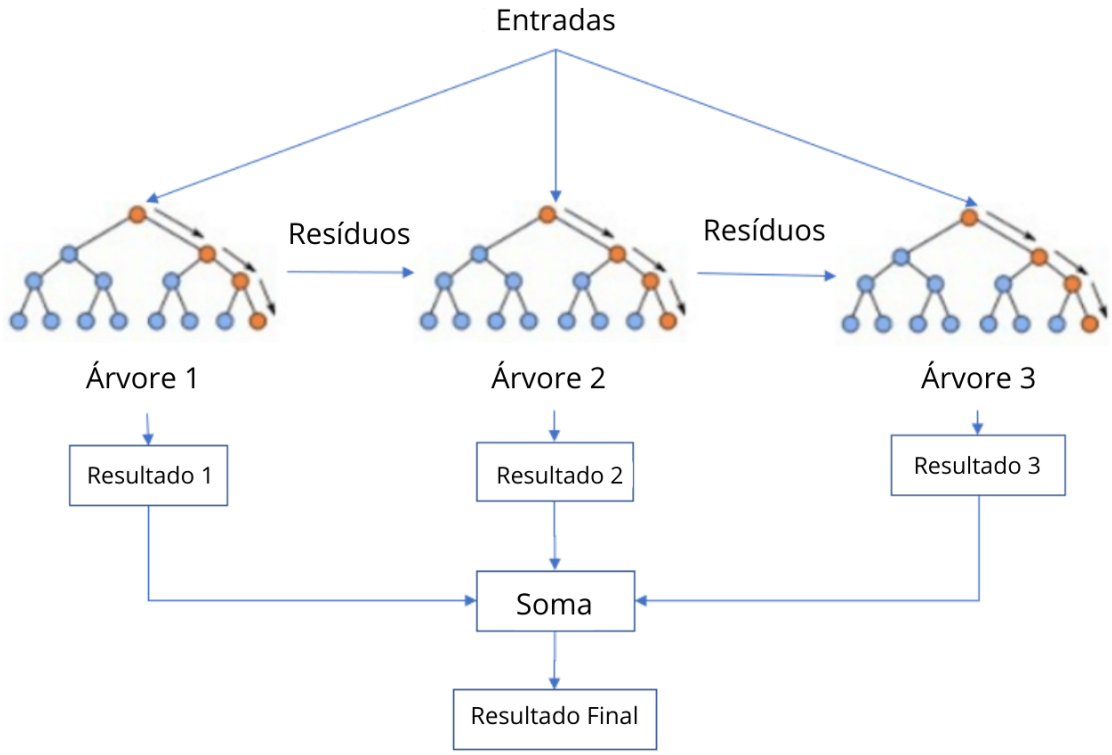


Fonte: SANDUNIL et al. (2024). (adaptado)

Em contraste, os métodos de gradient boosting trabalham de forma sequencial, onde cada árvore depende do resultado da anterior, resultado em um modelo ainda mais preciso e eficiente (DJON et al., 2023), conforme a representação da Figura 2. Diferente do Random Forest, cada

modelo de árvore no XGBoost minimiza o resíduo do seu modelo de árvore anterior (WANG; CHAKRABORTY; CHAKRABORTY, 2020).

Figura 2 – Representação gráfica da arquitetura do XGBoost



Fonte: WANG; CHAKRABORTY; CHAKRABORTY (2020). (adaptado)

2.4 MACHINE LEARNING E SUAS APLICAÇÕES NA PREVISÃO DE CRIMES

Em situações onde existe uma grande quantidade de dados, com relações complexas entre as variáveis, torna-se difícil para seres humanos a interpretação e extração de informações a partir dos dados. Nesse caso, é indicado a utilização de algoritmos capazes de aprender com os dados, sem a necessidade de programa-los de forma específica (MAHESH, 2019). Métodos de machine learning possuem diversas aplicações, como reconhecimento facial, reconhecimento de voz, direção automática, diagnóstico médico, entre outros (ZHANG et al., 2022). No caso particular da segurança pública, algoritmos de machine learning são comumente utilizados na previsão de pontos quentes (ARAUJO, 2019) (BORGES et al., 2017), categorização de crimes (KHAN; ALI; ALHARBI, 2022) (KWON; JUNG; LEE, 2021), identificação de fatores influentes na criminalidade (DJON et al., 2023), predição de taxa de crimes (AZIZ; SHARMA; HUSSAIN, 2024), entre diversas outras aplicações. Furtado (2022) resume as aplicações de inteligências artificiais na segurança pública na figura 3.

Figura 3 – Resumo das aplicações de algoritmos de IA na segurança pública

Fonte: FURTADO (2022). (adaptado)

De acordo com (ARAUJO, 2019), não existe um consenso na literatura sobre qual algoritmo de machine learning é o mais adequado para problemas de predição de crimes. No entanto, os algoritmos baseados em Random Forest e Gradient Boosting são amplamente considerados apropriados, conforme evidenciado pelos trabalhos do próprio Araujo (2019), além de Djon et al. (2023), Aziz, Sharma e Hussain (2024), Khan, Ali e Alharbi (2022), Zhang et al. (2022), entre outros. Em particular, os modelos baseados em Gradient Boosting são frequentemente apontados como os mais precisos para este tipo de problema, como nos trabalhos de Zhang et al. (2022), Djon et al. (2023), e Khan, Ali e Alharbi (2022).

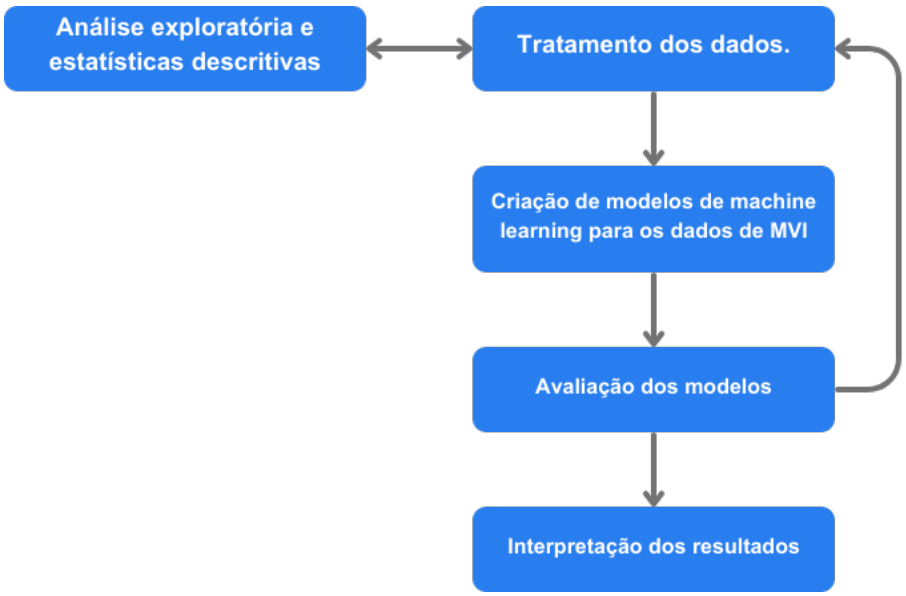
3 MATERIAIS E MÉTODOS

Neste capítulo serão apresentados os procedimentos e técnicas utilizados no desenvolvimento deste estudo, além de uma descrição dos dados obtidos para análise. A metodologia utilizada visa aplicar algoritmos preditivos de machine learning otimizado para os dados obtidos permitindo uma análise robusta dos resultados.

3.1 ETAPAS DA PESQUISA

O presente estudo segue as etapas descritas no seguinte fluxograma:

Figura 4 – Etapas do estudo



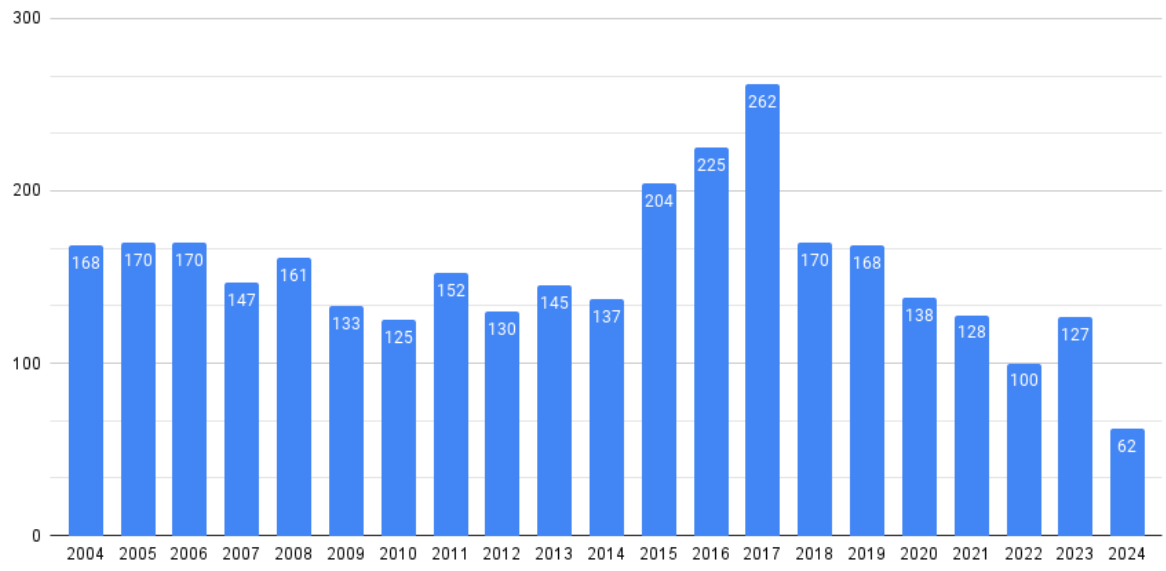
Fonte: o autor (2024).

Inicialmente são aplicadas técnicas de análise exploratória e estatísticas descritivas para a compreensão dos dados e identificação de padrões de crescimento/decrescimento ao longo da série histórica. Paralelamente são feitas operações de tratamento, limpeza e transformação dos dados de acordo com o observado nas análises assegurando a qualidade e adequação do conjunto de dados para a aplicação dos modelos de regressão, além da criação de novas variáveis a partir dos dados já existentes. Na sequência, são aplicados os algoritmos de machine learning e os resultados são avaliados como satisfatórios ou se são necessárias novas modificações. Finalmente, os resultados são analisados em profundidade e interpretados.

3.2 DESCRIÇÃO DOS DADOS

O indicador criminal Mortes Violentas Intencionais (MVI), recentemente implementado pela Secretaria de Defesa Social de Pernambuco (SDS) por meio da Portaria Nº 1066 de 11 de março de 2023, abrange os seguintes tipos de crimes: Homicídio doloso; Feminicídio; Latrocínio; Lesão corporal seguida de morte; Outros crimes resultantes em morte; Morte por intervenção de agente do Estado, sendo este indicador correspondente ao antigo indicador Crimes Violentos Letais Intencionais (CVLI). Após a mudança, o CVLI deixa de contabilizar os casos de Morte por intervenção de agente do Estado. Os dados obtidos através da SECOP são provenientes de três fontes principais: a própria SECOP, a 14ª Delegacia Seccional de Polícia Civil (14ª Desec) e SDS, e são referentes aos MVIs registrados em Caruaru entre o período de janeiro de 2017 a junho de 2024 e totalizando 1154 casos de MVI no município no período mencionado. Vale ressaltar que, de acordo com dados históricos da SDS, o ano de 2017 apresentou um pico significativo nos crimes de morte, sendo o ano mais violento de toda a série histórica desde 2004, quando os dados se iniciam. Portanto, o período analisado representa uma queda expressiva nos números de MVI.

Figura 5 – Série histórica dos dados de MVI - SDS/PE (dados parciais em 2024)



Fonte: SDS (2024).

Kwon, Jung e Lee (2021) argumenta que crimes de oportunidade, como roubo e furto, são mais afetados por variáveis geográficas, enquanto crimes de morte, como os crimes inclusos no indicador MVI, são influenciados por sentimentos pessoais e miram em vítimas específicas. Portanto, ainda que os dados disponibilizados publicamente pela SDS/PE abrangem um período de tempo maior, optou-se pela utilização dos dados disponibilizados pela SECOP, em razão da maior quantidade de informações em cada registro, incluindo informações tanto sobre o perfil

das vítimas quanto sobre variáveis geográficas e o perfil dos crimes. Algumas das principais informações incluem o georreferenciamento, a natureza jurídica dos crimes, a motivação, os antecedentes criminais das vítimas, o gênero das vítimas, o horário, entre outros aspectos relevantes. Abaixo, uma explicação mais detalhada sobre os dados obtidos.

- **Coordenadas, Bairro, TGS, Zona**: Essas colunas dizem respeito à localização do crime. As coordenadas são utilizadas para determinar o bairro/sítio onde ocorre cada crime. Utiliza-se a denominação TGS (Território de Gestão Sustentável) para determinadas as zonas que compõem o município (figura 6) que apresentam diferentes características geográficas e socioeconômicas, e que apresentam diferentes tipos de problemas (LYRA RAQUEL E PINHEIRO, 2020), semelhante às macrorregiões do Brasil. A classificação da Zona (rural ou urbana) é determinada a partir do TGS: TGSs de 1 a 4 são considerados Zona Rural, enquanto TGSs de 5 a 9 são considerados Zona Urbana. Há uma exceção para o bairro Canaã, que está localizado no TGS 2, no entanto é considerado Zona Urbana.

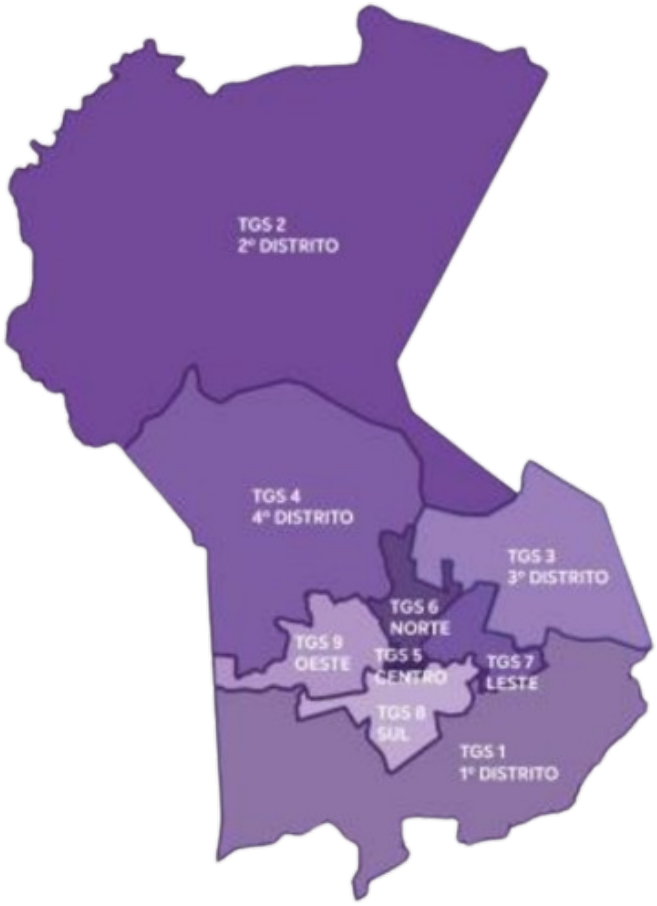
- **Natureza Jurídica do Fato**: Esta é determinada pela Polícia Civil conforme a portaria nº 229, de 10 de Dezembro de 2018, do Diário Oficial da União.

- **Motivação do Crime e Grupo de Motivação**: Determinada pela Polícia Civil de acordo com o catálogo de motivações da portaria nº 6195, de 30/12/2019 da SDS, bem como o grupo de motivações.

- **Possui Envolvimento com Drogas**: Esta classificação refere-se a crimes com qualquer motivação relacionada a drogas ou quando é constatado pelo Instituto de Medicina Legal (IML) o consumo de drogas por parte da vítima.

- **Outras Informações do Perfil das Vítimas e dos Crimes**: Como gênero, idade, data e hora do fato, e antecedentes criminais, são determinadas com base na investigação conduzida pela Polícia Civil.

Figura 6 – Territórios de Gestão Sustentável (TGS) de Caruaru

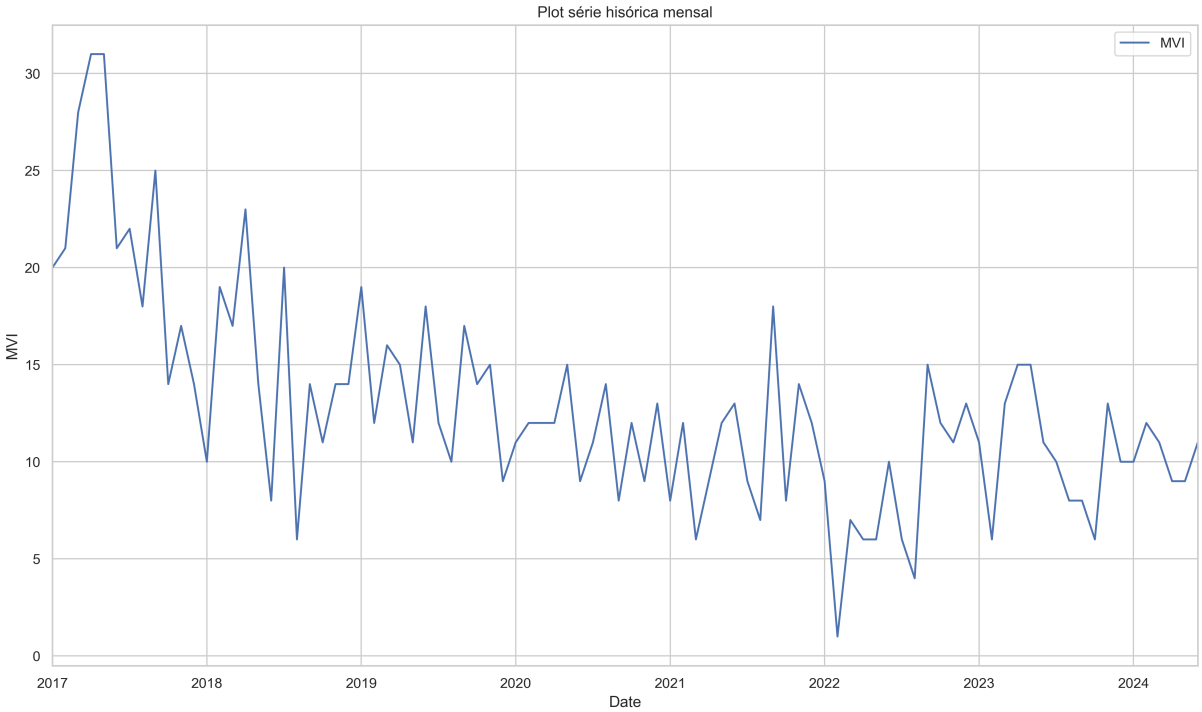


Fonte: LYRA RAQUEL E PINHEIRO (2020).

3.3 ANÁLISE EXPLORATÓRIA

Os dados coletados compreendem o período de janeiro de 2017 a junho de 2024, totalizando um período de 7 anos e meio. Os dados foram agrupados em meses, conforme o gráfico da Figura 7:

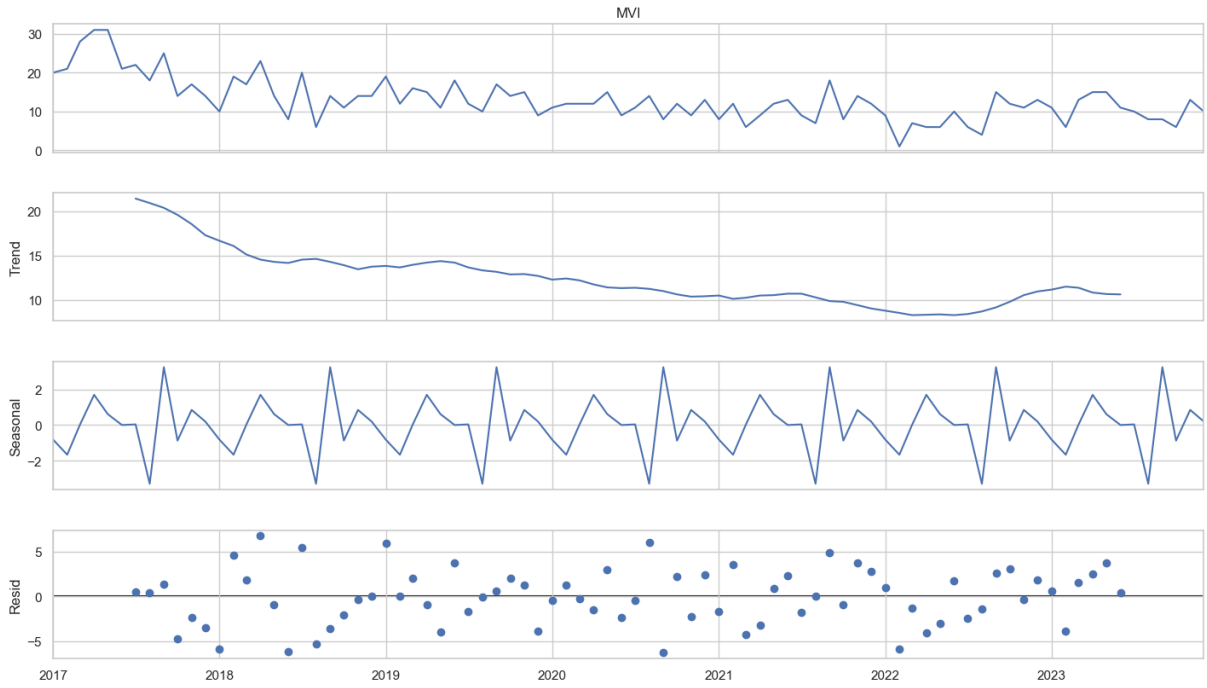
Figura 7 – MVI por mês 2017 a 2024



Fonte: o autor (2024).

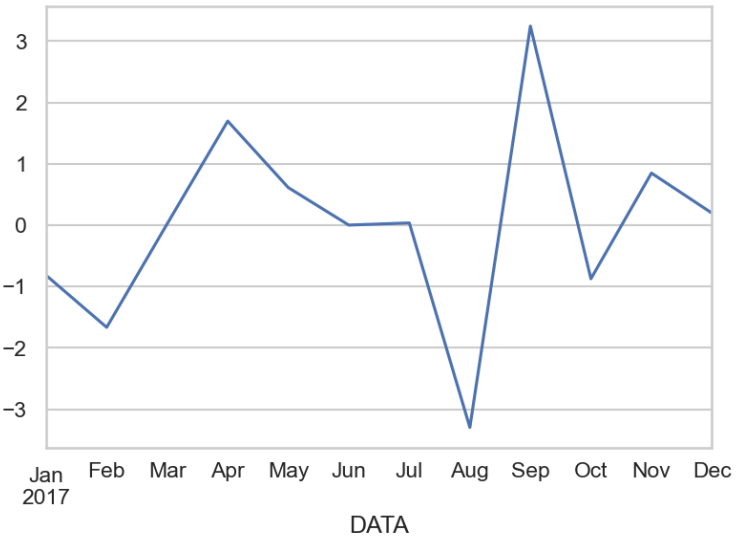
Analisando os valores mensais de MVI no gráfico da Figura 7 é possível identificar uma clara tendência de queda entre os anos de 2017 e 2020, com os dados apresentando flutuações abruptas. Nos anos subsequentes, entretanto, a série demonstra uma maior estabilidade, com variações menos expressivas, situando-se geralmente entre 5 e 15 MVIs por mês. Embora ainda existam flutuações, elas ocorrem em uma faixa mais contida. Além disso, é possível observar mudanças nos padrões de sazonalidade ao longo dos anos. A decomposição da série (Figura 8) revela um padrão sazonal característico, conforme ilustrado na Figura 9. Esse padrão evidencia um aumento nos MVIs no início do ano, seguido por uma redução até agosto, e um novo aumento no final do ano. No entanto, este comportamento sazonal não se mantém constante ao longo de toda a série histórica. Nos anos de pandemia de COVID-19, especificamente 2020 e 2021, a série exibiu maior estabilidade, com menores flutuações sazonais. Já em 2022, ano marcado pelo retorno das atividade normais, o pico de crimes migrou para o final do ano, e em 2023 o padrão sazonal retornou ao comportamento observado nos primeiros anos da série, com picos no início e final do ano. Essas alterações refletem as mudanças sociais ocasionadas pela pandemia, que influenciaram diretamente a dinâmica de fenômenos sociais, como a ocorrência de crimes.

Figura 8 – Decomposição da série histórica de MVI - modelo aditivo



Fonte: o autor (2024) ().

Figura 9 – Recorte do padrão de sazonalidade da série histórica de MVI

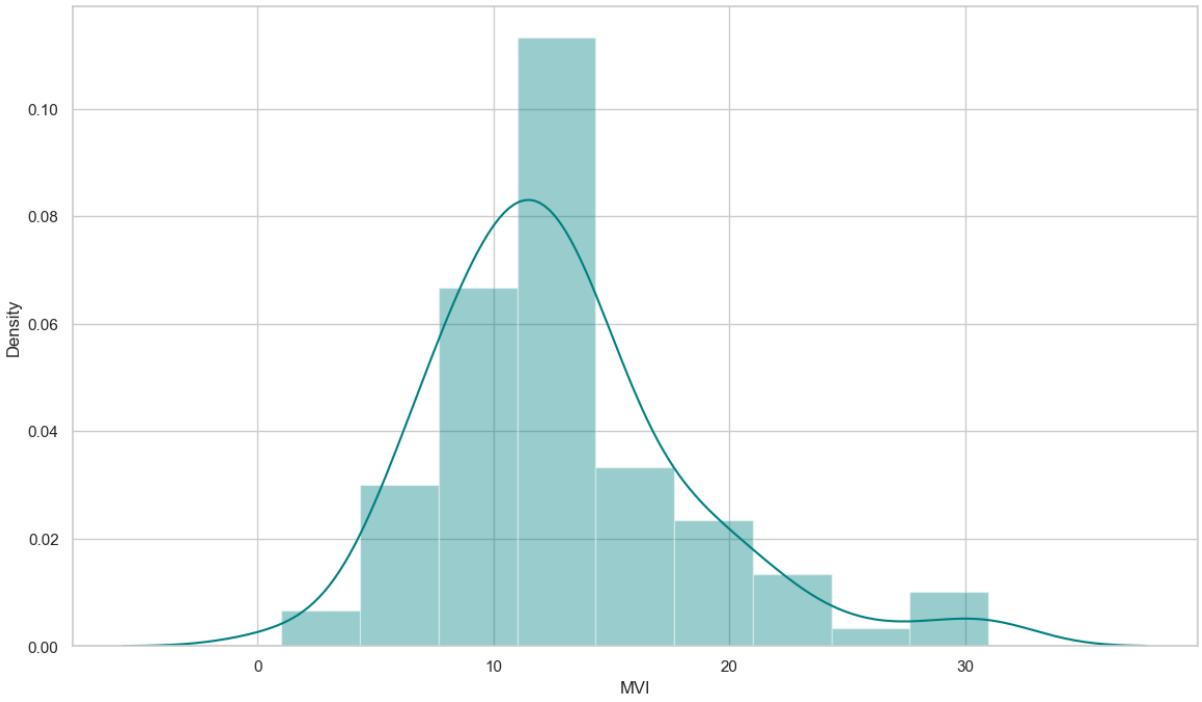


Fonte: o autor (2024) ().

Aplicando o teste Augmented Dickey-Fuller (ADF), e assumindo um nível de significância (α) de 0.05, obteve-se um p-value de 0.0353, indicando a estacionariedade dos dados. Obteve-se, também, um p-value de aproximadamente 0 no teste de Shapiro-Wilk, indicando a não normalidade dos dados além de um valor de 1.108910 na métrica *skewness* indicando uma distribuição de dados com alta assimetria positiva (skewed to the right), Figura 10. O gráfico de autocorrelação da Figura 11 apresenta uma autocorrelação progressivamente menor, indicando

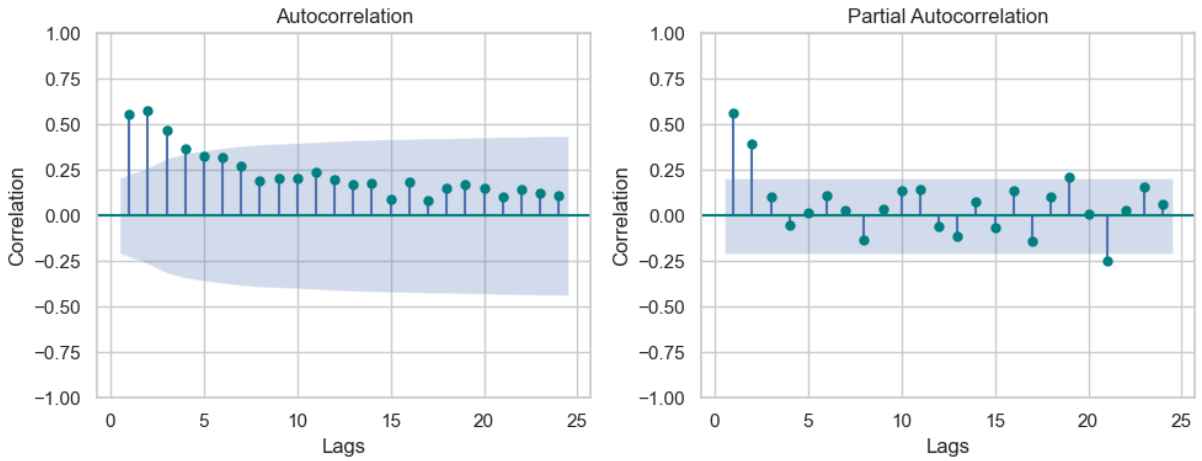
que os valores mais recentes têm maior influência sobre os valores de MVI, o que é razoável de se esperar dado a natureza dos dados. Em particular os lags de 1 a 4 podem ser considerados estatisticamente significativos, corroborando com o padrão sazonal em 3 ciclos ao longo do ano, identificado anteriormente. Pelo gráfico PACF (Figura 11) identificas como lags significativos os lags 1, 2, 19 e 21, sendo estes os lags são os mais relevantes para modelos de previsão nesse contexto. No entanto, considerando que o objetivo deste estudo é realizar previsões para um horizonte de 6 meses, os lags 1 e 2 não serão utilizados, uma vez que esses valores não estariam disponíveis no momento da previsão.

Figura 10 – Gráfico de distribuição dos valores mensais de MVI



Fonte: o autor (2024).

Figura 11 – Gráficos ACF (esquerda) e PACF (direita) da série de MVI

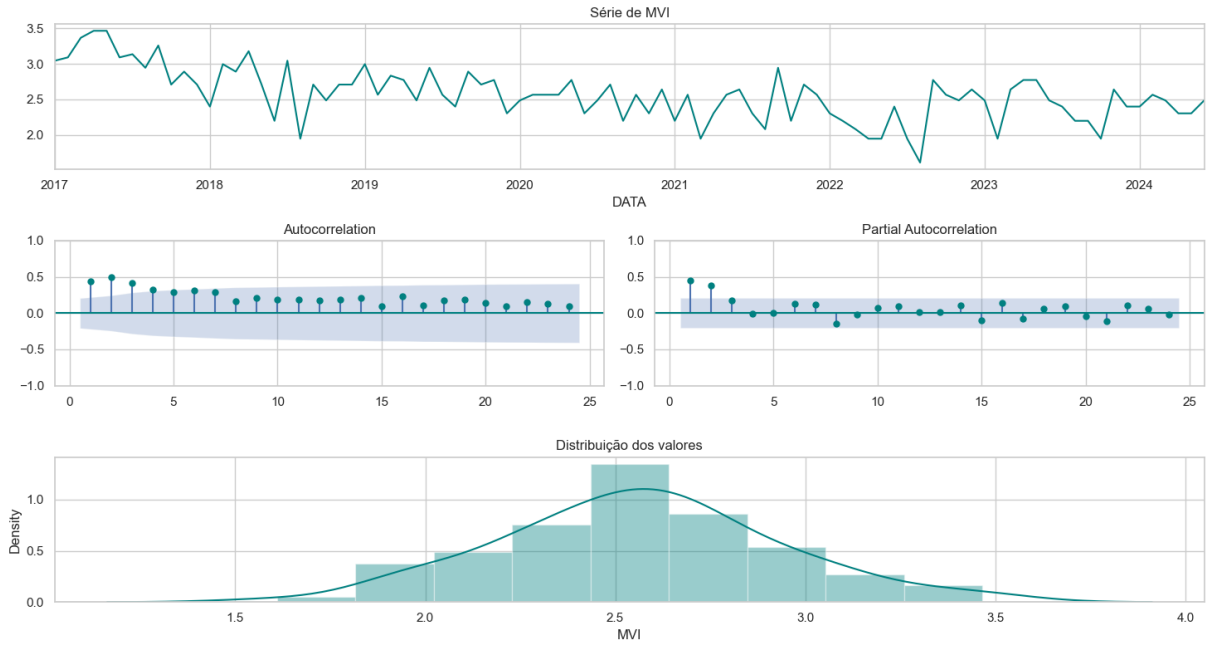


Fonte: o autor (2024).

Aplicando uma transformação logarítmica obtém-se uma série estacionária e normalmente distribuída (Figura 12) (p-values de 0.0379 e 0.4740 para os testes ADF e Shapiro-Wilk, respectivamente), sustentado ainda pela interpretação do gráfico de autocorrelação (ACF), o qual não apresenta um padrão de crescimento/decrescimento, conforme a Figura 12. De forma similar à série original, o gráfico ACF da série transformada apresenta oslags de 1 a 4 como estatisticamente significativos, enquanto o gráfico PACF mostra os lags 8 e 21 como significativos.

Apesar do prejuízo às informações e interpretabilidade dos dados, optou-se por aplicar a transformação logarítmica $x = log(x + 1)$ em todos os dados, tornando a série normalmente distribuída, além de remover as ocorrências de valores zeros nas variáveis explicativas, beneficiando a usabilidade dos dados nos modelos de machine learning. Após a transformação dos dados, identificou-se um outlier referente ao mês de fevereiro de 2022, período em que foi registrado apenas um homicídio, o menor valor de toda a série histórica. Esse resultado pode refletir mudanças comportamentais no contexto pós-pandemia, uma vez que fevereiro é tradicionalmente um mês de férias, e, em 2022, as restrições a aglomerações foram mais brandas. No entanto, para a utilização dos dados, o outlier foi substituído pela média simples entre o valor anterior e o posterior da série.

Figura 12 – Gráficos ACF, PACF e distribuição dos valores de MVI aplicando a transformação logarítmica

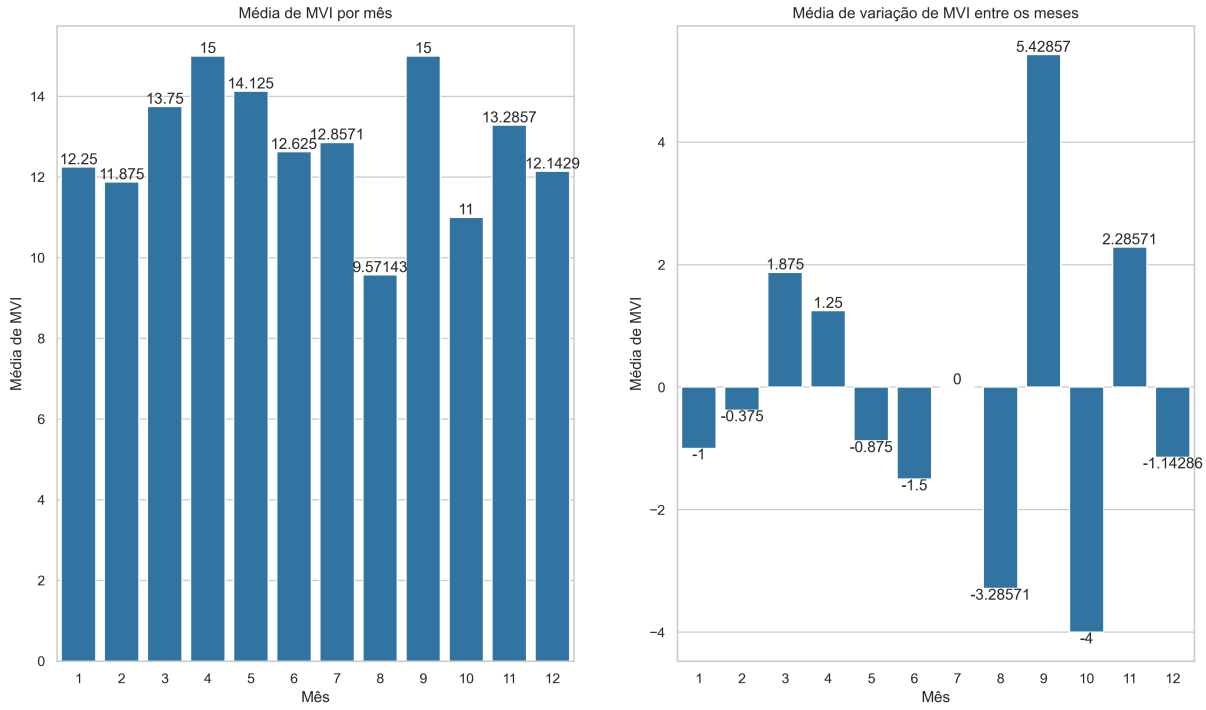


Fonte: o autor (2024).

A distribuição de MVI pelos meses do ano, Figura 13, mostra o padrão de sazonalidade nos casos de MVI descrito anteriormente nesta secção, com um "aquecimento"entre os meses de março e maio, e uma queda no mês agosto. Um ponto que chama atenção é o mês setembro, o qual apresenta a maior média de MVI logo após o mês com a menor média e anterior do mês

com a segunda menor média, resultando nas maiores variações entre meses e indicando uma possível influência de fatores sazonais específicos.

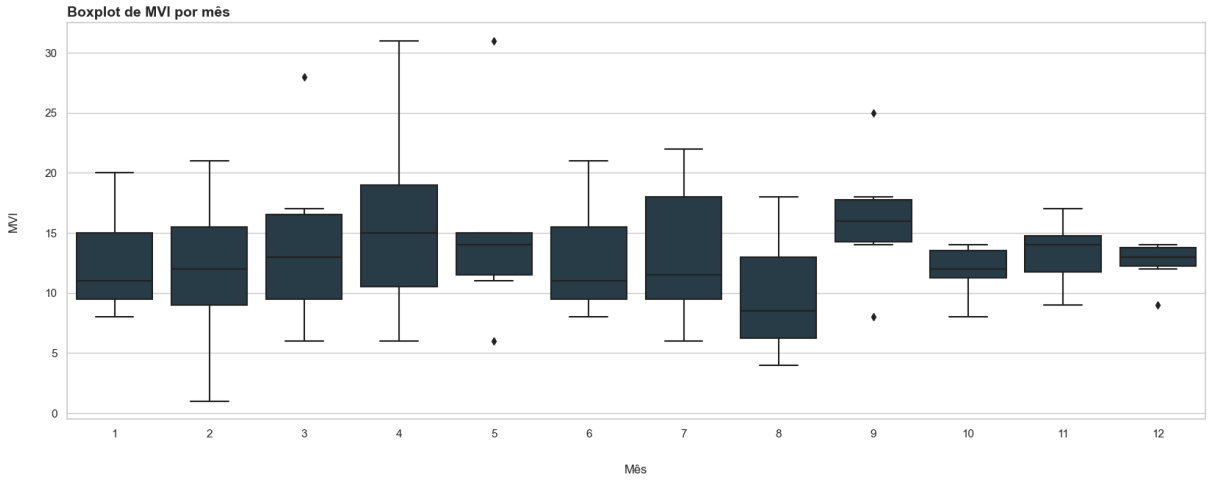
Figura 13 – Detalhamento dos números de MVI por mês



Fonte: o autor (2024).

Através do boxplot da Figura 14 é possível identificar uma maior constância nos números dos últimos quarto meses do ano, com alguns outliers, enquanto que nos demais meses é possível identificar uma variação muito maior. Em particular, os meses de fevereiro e abril destacam-se com as maiores variações, o que torna mais difícil realizar previsões precisas para esses períodos. Estes fatos refletem as mudanças nos padrões de sazonalidade ao longo dos anos citadas anteriormente, onde se teve, principalmente, uma mudança no comportamento dos casos de MVI nos primeiros meses do ano, causando uma maior variabilidade.

Figura 14 – Boxplot de MVI por mês



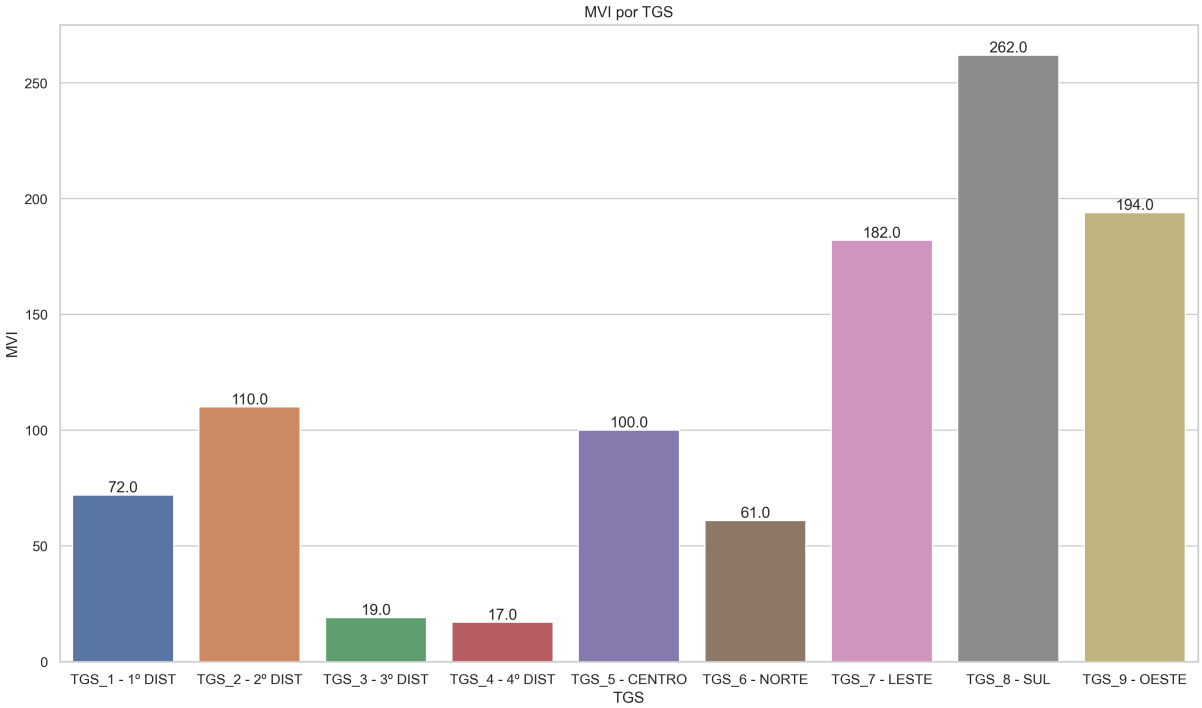
Fonte: o autor (2024).

A distribuição de MVI por TGS, Figura 15, mostra uma grande concentração dos crimes na zona urbana do município, em especial os TGSs 7, 8 e 9. A concentração de crimes nesses setores pode ser atribuída a diversos fatores, como a densidade populacional, renda familiar média, e possivelmente menor vigilância em comparação com áreas residenciais mais ao centro da cidade. Isso fica evidente ao analisar o mapa da calor de MVI no município (Figura 16), onde é possível identificar alguns pontos quentes que são atribuídos a certos bairros/localidades, os principais pontos identificados são:

- Ponto quente 1: Bairro José Carlos de Oliveira - TGS 9
- Ponto quente 2: Bairros Monte Bom Jesus / Centenário / Divinópolis - TGS 5
- Ponto quente 3: Parque 18 de Maio / Bairro Santa Rosa - TGS 8
- Ponto quente 4: Bairro Santa Rosa - TGS 8
- Ponto quente 5: Bairros Riachão / Salgado - TGS 7
- Ponto quente 6: Bairros Salgado / São João da Escócia - TGS 7
- Ponto quente 7: Bairro Cidade Jardim - TGS 7

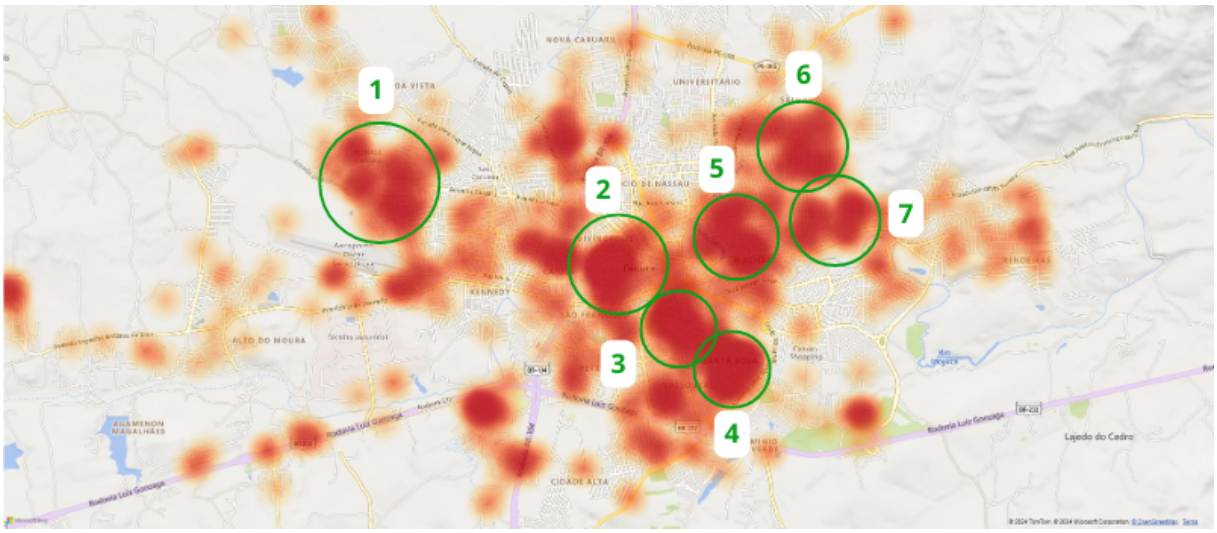
Com isso, o TGS 7 - Leste possui o maior número de pontos quentes com um total de 3 pontos quentes, enquanto o TGS 8 - possui o maior número de MVIs, distribuídos entre 2 pontos quentes.

Figura 15 – Detalhamento dos números de MVI por TGS



Fonte: o autor (2024).

Figura 16 – Mapa de calor dos dados de MVI



Fonte: o autor (2024).

3.4 PREPARAÇÃO DA BASE DE DADOS

Dado a natureza dos dados, algumas informações cruciais como a natureza jurídica, objeto utilizado, motivação, data, entre outros, são necessariamente informadas pela Polícia Civil, e, em geral, a base de dados possui uma pequena porção de valores em branco, portanto o principal foco nesta etapa foi a padronização dos dados, ondem foi feita a alteração manual de vários

dados com escritas divergentes, exemplo: bairro Nossa Senhora das Dores e Centro, ambas as nomenclaturas se referem ao mesmo bairro, no entanto ambas são comumente utilizadas.

A priori, foram descartadas algumas colunas da base de dados as quais não seriam relevantes para o estudo, como ID e nome do logradouro do fato. Além disso, foram removidas as informações de *Bairro de origem da vítima* e *Orientação sexual da vítima*, as quais, especificamente, apresentaram um grande percentual de valores em branco (acima de 60%) indicando uma falha na qualificação destas informações.

A coluna *Antecedentes criminais da vítima* apresenta 19,5% de valores em branco, inicialmente foi feita a classificação de Possui ou Não Possui Antecedente Criminais gerando uma coluna binária. Para contornar os dados em branco utilizou-se 2 metodologias: 1 - utilizou-se do algoritimo K-nearest neighbors (KNN) (k=26, acurácia = 0.587) para prever os valores não informados da classificação binária; 2 - os valores em branco foram considerados como Não Possui Antecedente Criminais. A informação sobre os antecedentes criminais da vítima são de extrema importância para a qualificação dos crimes, no entanto a proporção de valores em branco mostram uma negligência no preenchimento de tais dados, resultando em uma limitação no modelo de previsão.

Na tabela 1 um resumo das modificações realizadas.

O conjunto de dados utilizado contém informações de contagem, onde cada linha representa um caso de MVI. Para o tratamento das variáveis categóricas, foi aplicado o método de codificação *One Hot Encoding* (ou *Dummy Encoding*), que converteu cada categoria de uma feature em n colunas binárias, sendo n o número de categorias dessa feature. Posteriormente, os dados foram agrupados por mês, dado que este intervalo de tempo é comumente utilizado pelos agentes de segurança pública para a avaliação de resultados e definição de metas e estratégias, resultando em uma série temporal mensal.

3.4.1 Engenharia de features

Como sugerido por Almeida (ALMEIDA, 2023), serão utilizadas as seguintes técnicas para criação de novas features:

- **Observações passadas (lags)**: Valores registrados em períodos (meses) passados de cada variável, dependentes e independentes. Os lags foram selecionados de acordo com lags de interesse, (o lag 12 é comumente utilizado por se referir ao mesmo mês do ano anterior, uma sazonalidade conhecida empiricamente); de acordo com análise dos gráficos ACF e PACF descritas anteriormente; e de acordo com o modelo utilizado. Os lags utilizados foram: 6, 8, 12, 18, 19 e 24.

- **Estatísticas móveis (média, desvio padrão, valor mínimo, valor máximo)**: As estatísticas móveis são calculadas em uma janela de valores anteriores à cada observação. O tamanho da janela de valores influência na suavidade da curva da nova série criada: inter-

Tabela 1 – Resumo das modificações na base de dados por coluna

Coluna	**Modificação**
ID	Removido
Data	Utilizado para formular a série histórica, e definir o dia da semana, semana do mês, etc.
Horário	Agrupado em Turno (madrugada, manhã, tarde e noite)
Coordenadas	Utilizado para definir o bairro exato, de acordo com o plano diretor do município
Bairro	Substituído pelo bairro definido através das coordenadas
Logradouro	Removido
Natureza Jurídica	Sem alteração
Sexo da vítima	Sem alteração
Idade da vítima	Agrupado em faixas etárias
Objeto utilizado	Sem alteração
Motivação do crime	Agrupados por grupos de motivações
Antecedentes criminais da vítima	Previsão/substituição dos valores em branco e Substituído pela classificação booleana *Possui Antecedentes*
Bairro de residência da vítima	Removido
Orientação sexual da vítima	Removido
Consumo de substância pela vítima	Substituído pela classificação booleana *Envolvimento com Narcóticos*
Local do fato	Sem alteração
Cor da vítima	Sem alteração

valos maiores criam curvas mais suaves, enquanto intervalo menores criam curvas mais sensíveis. Os tamanhos de janelas utilizados foram: 6, 12, 18 e 24.

- **Médias móveis ponderadas exponencialmente**: Similar à média móvel, no entanto, define pesos maiores para valores mais recentes da série histórica de forma exponencial. Os valores de alph utilizados foram: 0.9, 0.8, 0.7 e 0.5.

- **Tendência e Sazonalidade**: Valores de tendência e sazonalidade obtidos através das decomposição da série de MVI, a partir do método aditivo.

3.5 MACHINE LEARNING

Conforme exposto na secção 2.3, algoritmos baseados em Random Forest e Gradient Boosting são amplamente considerados apropriados para a tarefa de previsão de crimes. Como isso, neste trabalho foram aplicados os algoritmos de machine learning Random Forest, da biblioteca SciKit-Learn, e Extreme Gradient Boosting, da biblioteca XGBoost, para realizar a regressão dos dados e prever o número de casos de MVI para os próximos seis meses.

Além da previsão, ambos os algoritmos foram utilizados para identificar as variáveis mais influentes no modelo. Os modelos foram treinados sobre a mesma base de dados, empregando técnicas como validação cruzada para garantir a robustez das previsões e o uso de Recursive Feature Elimination (RFE) para a seleção das variáveis mais relevantes. Cada algoritmo foi ajustado de acordo com suas especificidades, seguindo uma metodologia personalizada para otimizar o desempenho preditivo.

3.5.1 Seleção de Features

Para a seleção de features, foi utilizado o algoritmo de seleção recursiva de features (Recursive Feature Elimination ou RFE) da biblioteca SciKit-Learn para a seleção das features mais relevantes para os modelos de machine learning, através de testes em conjuntos de features cada vez menores, de forma recursiva, até chegar à quantidade determinada de features (SCIKIT-LEARN, 2024). O principal objetivo da utilização deste algoritmo é reduzir a quantidade total de features utilizadas pelo modelo final de machine learning, reduzindo consideravelmente o custo computacional para treinar tais modelos, além da possibilidade de otimizar seu desempenho pela eliminação de features ruins.

Após a etapa anterior obteve-se um dataset com um conjunto de 627 features X além da variável independente y. Semelhante à metodologia de MALTA (2022), foram testados subconjuntos de features com diferentes dimensões, selecionando o subconjunto que apresente o menor Erro Quadrático Médio (MSE), priorizando o subconjunto com menor dimensão, mesmo que este apresente um MSE até 5% superior ao melhor modelo.

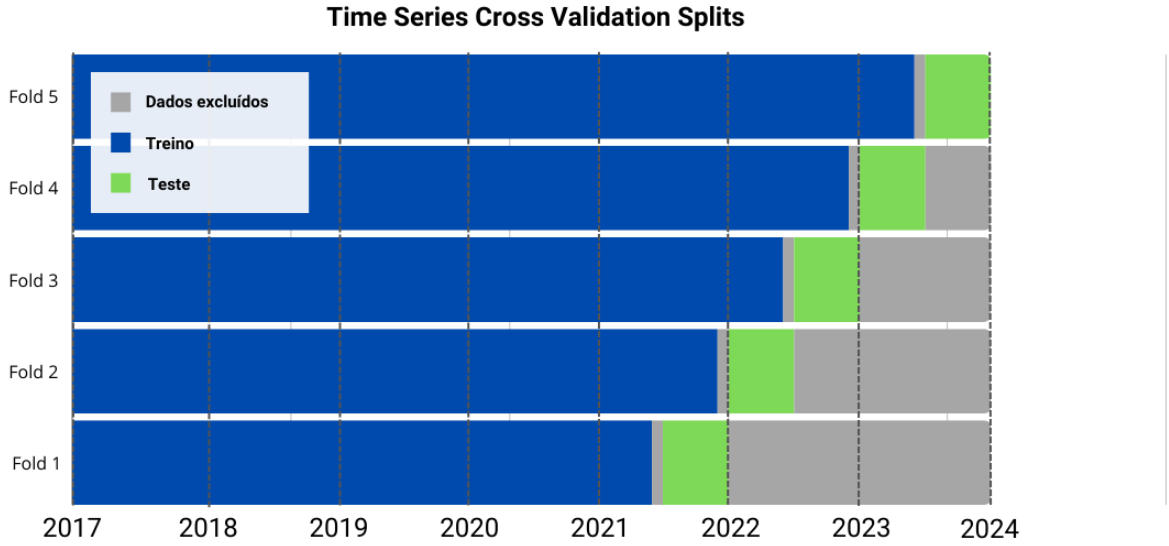
3.5.2 Validação cruzada

A validação cruzada (cross-validation ou CV) é uma técnica amplamente utilizada para validação de modelos de machine learning, e, ao contrário da validação simples, consiste em utilizar diferentes conjuntos de dados na etapa de validação (ARLOT; CELISSE, 2010), afim de minimizar a influência de qualquer viés nos dados de teste, e garantir que o modelo testado pode ser generalizado para novos dados (HASTIE; TIBSHIRANI; FRIEDMAN, 2001; BERGMEIR; HYNDMAN; KOO, 2018).

Neste trabalho, utilizou-se do método de validação cruzada para séries históricas *TimeSeriesSplit* da biblioteca SciKit-Learn. Este método se assemelha ao método mais comum de validação cruzada, o método K-fold, no entanto adaptado para na k-ésima divisão, retorna as

primeiras k dobras como conjunto de treino e a (k+1)-ésima dobra como conjunto de teste (SCIKIT-LEARN, 2024), conforme a Figura 17. Optou-se, também, por não utilizar o mês entre os conjuntos de treino e teste (Gap = 1 mês) no intuito de evitar overffiting e vieses nos conjuntos de teste. Como sugerido por Bergmeir, Hyndman e Koo (2018), após a validação cruzada verificou-se a correlação dos resíduos para cada um dos modelos de machine learning testado através do teste de Ljung–Box.

Figura 17 – Validação cruzada utilizada nos modelos de machine learning



Fonte: o autor (2024).

3.5.3 Ajuste dos hiperparâmetros

Na etapa de treinamento dos modelos, optou-se pela escolha dos hiperparâmetros através de otimização Bayesiana, utilizando o framework da biblioteca Optuna. Para cada modelo foi criado um estudo com 300 iterações tomando com objetivo minimizar o MSE gerado pelo modelo. Na tabela 2 os hiperparâmetros otimizados para cada modelo.

Tabela 2 – Lista de hiperparâmetros otimizados para cada modelo

XGBoost	**Random Forest**
learning rate	max depth
max depth	min samples split
subsample	min samples leaf
XGBoost	-
colsample bytree	-
min child weight	-

3.6 AVALIAÇÃO DOS MODELOS

Para avaliar os resultados das previsões dos modelos, foram utilizadas as métricas de desempenho listadas abaixo. Em seguida, ambos os modelos modelos foram comparados com um modelo de referência (modelo base), que utilizou a média simples dos valores mensais de MVI da série histórica para prever os seis meses subsequentes.

- Erro quadrático médio (MSE) (Principal)
- Erro absoluto médio (MAE)
- Erro absoluto percentual médio (MAPE)
- Raiz quadrada do erro quadrático médio (RMSE)
- R-quadrado (R^2)

Além disso, para a análise e interpretação das regras de decisão dos modelos de machine learning, utilizou-se o framework de valores aditivos de Shapley, através da biblioteca SHAP. A interpretabilidade de métodos de conjunção de árvores é importante porém pode ser difícil de conseguir. O framework SHAP minimiza esse problema, possibilitando a analise da influência global de local de cada variável nas previsões geradas pelo modelo. A contribuição de cada variável para o output do modelo é calculada de acordo com sua contribuição marginal em comparação com outras variáveis medidas em valores SHAP (ZHANG et al., 2022).

3.7 SOFTWARES E RECURSOS UTILIZADOS

Tabela 3 – Lista de softwares utilizados no trabalho

Software	**Descrição**
Python	Linguagem de programação
Pandas	Biblioteca voltada para o processamento e análise de dados em formas de dataframes
Scikit Learn	Biblioteca voltada para o machine learn, disponibilizando ferramentas para processamento de dados, criação e seleção de modelos, entre outros.
XGBoost	Biblioteca voltada para a implementação de algoritmos de Gradiente Boosting.
SHAP	Biblioteca voltada para a interpretabilidade de modelos de machine learning baseada no valores de Shapley
Optuna	Framework para a otmização de hiperparâmetros de modelos de machine learning
Matplotlib	Biblioteca voltada para a criação de visualizações gráficas
Jupyter Notebooks	Ferramenta para o desenvolvimento interativo de códigos

4 RESULTADOS COMPUTACIONAIS

Nesta capítulo são apresentados os modelos de machine learning avaliados além dos resultados obtidos. Os dados compreendem o período de janeiro de 2017 a junho de 2024, sendo o período de janeiro a junho de 2024 reservado para o teste dos modelos (test set), resultando em uma divisão 93:7 dos dados. Os códigos foram executados em um computador AMD Ryzen 5 PRO 5650G, 3.90 GHz, 16GB RAM.

4.1 SELEÇÃO DE FEATURES

No método de eliminação recursiva de features, para cada regressor foram testadas subconjuntos com 8 diferentes tamanhos (q) e calculado o erro quadrático médio (MSE) gerado pelo modelo base, sem nenhum ajuste de hiperparâmetros, utilizando a base de treino com validação cruzada. Os resultados obtidos estão dispostos no quadro 4.1. Vale ressaltar que os valores apresentados estão na escala logarítmica, em razão da transformação aplicada aos dados, descrita no capítulo anterior.

Quadro 4.1 – Erro quadrático médio obtido para cada modelo testado

Regressor	q=8	q=15	q=20	q=25	q=35	q=50	q=100	q=150
XGBoost	0.290	**0.230***	0.246	0.229	0.243	0.228**	0.233	0.232
Random Forest	**0.288***	0.326	0.318	0.294	0.287	0.289	0.286	0.277**

*Modelo selecionado. **Menor MSE.

Conforme a regra de decisão mencionada, para cada regressor foi selecionado o modelo com menor número de features, desde que o MSE deste modelo não ultrapasse o menor MSE por mais de 5%. Para o XGBoost, foi selecionado o modelo com 15 features, com um MSE de 0,230 ($0,23 < 1,05 * 0,228 = 0,2394$). Para o Random Forest, foi selecionado o modelo com 8 features, com um MSE de 0,288 ($0,288 < 1,05 * 0,277 = 0,290$).

4.2 OTIMIZAÇÃO E RESULTADOS DOS MODELOS

Após a otimização dos hiperparâmetros dos modelos, foram obtidos os valores para cada hiperparâmetro descritos nos quadros 4.2 e 4.3.

Quadro 4.2 – Ajuste dos hiperparâmetros do modelo XGBoost

	learning rate	max depth	subsample	colsample by tree	min child weight
XGBoost	0.0946	3	0.585	0.335	1

Quadro 4.3 – Ajuste dos hiperparâmetros do modelo Random Forest

	max depth	min samples split	min samples leaf
RF	1	9	10

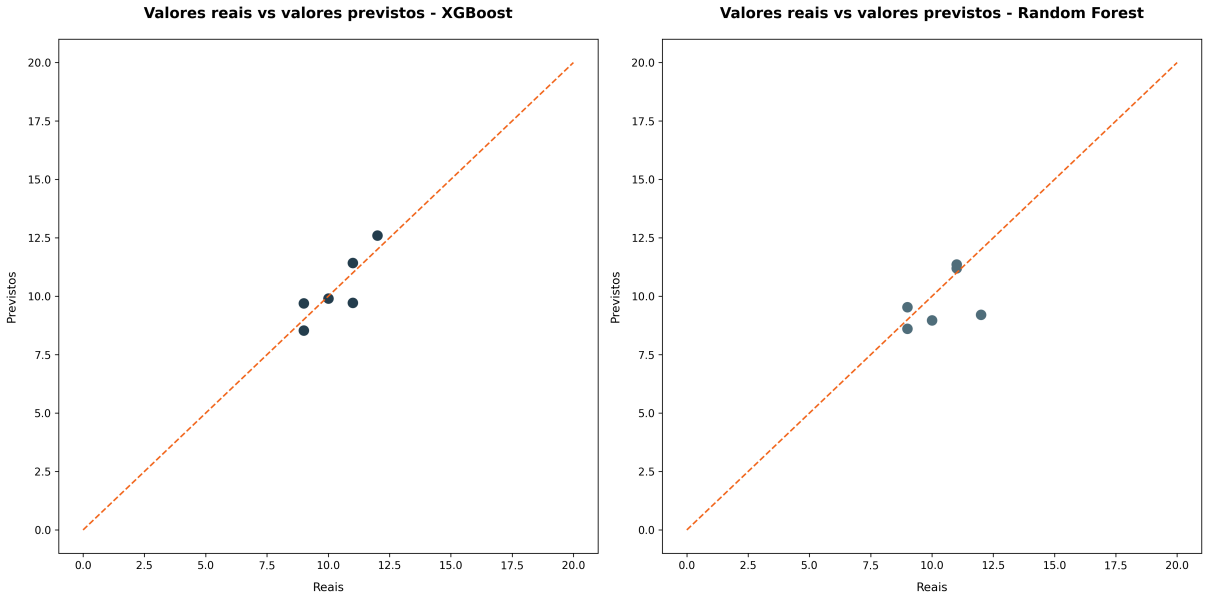
Os modelos ajustados foram aplicados para prever os números de MVI nos seis meses do conjunto de teste. As previsões geradas estão dispostas na Tabela 4, enquanto as métricas de avaliação são apresentadas na Tabela 5 (com valores na escala original).

Observa-se que o modelo baseado no algoritmo XGBoost produziu previsões mais próximas dos valores reais em todos os meses analisados, superando o desempenho do modelo Random Forest. Na figura 18 observa-se que os pontos gerados pelo modelo XGBoost (gráfico à esquerda) estão muito próximos da linha de identidade (onde os valores previstos seriam exatamente iguais aos valores reais), indicando que o modelo XGBoost conseguiu prever os valores de MVI com alta precisão, enquanto que para o modelo Random Forest, estes pontos estão mais dispersos, mostrando uma maior discrepância entre os valores previstos e os reais. Com isso, o XGBoost gerou um Erro Quadrático Médio (MSE) aproximadamente 2,6 vezes menor que o Random Forest. Além disso, pela ótica do valor total previsto para o período de 6 meses, o modelo XGBoost também se sobressai com uma estimativa quase exata.

Tabela 4 – Comparativo das previsões geradas pelos modelos com os valores reais de MVI

Mês	Valor real	XGBoost		Random Forest	
		Valor previsto	Erro %	Valor previsto	Erro %
jan / 2024	10	9.907845	-0.92%	8.424932	-10.30%
fev / 2024	12	12.600325	5.00%	9.053144	-23.30%
mar / 2024	11	11.421800	3.83%	11.294338	3.30%
abr / 2024	9	9.701710	7.80%	7.137709	-4.28%
mai / 2024	9	8.539045	-5.12%	9.608384	5.92%
jun / 2024	11	9.715205	-11.68%	10.637950	1.84%
Total	**62**	**61.88**	**-0.18%**	**58.88**	**-5.02%**

Figura 18 – Valores previstos vs valores reais



Fonte: o autor (2024).

Ao comparar ambos os modelos com o modelo de referência (Tabela 5), o qual utiliza a média simples dos valores mensais de MVI da série histórica como previsão para os próximos meses (ou seja, 13 MVIs para cada mês do conjunto de teste), ambos os modelos de machine learning apresentaram resultados superiores ao modelo mais simples, destacando o poder preditivo desses algoritmos. No entanto, em termos do coeficiente de determinação R^2, o Random Forest, assim como o modelo de referência, apresentou um resultado negativo, de -0,2931, indicando que tal modelo não se adequou bem aos dados ao falhar em capturar a variação dos valores reais. Por outro lado, o modelo XGBoost alcançou um R^2 de 0,6042, indicando que o mesmo foi capaz de explicar aproximadamente 60% da variação observada nos dados de teste. Esses resultados, juntamente com o menor Erro Quadrático Médio (MSE) do XGBoost, reforçam sua escolha como o modelo mais adequado, demonstrando maior capacidade de generalização e precisão nas previsões.

Tabela 5 – Métricas de avaliação obtidas em cada modelo

Modelo	**MSE**	**MAE**	**MAPE**	**RMSE**	**R^2**
XGBoost	0.4837	0.5936	5.7%	0.6955	0.6042
Random Forest	1.5804	0.8849	8.15%	1.2571	-0.2931
Modelo base	8.3333	2.6666	27.26%	2.8867	-5.818

Ao analisar os resíduos do modelo (Tabela 6, foi identificado um aumento no erro de predição para valores de MVI nos meses de maio e junho. Em maio, o erro pode ser explicado pela maior variação nos dados deste mês, conforme pode ser observado no boxplot da Figura 14, sendo maio o mês com o maior desvio padrão nos números de MVI. Observa-se o maior

erro do modelo no mês de junho, uma das possíveis causas é a variação nos 4 primeiros meses de junho da série histórica: em 2017 foram 21 MVIs reduzindo drasticamente para 8 MVIs em 2018, aumentando para 18 MVIs em 2019, e reduzindo novamente para 9 MVIs em 2020. O mês de junho é especialmente importante no contexto de Caruaru devido às tradicionais festas juninas, o maior evento do município, que atrai um grande número de visitantes. Durante todo o mês, as operações de segurança são intensificadas pelos órgãos responsáveis podendo impactar o números de MVI. Tais fatores podem influenciar a capacidade preditiva para o mês de junho, no entanto, pelo tamanho reduzido da amostra de dados de teste, entende-se que seria necessário mais valores para aprofundar tal análise.

Tabela 6 – Resíduos gerados pelo modelo XGBoost

Mês	Valor real	Previsto	Resíduo (absoluto)
jan / 2024	10	9.907845	0,092155
fev / 2024	12	12.600325	0,600325
mar / 2024	11	11.421800	0,4218
abr / 2024	9	9.701710	0,70171
mai / 2024	9	8.539045	0,460955
jun / 2024	11	9.715205	1,284795

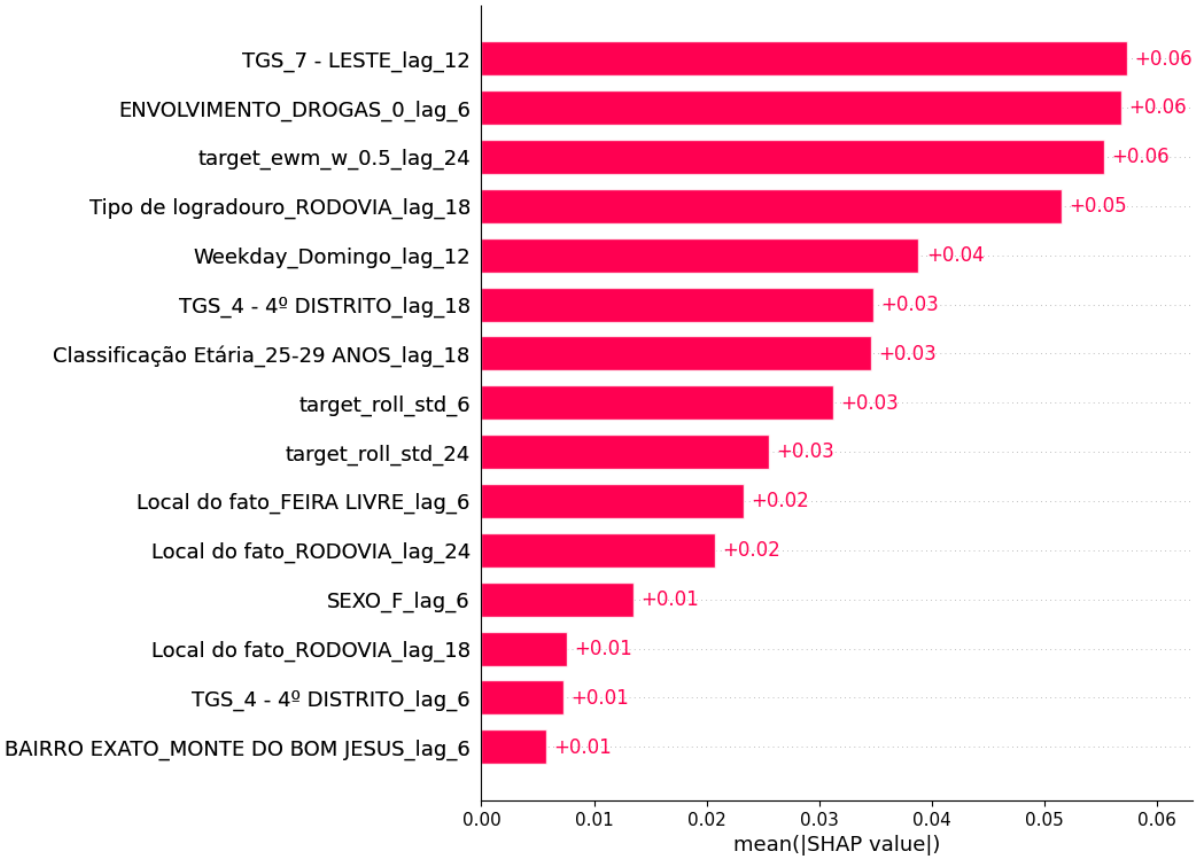
4.3 ANÁLISE DAS VARIÁVEIS INFLUENTES DO MODELO E VALORES SHAP

O framework SHAP pode ser utilizado para explicar o modelo tanto de forma global, ao mostrar a contribuição média de cada variável para os resultados, quanto de forma local, ao detalhar como cada previsão foi gerada (ZHANG et al., 2022). A Figura 19 apresenta o ranking dos valores globais SHAP para cada variável selecionada pelo modelo XGBoost na etapa de seleção de features (valores em escala logarítmica). Esses valores indicam a influência, em valores absolutos, de cada variável nos resultados previstos pelo modelo, podendo ser uma influência positiva ou negativa. As três primeiras posições incluem variáveis de naturezas distintas, o que evidencia a complexidade dos dados analisados: TGS 7 - Leste, uma variável geoespacial; "não possui envolvimento com drogas"(ENVOLVIMENTO_DROGAS_0), uma variável associada ao perfil do crime; e a média móvel ponderada exponencialmente com alpha de 0,5 (target_ewm_w_0.5), uma variável temporal. O modelo foi testado utilizando as duas abordagens de tratamento dos valores em branco da coluna de Antecedentes Criminais, mencionadas na seção 3.4 deste trabalho, no entanto em nenhuma das abordagens essa informação foi selecionada pelo modelo, e consequentemente gerando os mesmos resultados para ambas.

É comum que para se prever o número de casos de MVI de um dado mês ou avaliar o resultado obtido, utilize-se do resultado do mesmo mês do ano imediatamente anterior. Para

testar essa abordagem, os valores de MVI de 6, 12, 18 e 24 meses atrás (lags 6, 12, 18 e 24) foram incluídos no modelo como variáveis preditoras, no entanto nenhuma destas variáveis foi selecionada, indicando que a estratégia descrita não é um bom estimador para o número de MVI, e indicando também a existência de uma complexidade na relação entre o número de MVI e outros fatores.

Figura 19 – Valores SHAP globais (escala logarítmica)



Fonte: o autor (2024).

Além vaiáveis já mencionadas, o ranking inclui outras variáveis relevantes, como TGS_4 - 4º Distrito, médias móveis com janelas de 6 e 24 períodos (target_roll_std_6 e target_roll_std_24), e a classificação etária de 25 a 29 anos. A Tabela 7 apresenta a classificação de cada variável selecionada, enquanto a Tabela 8 mostra o somatório dos valores SHAP agrupados por categoria. Observa-se uma dominância das variáveis geoespaciais, que representam 8 das 15 variáveis selecionadas, com uma influência total de 0,2084. Esse valor é aproximadamente 38% superior à influência das variáveis temporais e cerca de 98% maior que a das variáveis relacionadas ao perfil do crime ou da vítima.

Na Figura 20 é possível analisar de forma detalhada a influência de cada variável nas previsões. No gráfico, para cada variável é mostrado seis pontos, correspondentes às seis previsões geradas pelo modelo. O eixo X representa o impacto que tal variável teve em cada previsão, enquanto a cor representa o valor da variável para tal ponto. A variável TGS 7 - Leste_lag12

Tabela 7 – Classificação das variáveis selecionadas pelo modelo XGBoost

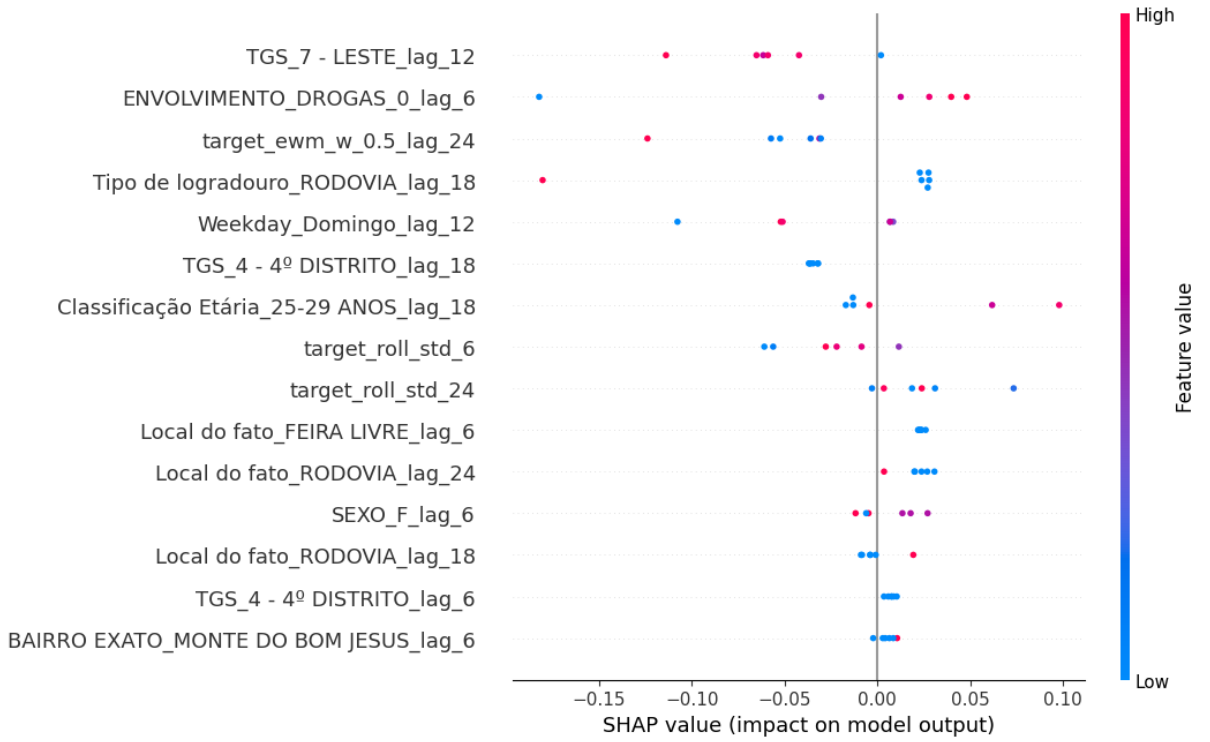
Variável	**Classificação**
TGS_7 - LESTE_lag_12	Geoespacial
ENVOLVIMENTO_DROGAS_0_lag_6	Perfil do crime / vítima
target_ewm_w_0.5_lag_24	Temporal
Tipo de logradouro_RODOVIA_lag_18	Geoespacial
Weekday_Domingo_lag_12	Temporal
TGS_4 - 4º DISTRITO_lag_18	Geoespacial
Classificação Etária_25-29 ANOS_lag_18	Perfil do crime / vítima
target_roll_std_6	Temporal
target_roll_std_24	Temporal
Local do fato_FEIRA LIVRE_lag_6	Geoespacial
Local do fato_RODOVIA_lag_24	Geoespacial
SEXO_F_lag_6	Perfil do crime / vítima
Local do fato_RODOVIA_lag_18	Geoespacial
TGS_4 - 4º DISTRITO_lag_6	Geoespacial
BAIRRO EXATO_MONTE DO BOM JESUS_lag_6	Geoespacial

Tabela 8 – Valores SHAP por cada classificação de variáveis

Classificação	**Valor SHAP**
Geoespacial	0,2084
Temporal	0,1508
Perfil do crime / vítima	0,1049

refere-se ao número de MVIs registrados no mesmo mês do ano anterior. Pelo heatmap da Figura 16, observa-se que o TGS 7 possui 3 pontos quentes de MVI, indicando que tal área possui um grande índice de violência. Conforme observado na figura, os maiores valores dessa variável (em vermelho) resultam em uma influência negativa nas previsões de MVI, evidenciando uma relação inversamente proporcional. Embora pareça contra-intuitivo, este comportamento pode indicar um ciclo ou padrão temporal específico de oscilação dos crimes nesta área, podendo indicar a efetividade de políticas de segurança ou ações de intervenção implementadas em resposta à situações de aumento de violência. No entanto, tal oscilação pode indicar, também, um comportamento de realocação dos crimes, sugerindo que a violência se desloca entre diferentes áreas da cidade conforme ações são implementadas.

Figura 20 – Distribuição dos valores SHAP de cada variável para cada previsão (escala logarítmica)



Fonte: o autor (2024).

A mesma relação é observada na variável "target_ewm_w_0.5_lag_24", que representa o valor da e a média móvel ponderada exponencialmente (EWMA) com alpha de 0,5 24 meses atrás. Um valor elevado dessa variável tende a reduzir as previsões do modelo, sugerindo uma tendência de queda nos números de MVI ao longo da série histórica. Em contraste com o mencionado anteriormente sobre a utilização dos valores de MVI de meses atrás como variáveis preditoras, a EWMA se mostra um estimador melhor para o número de MVI. Isso pode ser atribuído à característica da EWMA de dar maior peso aos valores mais recentes, o que é particularmente relevante em séries temporais com tendência, como é o caso da série histórica analisada neste estudo, conforme mostrado na Figura 5.

Por outro lado, para a variável "ENVOLVIMENTO_DROGAS_0_lag_6", a qual refere-se ao número de MVIs não relacionados a entorpecentes seis meses atrás, observa-se um impacto positivo nas previsões geradas quando o valor registrado é alto, caracterizando uma relação proporcional, e sugerindo um padrão temporal em que meses que registram um menor índice de mortes envolvendo entorpecentes são seguidos por um aumento no quantitativo geral de MVI. Tal comportamento pode ser explicado por fatores como a sazonalidade dos crimes relacionados ao tráfico de drogas, ou ainda por dinâmicas retomada de atividades criminosas após um período de maior repressão.

De maneira geral, observa-se uma oscilação da influência gerada por valores semelhantes de uma mesma variável, indicando padrões complexos de interações entre as variáveis selecio-

nadas. Além disso, a limitada quantidade de dados de teste contribui para aumentar a incerteza na interpretação das decisões do modelo de machine learning e, com isso, a interpretação dos padrões pode ser influenciada pelas variações locais e ruídos, dificultando a generalização das conclusões e a identificação de padrões robustos.

5 CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho teve como objetivo desenvolver um modelo capaz de prever os números do indicador de Mortes Violentas Intencionais (MVI) no município de Caruaru, além de identificar os fatores que influenciam esse indicador. Para isso, foram aplicadas técnicas de tratamento de dados e machine learning em uma base de dados detalhada, composta por casos passados de MVI registrados pela Secretaria de Ordem Pública do município.

Foi desenvolvido um modelo de regressão com machine learning baseado em Gradient Boosting, o qual utiliza valores passados (lags) das variáveis registradas para prever casos futuros e identificar padrões subjacentes. Ao comparar os resultados desse modelo com os obtidos por outro modelo, baseado no algoritmo Random Forest, constatou-se que o Gradient Boosting apresentou melhor desempenho para o problema em questão, corroborando com os resultados obtidos por Zhang et al. (2022), Djon et al. (2023), e Khan, Ali e Alharbi (2022).

A avaliação da eficácia do modelo XGBoost desenvolvido revelou previsões para os primeiros seis meses de 2024 com um erro absoluto percentual médio (MAPE) de 5,7% e um coeficiente de determinação (R^2) de 0,6042. Esses resultados indicam que o modelo foi capaz de prever os números de MVI com uma precisão satisfatória, além de explicar aproximadamente 60% da variação observada nos dados de teste. Ademais, ao somar as previsões para os seis meses, o total previsto de 61,88 MVIs foi muito próximo do valor real observado de 62, resultando em um erro de apenas 0,18%, o que demonstra a robustez do modelo desenvolvido.

Por meio de um algoritmo de seleção recursiva de variáveis explicativas, foram selecionadas 15 variáveis: 8 relacionadas a informações geoespaciais dos casos de MVI, 4 a dados temporais, e 3 ao perfil do crime ou da vítima. As variáveis geoespaciais apresentaram uma influência (valor SHAP) aproximadamente 38% maior em comparação com as variáveis temporais e cerca de 98% superior em relação às variáveis de perfil do crime ou da vítima. Esses resultados destacam a predominância das variáveis geoespaciais e temporais sobre as de perfil do crime ou da vítima, sem, contudo, descartar a relevância destas últimas.

Através da análise dos valores SHAP individuais de cada variável, observa-se uma relação inversamente proporcional entre TGS 7 - Leste_lag_12 e os valores previstos de MVI, indicando um padrão temporal específico de oscilação dos crimes nesta área ou ainda uma realocação dos crimes, sugerindo que a violência se desloca entre diferentes áreas da cidade conforme ações são implementadas. Ainda, observa-se uma relação proporcional MVIs não relacionados a entorpecentes de meses atrás com os valores previstos, indicando uma sazonalidade dos crimes relacionados ao tráfico de drogas. De maneira geral, os resultados indicam características sazonais, indicando a falta de medidas que ataquem as raízes dos problemas, permitindo a reincidência de crimes com as mesmas características. Esses resultados reforçam a necessidade de ações policiais direcionadas à repressão e desarticulação de grupos criminosos na região, visando evitar tanto a reincidência quanto a migração dos delitos para outras áreas.

No entanto, foi identificada uma limitação na análise das variáveis explicativas devido à quantidade reduzida de dados de teste, o que torna a análise mais suscetível a variações locais e ruídos. Portanto, como sugestão para trabalhos futuros:

- Explorar diferentes proporções na divisão entre dados de treino e teste, ampliando a amostra de dados de teste, ao custo de diminuir a precisão do modelo.
- Utilizar modelos de classificação para a interpretação para definir a influência de cada variável nos números de MVI, complementando a base de dados com dados socioeconômicos da população do município de Caruaru, além de dados geográficos.

REFERÊNCIAS

ALDADO, M. R. **POLÍTICAS PÚBLICAS DE SEGURANÇA: UMA ANÁLISE DO OBSERVATÓRIO MUNICIPAL DE SEGURANÇA PÚBLICA DE PELOTAS**. 2021.

ALMEIDA, P. **Store Item Demand Forecasting**. [S.l.]: GitHub, 2023. https://github.com/allmeidaapedro/Store-Item-Demand-Forecasting/blob/main/README.md.

ALVES, L. G. A. et al. Distance to the scaling law: A useful approach for unveiling relationships between crime and urban metrics. **PLOS ONE**, Public Library of Science, v. 8, n. 8, p. 1–8, 08 2013. Disponível em: https://doi.org/10.1371/journal.pone.0069580.

ALVES, L. G. A.; RIBEIRO, H. V.; RODRIGUES, F. A. Crime prediction through urban metrics and statistical learning. **Physica A**, Elsevier BV, v. 505, p. 435–443, set. 2018.

ARAUJO, A. de. **Predspot: Predicting Crime Hotspots with Machine Learning**. Tese (Doutorado) — Universidade Federal do Rio Grande do Norte - UFRN, 09 2019.

ARLOT, S.; CELISSE, A. A survey of cross-validation procedures for model selection. **Statistics Surveys**, Amer. Statist. Assoc., the Bernoulli Soc., the Inst. Math. Statist., and the Statist. Soc. Canada, v. 4, n. none, p. 40 – 79, 2010. Disponível em: https://doi.org/10.1214/09-SS054.

AZIZ, R. M.; SHARMA, P.; HUSSAIN, A. Machine learning algorithms for crime prediction under indian penal code. **Ann. Data Sci.**, Springer Science and Business Media LLC, v. 11, n. 1, p. 379–410, fev. 2024.

BECKER, G. S. Crime and punishment: An economic approach. In: ____. [S.l.]: NBER, 1974. cap. 2, p. 1–54.

BELYADI, H.; HAGHIGHAT, A. Chapter 5 - supervised learning. In: BELYADI, H.; HAGHIGHAT, A. (Ed.). **Machine Learning Guide for Oil and Gas Using Python**. Gulf Professional Publishing, 2021. p. 169–295. ISBN 978-0-12-821929-4. Disponível em: https://www.sciencedirect.com/science/article/pii/B9780128219294000044.

BERGMEIR, C.; HYNDMAN, R. J.; KOO, B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. **Computational Statistics and Data Analysis**, v. 120, p. 70–83, 2018. ISSN 0167-9473. Disponível em: https://www.sciencedirect.com/science/article/pii/S0167947317302384.

BOGOMOLOV, A. et al. Once upon a crime. In: **Proceedings of the 16th International Conference on Multimodal Interaction**. New York, NY, USA: ACM, 2014.

BORGES, J. et al. Feature engineering for crime hotspot detection. In: **2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)**. [S.l.: s.n.], 2017. p. 1–8.

CARUARU. Decreto nº 025, de 04 de maio de 2017. institui, no âmbito municipal de caruaru, o comitê permanente municipal juntos pela segurança. **Diário Oficial [da] Prefeitura de Caruaru**, Brasília, DF, 2017. Disponível em: https://caruaru.pe.gov.br/decreto-no-025-de-04-de-maio-de-2017/.

CERQUEIRA, D. "custo de bem-estar da violência e criminalidade no brasil. **Anuário Brasileiro de Segurança Pública 2017**, Fórum Brasileiro de Segurança Pública, 2017.

COSTA, M. R. d. A violência urbana é particularidade da sociedade brasileira? **São Paulo em Perspect.**, FapUNIFESP (SciELO), v. 13, n. 4, p. 3–12, dez. 1999.

CUNHA, H. M. da. **METODOLOGIA PARA ANÁLISE DE INDICADORES CHAVE DE DESEMPENHO RELACIONADOS À SEGURANÇA PÚBLICA EM MUNICÍPIOS**. Monografia (TCC - Bacharel em Engenharia de Produção) — UNIVERSIDADE FEDERAL DE PERNAMBUCO, 2022.

DJON, D. et al. **A Comparative Analysis of Multiple Methods for Predicting a Specific Type of Crime in the City of Chicago**. 2023. Disponível em: https://arxiv.org/abs/2304.13464.

DURANTE, M. et al. Modernização dos modelos de policiamento: uso das estatísticas na segurança pública para além da análise dos registros de ocorrência na polícia (df – 2015 a 2018). **Estatísticas de segurança pública: produção e uso de dados criminais no Brasil**, Fórum Brasileiro de Segurança Pública, 2022.

FURTADO, V. Inteligência artificial na segurança pública: conceitos, perspectivas e desafios. **Estatísticas de segurança pública: produção e uso de dados criminais no Brasil**, Fórum Brasileiro de Segurança Pública, 2022.

FóRUM BRASILEIRO DE SEGURANçA PúBLICA. **Anuário Brasileiro de Segurança Pública 2024**. 2024. Disponível em: https://publicacoes.forumseguranca.org.br/handle/123456789/253. Acesso em: 04 out. 2024.

GORDON, M. B. et al. Crime and punishment: the economic burden of impunity. **Eur. Phys. J. B**, Springer Science and Business Media LLC, v. 68, n. 1, p. 133–144, mar. 2009.

GUITARRARA, P. **Tipos de Violência**. Brasil Escola, 2024. Acesso em: 06 out. 2024. Disponível em: https://brasilescola.uol.com.br/sociologia/tipos-de-violencia.htm.

GUITARRARA, P. **Violência urbana**. Brasil Escola, 2024. Acesso em: 06 out. 2024. Disponível em: https://brasilescola.uol.com.br/geografia/violencia-urbana.htm.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. Springer, 2001. (Springer series in statistics). ISBN 9780387952840. Disponível em: https://books.google.com.br/books?id=VRzITwgNV2UC.

IBGE. **Portal do IBGE**. 2024. Disponível em: https://www.ibge.gov.br/.

JAITMAN, L. et al. **The costs of crime and violence: New evidence and insights in Latin America and the Caribbean**. [S.l.], 2017.

KHAN, M.; ALI, A.; ALHARBI, Y. Predicting and preventing crime: A crime prediction model using san francisco crime data by classification techniques. **Complexity**, Wiley, v. 2022, n. 1, p. 1–13, jan. 2022.

KIM, S. et al. Crime analysis through machine learning. In: **2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)**. [S.l.]: IEEE, 2018.

KUROKI, M. Crime victimization and subjective well-being: Evidence from happiness data. **J. Happiness Stud.**, Springer Science and Business Media LLC, v. 14, n. 3, p. 783–794, jun. 2013.

KWON, E.; JUNG, S.; LEE, J. Artificial neural network model development to predict theft types in consideration of environmental factors. **ISPRS Int. J. Geoinf.**, MDPI AG, v. 10, n. 2, p. 99, fev. 2021.

LIRA, P.; CABALLERO, B.; NASCIMENTO, D. C. do. Informação qualificada a partir de estatísticas criminais oficiais: avanços e desafios nacionais e a experiência do espírito santo. **Estatísticas de segurança pública: produção e uso de dados criminais no Brasil**, Fórum Brasileiro de Segurança Pública, 2022.

LYRA RAQUEL E PINHEIRO, R. **PARA CARUARU SEGUIR EM FRENTE Uma Construção Coletiva**. 2020. https://divulgacandcontas.tse.jus.br/candidaturas/oficial/2020/PE/23817/426/candidatos/384682/5_1601006690919.pdf. Acesso em: 2024-08-29.

MAHESH, B. Machine learning algorithms -a review. **International Journal of Science and Research (IJSR)**, v. 9, 01 2019.

MALTA, G. **An interpretable machine learning approach for predicting sleep quality in three temporal waves throughout the COVID-19 pandemic**. Monografia (TCC - Bachelor in Computer Science) — Universidade Federal do Rio Grande do Sul, 2022.

MJSP. **Dados Nacionais de Segurança Pública**. Ministério da Justiça e Segurança Pública, s.d. Acesso em: 07 out. 2024. Disponível em: https://www.gov.br/mj/pt-br/assuntos/sua-seguranca/seguranca-publica/estatistica.

PARENTE, R. V. R. M. **Reducing crime could significantly boost investment, productivity, and GDP growth in Latin America and the Caribbean**. International Monetary Fund (IMF), 2023. Acesso em: 07 out. 2024. Disponível em: https://www.imf.org/en/Blogs/Articles/2023/12/18/latin-america-can-boost-economic-growth-by-reducing-crime.

RABBANI, N. et al. Applications of machine learning in routine laboratory medicine: Current state and future directions. **Clinical Biochemistry**, v. 103, p. 1–7, 2022. ISSN 0009-9120. Disponível em: https://www.sciencedirect.com/science/article/pii/S0009912022000595.

SAHA, D.; MANICKAVASAGAN, A. Machine learning techniques for analysis of hyperspectral images to determine quality of food products: A review. **Current Research in Food Science**, v. 4, p. 28–44, 2021. ISSN 2665-9271. Disponível em: https://www.sciencedirect.com/science/article/pii/S2665927121000034.

SANDUNIL, K. et al. Effects of tuning decision trees in random forest regression on predicting porosity of a hydrocarbon reservoir. a case study: volve oil field, north sea. **Energy Adv.**, RSC, v. 3, p. 2335–2347, 2024. Disponível em: http://dx.doi.org/10.1039/D4YA00313F.

SAURET, G. V. Vidas salvas: metodologias de cálculo aplicadas ao pacto pela vida. **Estatísticas de segurança pública: produção e uso de dados criminais no Brasil**, Fórum Brasileiro de Segurança Pública, 2022.

SCIKIT-LEARN. **SKlearn Documentation**. 2024. https://scikit-learn.org/stable/. Acesso em: 2024-08-29.

SDS. **Pacto pela Vida completa 14 anos de êxito na redução da violência em PE**. Governo do Estado de Pernambuco, 2021. Acesso em: 05 out. 2024. Disponível em: https://www.sds.pe.gov.br/noticias/11337-pacto-pela-vida-completa-14-anos-de-exitos-na-reducao-da-violencia-em-pernambuco.

SDS, S. de Defesa Social de P. **Estatísticas**. 2024. Acesso em: 2024-08-29. Disponível em: https://www.sds.pe.gov.br/estatisticas.

SHINDE, P. P.; SHAH, S. A review of machine learning and deep learning applications. In: **2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)**. [S.l.]: IEEE, 2018.

WANG, W.; CHAKRABORTY, G.; CHAKRABORTY, B. Predicting the risk of chronic kidney disease (ckd) using machine learning algorithm. **Applied Sciences**, v. 11, p. 202, 12 2020.

WELLE, D. **Brasil lidera ranking de homicídios no mundo, mostra estudo da ONU**. 2023. https://www.terra.com.br/noticias/brasil/brasil-lidera-ranking-de-homicidios-no-mundo-mostra-estudo-da-onu,66fce2c84af2fa9e6961eab566b6a9335rz1ku28.html#:~:text=O%20Brasil%20lidera%20o%20ranking,4%25%20deles%20ocorreram%20no%20Brasil. Acesso em: 04 out. 2024.

ZHANG, X. et al. Interpretable machine learning models for crime prediction. **Computers, Environment and Urban Systems**, v. 94, p. 101789, 2022. ISSN 0198-9715. Disponível em: https://www.sciencedirect.com/science/article/pii/S0198971522000333.