UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

LEONARDO OSTJEN COUTO

Identificação de Padrões de Crime: Uma Abordagem Data-Driven para o Roubo de Celulares no RS

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação

Orientador: Prof. Dr. Renata Galante

Porto Alegre
2024

AGRADECIMENTOS

A minha família, verdadeiro pilar que proporcionou toda a estrutura e suporte necessários para que eu pudesse me dedicar plenamente e aproveitar cada instante deste longo período de graduação

Aos excelentes professores que estiveram presentes nesta jornada, em especial, a Professora Renata Galante, minha orientadora e o Professor Matheus Santos, que tornou-se um grande mentor durante este período.

Aos meus amigos, e em particular ao meu grande amigo João Vitor, agradeço pela amizade constante e pelo apoio incansável ao longo desta jornada. Sua presença foi uma fonte de força e ânimo em momentos desafiadores.

Ao Tenente Coronel Roberto Donato, da Brigada Militar do Rio Grande do Sul, pela sua colaboração e presteza em ajudar na compreensão e desenvolvimento deste trabalho.

A todos que, de alguma forma, contribuíram para este trabalho e para minha formação, meu mais sincero obrigado. Este é um momento especial compartilhado com cada um de vocês, e agradeço por fazerem parte desta conquista.

RESUMO

Este trabalho investiga o fenômeno do roubo de celulares no estado do Rio Grande do Sul, adotando uma abordagem baseada em dados para identificar padrões de crime. A análise faz uso de diversas técnicas de ciência de dados, incluindo aprendizado de máquina, visualização de dados e métodos estatísticos para identificar e compreender os padrões dentro desse tipo de delito no estado do Rio Grande do Sul.

Ao empregar essas técnicas, a análise busca identificar padrões comportamentais e geográficos associados ao roubo de celulares, um crime que, em virtude de sua crescente incidência, evoluiu para uma preocupação expressiva no contexto da segurança pública. Graças a Lei de Acesso à Informação, foi possível obter dados necessários para esta pesquisa com a Brigada Militar do RS.

O estudo visa compreender a dinâmica desse crime, buscando não apenas analisar incidentes isolados, mas também compreender o panorama geral. Isso envolve uma análise abrangente da demografia tanto das vítimas quanto dos criminosos, considerando variáveis como idade, gênero e localização geográfica.

Palavras-chave: Ciência de dados. dados governamentais. visualização de dados. aprendizado de máquina.

Identifying Crime Patterns: A data-driven approach to cellphone theft in Rio Grande Do Sul, Brazil

ABSTRACT

This study investigates the phenomenon of mobile phone theft in the state of Rio Grande do Sul, adopting a data-driven approach to identify crime patterns. The analysis utilizes various data science techniques, including machine learning, data visualization, and statistical methods, to identify and comprehend patterns within this type of crime in the state of Rio Grande do Sul.

By employing these techniques, the analysis aims to identify behavioral and geographical patterns associated with mobile phone theft, a crime that, due to its increasing incidence, has evolved into a significant concern in the context of public security. Thanks to the Brazilian Access to Information Law, necessary data for this research were obtained from the state police.

The study seeks to understand the dynamics of this crime, aiming not only to analyze isolated incidents but also to comprehend the bigger picture. This involves a comprehensive analysis of the demographics of both victims and criminals, considering variables such as age, gender, and geographical location.

Keywords: Public Data. Data Science.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
RS	Rio Grande do Sul
CRPO	Comando Regional de Polícia Ostensiva Central
IDH	Índice de Desenvolvimento Humano
POA	Porto Alegre

SUMÁRIO

1 INTRODUÇÃO

No panorama atual de segurança pública, o crescente índice de roubos de celulares desponta como um desafio de extrema importância, demandando abordagens inovadoras para lidar com este problema que se tornou cada vez mais prevalente na questão da segurança pública. O estado do Rio Grande do Sul não é imune a esse cenário, enfrentando uma escalada na incidência de roubos de dispositivos móveis, o que impacta diretamente a sensação de segurança da população.

Dada a gravidade do problema, percebe-se a urgência em encontrar soluções eficientes para enfrentar o desafio crescente dos roubos de celulares. Neste contexto, a Ciência de Dados emerge como uma ferramenta poderosa para a análise aprofundada de dados relacionados a crimes, oferecendo *insights* cruciais para o entendimento do problema de uma forma generalizada e sistemática, assim podendo desenvolver estratégias mais eficazes no combate a essas ocorrências.

O objetivo deste trabalho é desenvolver e aplicar abordagens baseadas em dados para identificar padrões e ter um compreendimento geral dos incidentes de roubo de celulares no estado do Rio Grande do Sul, potencialmente podendo contribuir para a formulação de políticas públicas mais efetivas e o aprimoramento das ações de segurança. Para atingir esse propósito, foram empregadas técnicas de processamento de dados, análise estatística, visualização de dados e algoritmos de aprendizado de máquina.

A estrutura organizada em capítulos visa proporcionar uma compreensão abrangente do problema. O Capítulo 2 aborda a fundamentação teórica, explorando conceitos essenciais de Ciência de Dados, Aprendizado de Máquina e suas aplicações específicas na segurança pública. No Capítulo 3, são apresentados os trabalhos relacionados, contextualizando a pesquisa dentro do panorama acadêmico atual. O Capítulo 4 se dedica à análise demográfica, oferecendo *insights* sobre os locais e perfis mais suscetíveis aos roubos de celulares no Rio Grande do Sul. A metodologia adotada para a identificação de padrões é detalhada no Capítulo 5, delineando as etapas do processo. Os resultados dos experimentos e sua análise são apresentados no Capítulo 6, destacando os principais achados e conclusões derivadas da aplicação das técnicas de Ciência de Dados e Aprendizado de Máquina. Por fim, no Capítulo 7, discutiremos as implicações práticas dos resultados, fornecendo uma conclusão abrangente e sugerindo possíveis direções para futuras pesquisas, consolidando assim o presente trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, são abordados os princípios e instrumentos empregados no desenvolvimento deste trabalho. Para facilitar o entendimento, o capítulo foi dividido entre conceitos e tecnologias, representados nas seções 2.1 e 2.2.

2.1 Conceitos

Nesta seção, são apresentados os conceitos teóricos utilizados ao decorrer do desenvolvimento do trabalho. São explicados conceitos fundamentais de ciência de dados como o ciclo de vida de um projeto até os tipos de abordagem de aprendizado de máquina e suas métricas de sucesso.

2.1.1 Ciclo de Vida de um projeto de Ciência de dados e o TDSP

O Processo de Ciência de Dados em Equipe (TDSP)(MICROSOFT, 2018) é um guia abrangente e estruturado criado pela Microsoft que direciona a execução eficiente de projetos de Ciência de Dados. Entre as diretivas propostas pela TDSP, recomenda-se a aplicação de um ciclo de vida específico para estruturar os projetos nessa área. Para este trabalho, optamos por uma variação desse processo. A seguir, apresentamos uma breve explicação de cada etapa

1. **Entendimento do Negócio**: Na fase inicial, a definição clara dos objetivos do projeto é essencial. Isso inclui identificar as metas, as métricas de sucesso e as partes interessadas envolvidas. O TDSP enfatiza a importância de estabelecer uma base sólida para orientar todas as etapas subsequentes.
2. **Exploração e Entendimento dos Dados**: A fase de exploração e entendimento de dados os objetivos principais são produzir um conjunto de dados limpo e de alta qualidade para criar um ambiente apropriado para modelagem. Isso envolve a ingestão dos dados, a exploração para avaliar a qualidade e, em projetos em estágios mais avançados, a configuração de um pipeline de dados para atualização regular. São utilizadas técnicas de sumarização e visualização para compreender a qualidade dos dados antes do treinamento do modelo.
3. **Modelagem**: A fase de modelagem é dedicada ao desenvolvimento, validação e

otimização de modelos de Machine Learning. O TDSP incentiva uma abordagem iterativa, explorando diversas técnicas e algoritmos para encontrar a solução mais adequada aos objetivos do projeto.

4. **Implantação**: Após o desenvolvimento do modelo, a fase de implantação se concentra na integração eficiente do modelo no ambiente de produção. O TDSP fornece diretrizes para garantir que a implementação seja realizada de maneira eficaz e sustentável, minimizando desafios operacionais.
5. **Aceitação do Cliente**: Validação do cliente para verificar se o sistema atende às necessidades das partes interessadas e se o projeto obteve sucesso ou não.

2.1.2 Aprendizado de Máquina

O aprendizado de máquina é um campo de estudo na área de inteligência artificial, focada no desenvolvimento de algoritmos capazes de aprender e aprimorar seu desempenho por meio da interação com dados. Ao contrário da programação tradicional, em que as instruções são definidas de forma explícita, o aprendizado de máquina permite que os sistemas se ajustem e evoluam com base na experiência adquirida (MAHESH, 2020). Neste Trabalho, fizemos experimentos utilizando dois paradigmas distintos de aprendizado, o supervisionado e o não supervisionado.

2.1.3 Aprendizado Supervisionado

A aprendizagem supervisionada é uma categoria de aprendizado de máquina que utiliza conjuntos de dados rotulados para treinar algoritmos a fim de prever resultados e reconhecer padrões.O objetivo do algoritmo é construir um classificador que possa determinar corretamente a classe de novos exemplos ainda não rotulados (LUDERMIR, 2021).

Durante a criação de um modelo de aprendizado supervisionado, é fundamental dividir os dados de entrada em conjuntos de treinamento e teste. Os dados de treinamento desempenham o papel de exemplos que alimentam o modelo, permitindo-lhe aprender padrões e realizar previsões. Por outro lado, os dados de teste são necessários para poder avaliar o desempenho do modelo, servindo como um conjunto independente que verifica a capacidade do modelo de realizar predições precisas em cenários não vistos durante o treinamento. Essa abordagem de divisão entre treinamento e teste é crucial para garantir

a generalização do modelo e aferir sua eficácia em situações do mundo real.

2.1.4 Aprendizado Não Supervisionado

O aprendizado não supervisionado é uma categoria de aprendizado de máquina que se desenvolve a partir de dados sem supervisão humana. Ao contrário do aprendizado supervisionado, os modelos de aprendizado não supervisionado recebem dados não rotulados e descobrem padrões e *insights* sem orientação explícita ou instrução (MAHESH, 2020). Utilizando algoritmos de autoaprendizagem, esses modelos exploram dados brutos e não rotulados, inferindo suas próprias regras e estruturando as informações com base em similaridades, diferenças e padrões, sem instruções específicas sobre como lidar com cada ponto de dados.

Os algoritmos de aprendizado não supervisionado destacam-se em tarefas de processamento mais complexas, como a organização de grandes conjuntos de dados em agrupamentos. Também são eficazes na identificação de padrões previamente não detectados nos dados, auxiliando na identificação de características úteis para categorização.

Embora o algoritmo em si não compreenda esses padrões com base em informações anteriores fornecidas, o pesquisador pode revisar os agrupamentos de dados e tentar classificá-los com base em sua compreensão do conjunto de dados. O aprendizado não supervisionado, assim, oferece uma abordagem valiosa para explorar e extrair *insights* de dados sem a necessidade de orientação humana prévia.

2.1.5 K-means

K-means é um algoritmo de agrupamento que busca particionar um conjunto de dados em k *clusters* distintos, onde k representa o número pré-definido de grupos desejados. O método se baseia em minimizar a variância intra-*cluster*, ou seja, a soma dos quadrados das distâncias entre os pontos de dados e o centro de seus respectivos *clusters* PEDREGOSA et al., 2011).

A lógica subjacente ao k-means reside na iteração entre atribuir pontos ao *cluster* mais próximo e recalcular o centroide do *cluster* com base nos pontos atribuídos. Esse processo é repetido até que os centróides e a alocação de pontos aos *clusters* se estabilizem. É importante ressaltar que o sucesso do k-means depende da escolha adequada do

número de *clusters*, uma vez que um valor inadequado pode resultar em agrupamentos pouco representativos. Estratégias como o método do cotovelo e validação cruzada são frequentemente empregadas para determinar o número ótimo de *clusters*.

Além de sua simplicidade e eficiência computacional, o k-means destaca-se por sua versatilidade e aplicabilidade em grandes conjuntos de dados. No entanto, é sensível à inicialização dos centróides, o que pode levar a diferentes soluções. Diversas variantes do algoritmo foram propostas para lidar com essas limitações, como o k-means++, que melhora a seleção inicial dos centróides, contribuindo para uma convergência mais rápida e resultados mais estáveis. Em síntese, o algoritmo k-means representa uma ferramenta valiosa na análise exploratória de dados, oferecendo uma abordagem robusta e eficaz para a tarefa desafiadora de agrupamento de dados não rotulados.

O k-means destaca-se não apenas por sua simplicidade e eficiência computacional, mas também por sua versatilidade e aplicabilidade em conjuntos de dados extensos. No entanto, sua sensibilidade à inicialização dos centróides pode resultar em soluções divergentes

2.1.5.1 Método do cotovelo

O método do cotovelo, também conhecido como *elbow method*, é uma técnica utilizada para determinar o número ideal de clusters em um conjunto de dados para o algoritmo k-means. O método basea-se em executar o algoritmo k-means para diferentes valores de k e calcular a soma dos quadrados das distâncias intra-cluster para cada k. Após isso, é criado um gráfico dessas somas de quadrados em relação aos valores de k. O ponto no gráfico onde ocorre uma mudança acentuada, assemelhando-se a um cotovelo, indica o número ideal de *clusters*, sugerindo que adicionar mais *clusters* não traria benefícios substanciais na explicação da variação dos dados.

Ao observar o gráfico gerado pelo método do cotovelo, é possível tomar decisões mais informadas sobre a quantidade de *clusters* a serem utilizados. Essa abordagem oferece uma métrica visual intuitiva para encontrar um equilíbrio entre a complexidade do modelo e a capacidade de explicar a estrutura subjacente dos dados

2.1.6 Teste Qui-Quadrado

O teste Qui-Quadrado (PEARSON, 1900), também conhecido como teste χ^2, é uma ferramenta estatística utilizada para avaliar a independência entre duas variáveis categóricas em um conjunto de dados. Este teste é aplicado quando se pretende investigar se a distribuição observada de frequências em uma tabela de contingência difere significativamente daquela esperada, assumindo que as variáveis são independentes.

A hipótese nula (H0) do teste Qui-Quadrado afirma que não há diferença significativa entre a distribuição observada e a esperada, indicando independência entre as variáveis. Por outro lado, a hipótese alternativa (H1) sugere que há uma associação significativa entre as variáveis.

A estatística do Qui-Quadrado é calculada pela soma dos quadrados dos desvios entre as frequências observadas e esperadas, normalizada pela esperada. Matematicamente, a fórmula é dada por:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Onde:

- χ^2 é a estatística de teste,
- O_i é a frequência observada na categoria i,
- E_i é a frequência esperada na categoria i.

A distribuição da estatística Qui-Quadrado segue uma distribuição qui-quadrado com $(r-1) \times (c-1)$ graus de liberdade, onde r é o número de linhas e c é o número de colunas na tabela de contingência.

2.2 Tecnologias

Nesta seção, abordamos as tecnologias utilizadas no desenvolvimento deste projeto. A linguagem de programação escolhida foi Python, uma escolha fundamentada no fato da linguagem ser o padrão da indústria em ciência de dados, oferecendo uma ampla variedade de bibliotecas e *frameworks* para a execução das mais diversas tarefas. O processo de desenvolvimento foi conduzido integralmente em notebooks, aprimorando a eficiência e a flexibilidade na análise e manipulação dos dados.

2.2.1 Python

Python (FOUNDATION, 2021) é uma linguagem de programação de alto nível e de propósito geral, que tem se destacado como uma ferramenta fundamental na área de ciência de dados, análise de dados e desenvolvimento de software. Criada por Guido van Rossum, Python é conhecida por sua sintaxe clara e legibilidade, facilitando a escrita de código conciso e eficiente. A linguagem possui uma comunidade ativa e um grande ecossistema de bibliotecas que contribuem para a versatilidade da linguagem, tornando-a uma escolha popular em diversas disciplinas acadêmicas e setores industriais.

Nesta seção também serão descritas todas as bibliotecas que auxiliaram no desenvolvimento do trabalho.

2.2.2 Jupyter notebook

Jupyter notebook é um ambiente de desenvolvimento interativo em que os usuários podem executar peças específicas de código e observar a saída em tempo real, ao mesmo tempo que mantém as partes já processadas em memória.

No contexto do desenvolvimento deste trabalho, a escolha da plataforma Jupyter Notebook desempenhou um papel fundamental. O Jupyter Notebook oferece uma abordagem interativa que combina elementos de código, texto explicativo e visualizações, proporcionando uma experiência dinâmica e eficiente no processo de desenvolvimento e comunicação de resultados.

2.2.3 Pandas

A biblioteca Pandas é uma ferramenta fundamental para análise de dados e para o desenvolvimento deste trabalho. Ela proporciona estruturas de dados flexíveis e eficientes para manipulação e análise de conjuntos de dados tabulares. Desenvolvida por Wes McKinney, o Pandas se destaca pela sua capacidade de lidar com dados heterogêneos e de diferentes tipos, permitindo a organização, limpeza e manipulação eficiente de informações. A fundação teórica do Pandas está ancorada em duas estruturas de dados principais: as Séries e os DataFrames.

As Séries são estruturas unidimensionais que armazenam dados de qualquer tipo,

acompanhados de um conjunto de rótulos associados, formando um índice. Essa abordagem fornece uma maneira poderosa e flexível de trabalhar com dados, permitindo operações rápidas e intuitivas.

DataFrames são a espinha dorsal do Pandas, pois incorporam princípios fundamentais da teoria de bancos de dados relacionais e fornecer métodos concisos para realizar tarefas comuns. A capacidade de indexação eficiente, juntamente com operações vetorizadas, impulsiona a eficiência computacional, tornando o Pandas uma escolha central para profissionais que buscam análise de dados em Python.

2.2.4 Matplotlib e Seaborn

Matplotlib e Seaborn são bibliotecas populares em Python utilizadas para visualização de dados. Matplotlib fornece uma ampla gama de funcionalidades para criar gráficos estáticos, como gráficos de linhas, barras, dispersão e histogramas. Sua flexibilidade permite personalizar praticamente todos os aspectos visuais dos gráficos, tornando-o uma ferramenta poderosa.Seaborn, por outro lado, é construído sobre o Matplotlib e oferece uma interface de alto nível para criar gráficos estatísticos atraentes com menos linhas de código. Seu conjunto de funções simplifica a criação de visualizações com paletas de cores agradáveis, estilos pré-definidos e métodos específicos para gráficos estatísticos, como boxplots e mapas de calor. Seaborn é especialmente útil para análise exploratória de dados, facilitando a geração rápida e atraente de gráficos informativos.

Neste trabalho, as duas bibliotecas foram usadas extensivamente para as tarefas de visualização de dados.

3 TRABALHOS RELACIONADOS

Neste capítulo, são introduzidos os trabalhos que utilizam de técnicas de ciência de dados para dados governamentais.

3.1 Algoritmos de Aprendizado de Máquina Aplicados a Dados Públicos para Obtenção de Insights em Segurança Pública

O trabalho de conclusão de curso de graduação da UFRGS em ciência da computação (KREMER, 2022) analisa diferentes algoritmos de aprendizado de máquina em dados de segurança pública em Porto Alegre, especificamente no delito de roubo de veículos.

Os experimentos de aprendizado não supervisionado revelaram a capacidade de agrupar subtrações de veículos com base em localização e valor, identificando padrões nas ocorrências. Paralelamente, os modelos supervisionados alcançaram uma acurácia notável de até 67% na predição da região da cidade onde os veículos subtraídos seriam encontrados. Esses resultados fornecem *insights* valiosos para o desenvolvimento de estratégias mais eficazes na prevenção e combate a crimes dessa natureza.

Destaca-se ainda a criação de uma visualização interativa por meio de *Dashboards*, facilitando a análise eficiente dos agrupamentos identificados. Essa ferramenta amplia a compreensão dos padrões identificados e também possibilita a obtenção de *insights* cruciais para o aprimoramento contínuo das estratégias de segurança pública.

3.2 Determinantes e predição de crimes de homicídios no Brasil: Uma abordagem de aprendizado de máquina

Na pesquisa conduzida na Universidade de São Paulo (LOPES; FELIX, 2019), técnicas de aprendizado de máquina foram empregadas para prever a incidência de homicídios em cidades brasileiras. Neste trabalho, foram testados modelos de regressão utilizando os algoritmos KNN, Árvore de Regressão, Florestas Aleatórias e Boosting.

Os resultados da análise indicam que, entre as 31 covariáveis consideradas, nove apresentam os maiores impactos na violência em nível nacional. Esses fatores incluem, em ordem decrescente de influência: tamanho da população jovem, saneamento básico,

tamanho da população total, população economicamente ativa, população urbana, PIB, mulheres chefes de família, pessoas pobres entre 0 e 14 anos e a proporção de pessoas que ganham até meio salário mínimo. Cada um desses fatores é discutido à luz da literatura econômica do crime. Além disso, destaca-se que o modelo de Florestas Aleatórias foi capaz de explicar, em média, 82% da variabilidade dos crimes de homicídios em nível nacional. Essa abordagem, fundamentada em métodos quantitativos robustos, oferece uma nova ferramenta para orientar os formuladores de políticas públicas na compreensão dos efeitos dos fatores sociais, econômicos e demográficos sobre o crime.

3.3 Data Science Aplicada à Análise Criminal Baseada nos Dados Abertos Governamentais do Brasil

A dissertação de mestrado do PPG em Ciência da Computação da Universidade Federal do Sergipe desenvolvida (PRADO, 2020), utiliza fundamentos de data science para analisar dados abertos governamentais relacionados aos crimes ocorridos nas Unidades Federativas (UFs) brasileiras e nos municípios de Minas Gerais.

A metodologia empregada inclui uma Revisão Sistemática quantitativa, seguida por dois experimentos controlados que exploram regras de associação entre estados, municípios, tipos de crimes, Regiões Integradas de Segurança Pública (RISPs), alvos de roubo e alvos de furto. Adicionalmente, a detecção de outliers e a criação de rankings contribuem para uma análise abrangente. Os resultados revelam padrões marcantes, indicando que o Paraná é consistentemente o estado mais perigoso, seguido pelo Rio de Janeiro. O estudo também destaca a utilidade da Data Science na realização de diagnósticos precisos e rápidos, oferecendo insights valiosos para o planejamento estratégico e a tomada de decisões em Segurança Pública.

3.4 DATA SCIENCE & SEGURANÇA PÚBLICA: padrões estatísticos sobre as ocorrências de flagrantes em roubo de celular na cidade de São Paulo.

Neste trabalho, o autor (VARGAS, 2019) analisa 1.001.006— um milhão, mil e seis— Boletins de Ocorrências (B.O.), do Registro Digital de Ocorrências (R.D.O.), acerca de roubo de celular na cidade de São Paulo no período de 10 anos, desde o primeiro dia do mês de janeiro do ano de 2010 e até o último dia do mês de dezembro de 2018, a

partir da base de dados da secretaria da Segurança Pública do Governo do Estado de São Paulo.

Na primeira parte do trabalho, a ênfase foi na inferência, ajustando modelos lineares convencionais e espaciais. Notavelmente, um modelo destacou-se ao explicar mais de 84% da variância nas ocorrências totais de roubo de celular na cidade, utilizando variáveis como ocorrências noturnas e envolvimento de motos. Esse desempenho robusto foi validado por testes de colinearidade, normalidade, análise de variância e auto correlação dos resíduos. Do ponto de vista da Política Pública, o objetivo era prever como diferentes variáveis influenciam os registros de Boletins de Ocorrência, visando melhorar informações para o policiamento e prevenção de roubos. Na segunda parte, focou-se em aprendizado de máquina para prever flagrantes em crimes de roubo de celular em São Paulo. O algoritmo Bagging, com 500 Árvores, destacou-se, prevendo 60,48% dos casos. Comparando vários algoritmos, concluiu-se que Random Forest e Bagging têm o melhor desempenho, prevendo cerca de 60,5% dos flagrantes nesse tipo de crime na cidade.

Tabela 3.1: Comparação entre os trabalhos relacionados

	KREMER, 2022	**LOPES; FELIX, 2019**	**PRADO, 2020**	**VARGAS, 2019**
Aprendizado Supervisionado	**x**	**x**		**x**
Aprendizado Não Supervisionado	**x**			
Demografia			**x**	
Realizado no RS	**x**			

4 DEMOGRAFIA DA BASE DE DADOS

A análise demográfica desempenha um papel crucial na compreensão dos padrões e dinâmicas no conjunto de dados deste trabalho. Neste capítulo, são apresentados e analisados os dados demográficos relacionados às vítimas e aos criminosos envolvidos nas ocorrências criminais do Rio Grande do Sul, proporcionando uma visão detalhada das características populacionais dentro do contexto deste conjunto de dados. A base de dados foi disponibilizada pela Brigada Militar do Rio Grande do Sul.

4.1 Demografia das vítimas

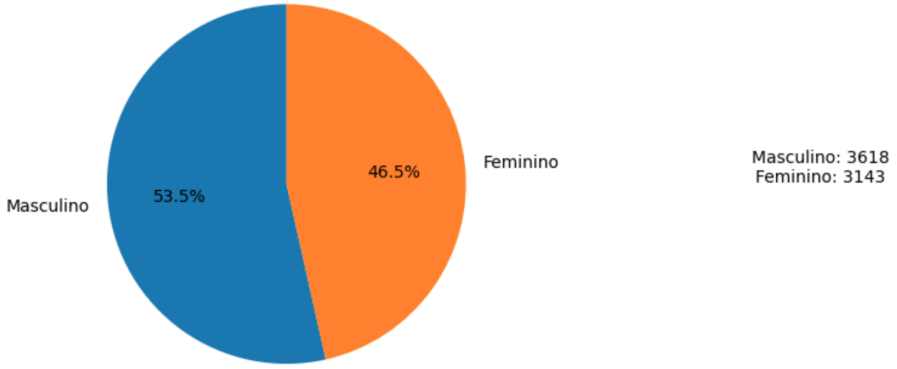
Nesta seção, abordamos exclusivamente a análise demográfica das vítimas de roubo de celulares. Ao concentrar nossa atenção nas características específicas relacionadas às vítimas, como idade, gênero, estado civil e nível de instrução, conseguimos ter um entendimento do cenário geral deste crime no estado do Rio Grande do Sul, e também identificar e investigar possíveis tendências e correlações.

4.1.1 Distribuição de sexo

A análise da distribuição de sexo entre as vítimas de roubo de celulares no Rio Grande do Sul revela uma relativa equidade, com 53.5% das vítimas sendo do sexo masculino e 46.5% do sexo feminino. Embora a disparidade não seja ampla, ainda é um elemento significativo a ser considerado na compreensão das dinâmicas desse tipo de crime na região.

A quase paridade entre os sexos pode indicar que o roubo de celulares não demonstra uma tendência clara de direcionamento com base no gênero das vítimas. Este resultado sugere a importância de uma abordagem equitativa nas estratégias de prevenção e combate ao crime, garantindo que medidas de segurança sejam eficazes para ambos os grupos. Na Figura 4.1 é apresentado o gráfico de setores correspondente à distribuição de sexo entre as vítimas.

Figura 4.1: Distribuição de sexo entre as vítimas

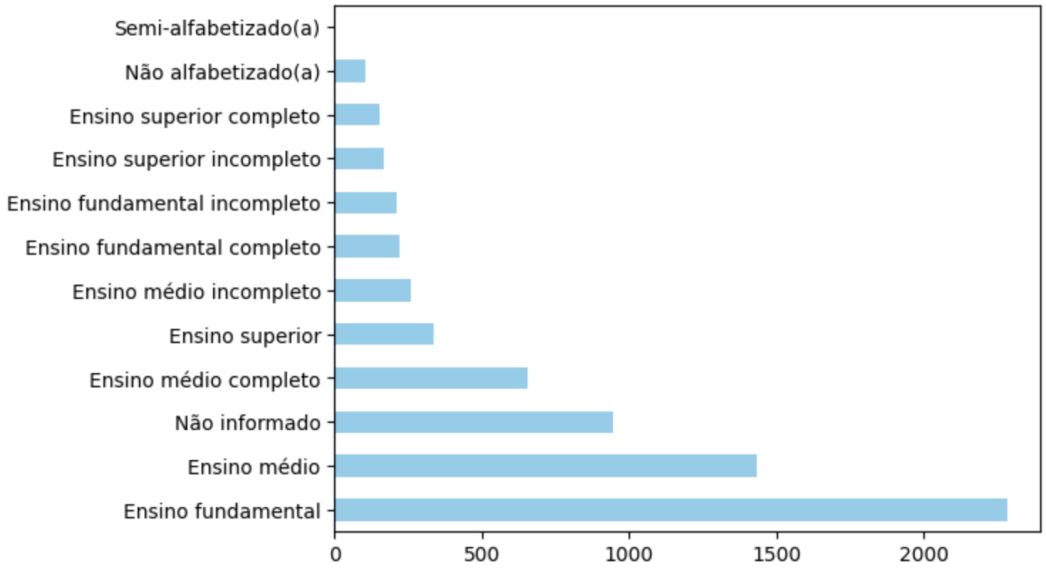


Fonte: Elaborado pelo autor

4.1.2 Grau de instrução

A análise do grau de instrução das vítimas de roubo de celulares no Rio Grande do Sul revela uma diversidade educacional. O ensino fundamental é a categoria mais comum, com 2283 vítimas, seguido pelo ensino médio, com 1431 vítimas. Destaca-se a presença significativa de vítimas que não forneceram informações sobre a escolaridade (944), indicando a necessidade de aprimorar a coleta de dados. Vítimas com ensino superior completo ou incompleto somam 502, sugerindo que o roubo de celulares não está restrito a grupos específicos de instrução. Na Figura 4.2 é apresentado o gráfico de barras horizontais correspondente à distribuição dos graus de escolaridade entre as vítimas.

Figura 4.2: Distribuição de Grau de Instrução

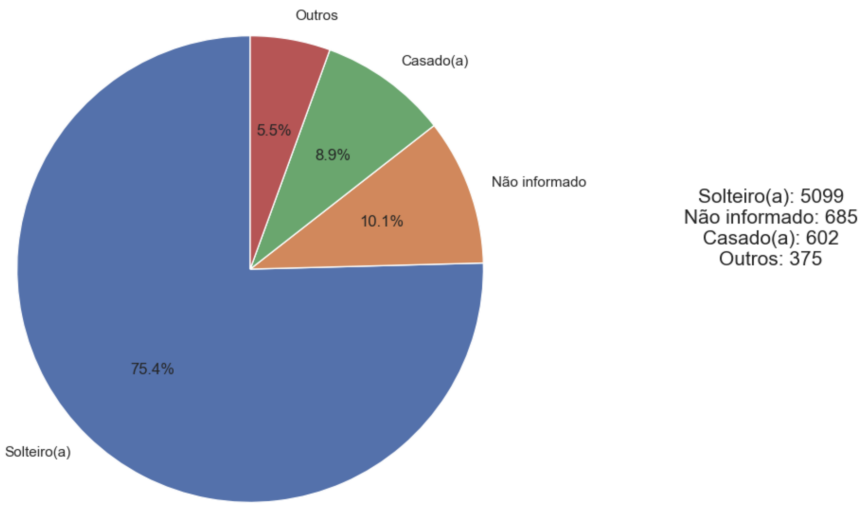


Fonte: Elaborado pelo autor

4.1.3 Estado Civil

A base de dados revela uma predominância de indivíduos solteiros, totalizando 5099 registros, seguidos por aqueles que não forneceram informações sobre o estado civil, com 685 casos. Os dados também indicam uma presença significativa de pessoas casadas (602), enquanto a categoria "Divorciado(a)"apresenta 200 casos, e "Separado(a)"e "Viúvo(a)"possuem 79 e 47 registros, respectivamente. Na Figura 4.3 é apresentado um gráfico de setores correspondente à distribuição dos estados civis das vítimas no conjunto de dados, porém reduzido para as 3 maiores categorias (solteiros, casados e não informado) que representam 95.5% das ocorrências e agrupando o resto dos estados civis presentes na categoria "outros".

Figura 4.3: Estado civil



Fonte: Elaborado pelo autor

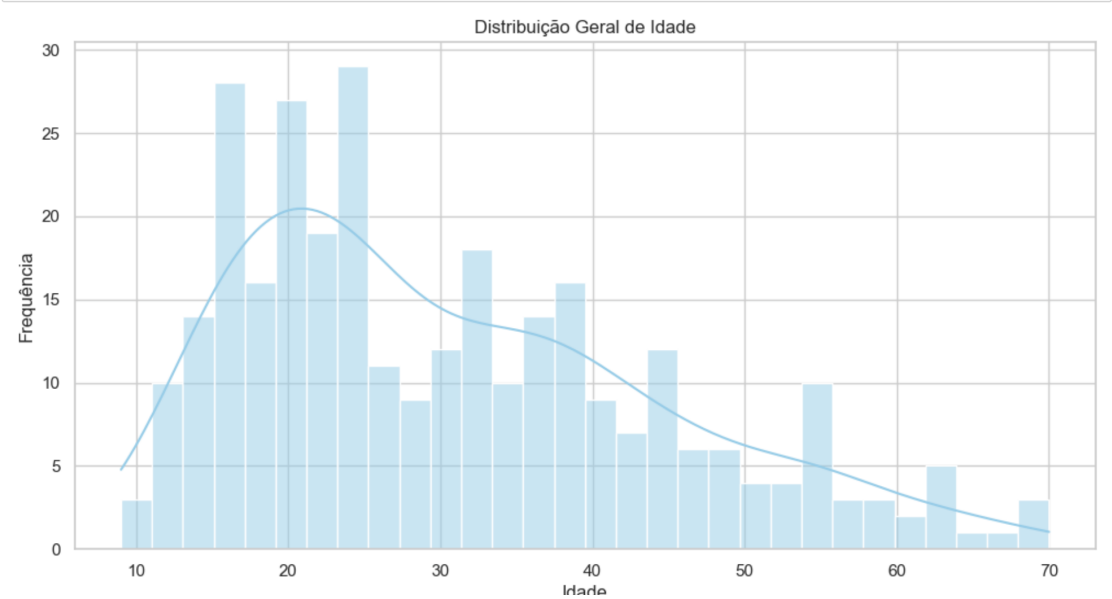
4.1.4 Idade

A análise da idade das vítimas de roubo de celulares no Rio Grande do Sul revela uma ampla distribuição etária. As idades mais comuns entre as vítimas estão concentradas em faixas relativamente jovens, com 17 casos registrados para a idade de 20 anos, seguidos por 16 casos para 16 anos e 15 casos para 25 anos. A presença significativa de casos em faixas etárias como 23, 24, 17 e 18 anos destaca a vulnerabilidade desse grupo demográfico específico.

Além disso, é importante observar que o fenômeno do roubo de celulares não se

limita a uma faixa etária específica, uma vez que casos são registrados em diversas idades, incluindo indivíduos mais velhos, como 63, 56 e 70 anos. Na figura 4.4 é apresentado o histograma da distribuição das idades entre as vítimas, acrescentado da curva da distribuição normal.

Figura 4.4: Distribuição de idade



Fonte: Elaborado pelo autor

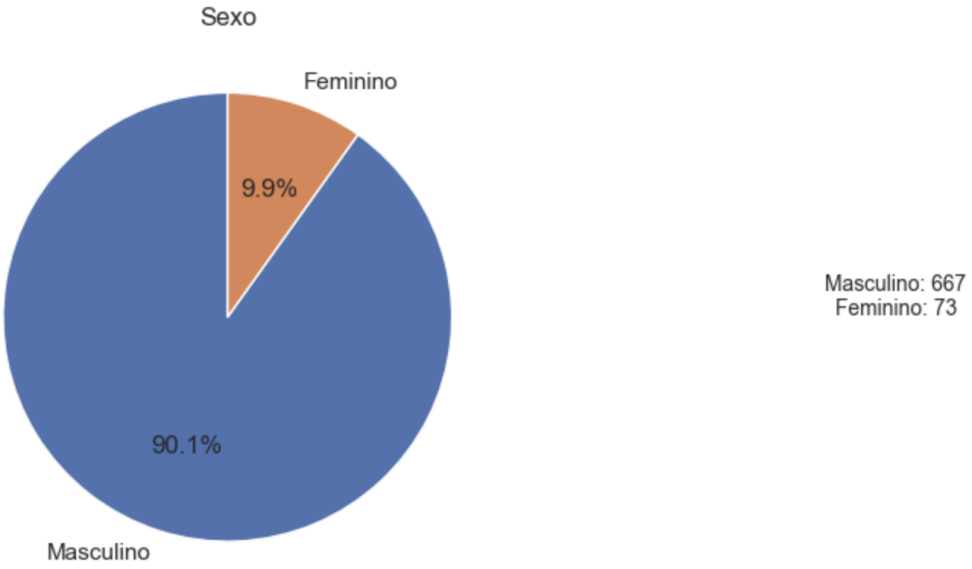
4.2 Demografia dos criminosos

Nesta seção, dirigimos nosso foco à análise demográfica dos criminosos envolvidos em ocorrências de roubo de celulares. Similar à abordagem para as vítimas, a compreensão aprofundada da demografia e dos perfis dos criminosos é de extrema relevância. Buscaremos desvendar o perfil dos perpetradores, explorando variáveis como sexo, idade, estado civil e suas inter-relações.

4.2.1 Distribuição de sexo

A distribuição de sexo entre os perpetradores de roubo de celulares no Rio Grande do Sul evidencia uma clara disparidade, com 667 casos associados ao sexo masculino e apenas 73 casos ao sexo feminino. Esta predominância masculina aponta para uma distinção notável no envolvimento de homens nesse tipo de crime no estado. Na Figura 4.6, podemos ver o gráfico de setores representando a distribuição de sexo entre os criminosos.

Figura 4.5: Distribuição de sexo



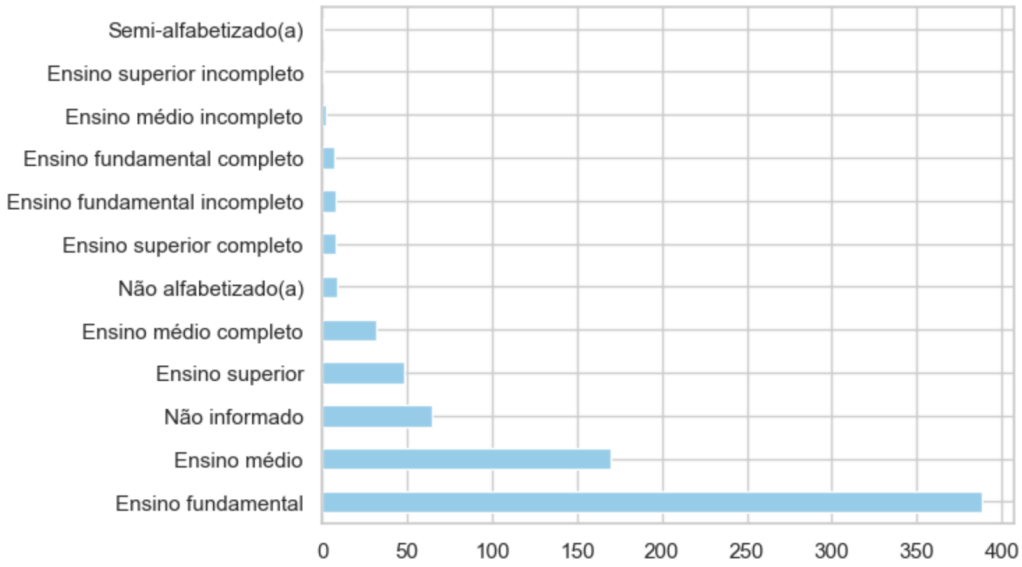
Fonte: Elaborado pelo autor

4.2.2 Grau de instrução

A análise do grau de instrução dos perpetradores de roubo de celulares no Rio Grande do Sul revela uma diversidade educacional considerável. A maioria dos criminosos possui educação até o ensino fundamental, com 388 casos, seguido pelo ensino médio, com 170 casos. Nota-se uma parcela significativa de casos em que a informação sobre o grau de instrução não foi fornecida (65).

É interessante observar que, embora exista representação em todos os níveis de instrução, a quantidade de perpetradores com ensino superior completo ou incompleto é extremamente baixa, com 8 e 1 casos, respectivamente. A presença de indivíduos não alfabetizados ou semi-alfabetizados é observada em 9 e 1 casos, respectivamente. Na Figura 4.7 é apresentado o gráfico horizontal de barras que representa a distribuição de grau de instrução entre os criminosos.

Figura 4.6: Grau de instrução

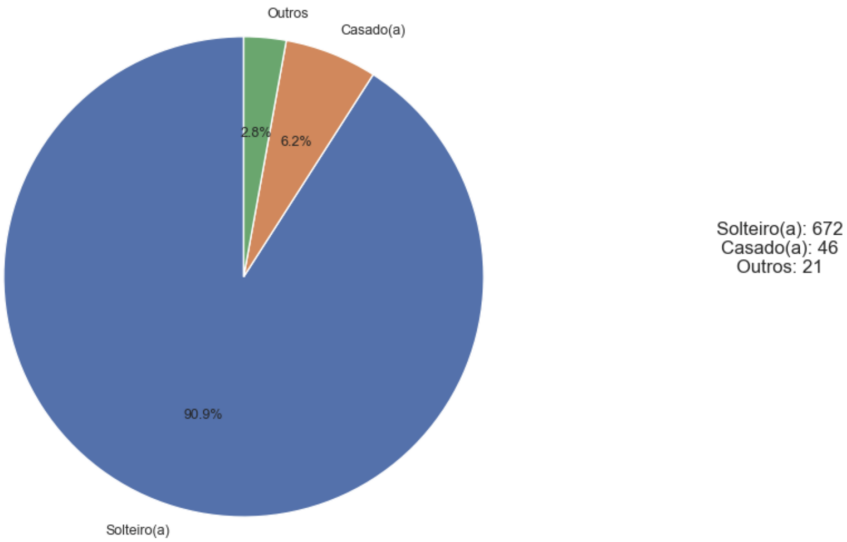


Fonte: Elaborado pelo autor

4.2.3 Estado Civil

A análise do estado civil dos perpetradores de roubo de celulares no Rio Grande do Sul revela uma predominância de indivíduos solteiros, totalizando 672 casos. Em contraste, há uma representação relativamente menor de criminosos casados (46 casos), amigados (7 casos), e aqueles que não informaram seu estado civil (5 casos). Divorciados e separados compõem 5 e 4 casos, respectivamente. Na Figura 4.8 é apresentado um gráfico de setores correspondente à distribuição dos estados civis dos criminosos, porém reduzido para as 2 maiores categorias (solteiros, casados) que representam 96.2% das ocorrências. As outras ocorrências foram incluídas no setor "outros".

Figura 4.7: Estado Civil



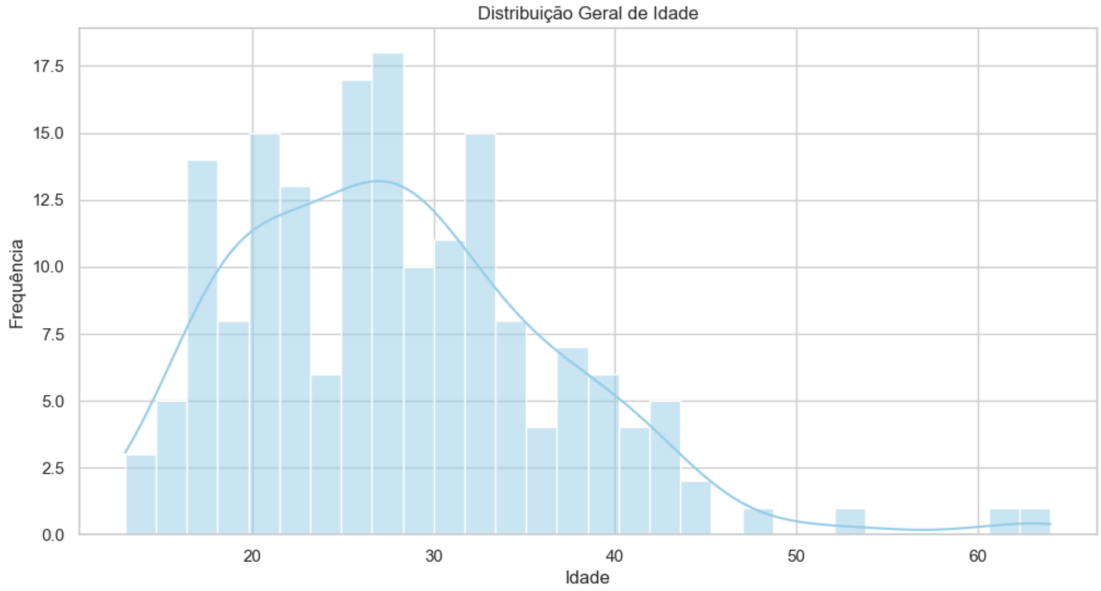
Fonte: Elaborado pelo autor

4.2.4 Idade

A análise da idade dos perpetradores de roubo de celulares no Rio Grande do Sul destaca uma variedade de faixas etárias envolvidas nesse tipo de crime. Observa-se que a faixa dos 20 anos é proeminente, com 12 casos registrados. Além disso, as idades de 25, 29, 28 e 30 anos apresentam uma frequência significativa, com 11, 10, 10 e 9 casos, respectivamente.

A distribuição abrange diversas faixas etárias, indicando que o envolvimento em crimes de roubo de celulares não está restrito a um grupo etário específico.

Figura 4.8: Grau de instrução



Fonte: Elaborado pelo autor

4.2.5 Considerações Finais

Em conclusão, a análise demográfica das vítimas e dos criminosos envolvidos em casos de roubo de celulares no Rio Grande do Sul oferece uma visão abrangente das características sociodemográficas relacionadas a esse tipo de crime. A distribuição equivalente entre os sexos das vítimas sugere que o roubo de celulares não demonstra uma tendência clara de direcionamento com base no gênero. A diversidade educacional e a predominância de casos entre aqueles com ensino fundamental entre as vítimas destacam a necessidade de estratégias de prevenção a diferentes níveis de instrução.

A análise do estado civil revela uma significativa presença de indivíduos solteiros entre as vítimas, sugerindo que esse grupo pode ser mais suscetível a esse tipo de crime. No entanto, é essencial considerar que o roubo de celulares afeta uma variedade de estados civis, indicando a complexidade dessa dinâmica criminal. Além disso, a análise demográfica dos criminosos destaca uma disparidade de gênero, com uma predominância significativa de homens envolvidos nesse tipo de crime.

Outrossim, a ampla distribuição etária tanto das vítimas quanto dos criminosos destaca a complexidade e a abrangência do fenômeno do roubo de celulares. Essas conclusões fornecem *insights* valiosos para o entendimento geral do problema, além de ajudar na formulação de hipóteses a serem respondidas com os dados disponíveis.

5 METODOLOGIA

O objetivo deste trabalho é realizar experimentos de ciência de dados para a identificação de padrões e obter *insights* na segurança pública do estado do Rio Grande do Sul, tendo o enfoque no crime de roubo de celulares. Este capítulo descreve a metodologia proposta para abordar o problema, obter os dados, realizar os experimentos e analisar os resultados.

5.1 Visão Geral

Este trabalho adota como estrutura metodológica o *Team Data Science Process* (TDSP) (MICROSOFT, 2018), uma metodologia desenvolvida pela Microsoft especificamente para orientar projetos de ciência de dados. A robustez e abrangência do TDSP proporcionam uma estrutura eficaz para condução de pesquisas científicas nesta área, sendo a base sobre a qual este projeto foi construído. Este capítulo tem como objetivo oferecer uma visão abrangente da metodologia utilizada todas as etapas do TDSP, enfatizando as variações incorporadas que se alinham às necessidades da pesquisa.

O TDSP é concebido como um processo colaborativo e interdisciplinar, envolvendo diversas etapas que englobam desde a compreensão do contexto até a implementação e manutenção de soluções baseadas em dados. A etapa inicial, conhecida como *Business Understanding*, destaca-se como a base fundamental de todo o processo, buscando a compreensão profunda dos objetivos científicos e das questões de pesquisa.

No contexto científico, a fase de Business Understanding é adaptada para contemplar a identificação e formulação precisa de questões científicas que orientarão o desenvolvimento da pesquisa. Este capítulo não apenas apresentará as etapas do TDSP, mas também abordará as nuances introduzidas para adequação ao escopo específico da pesquisa, garantindo uma aplicação precisa e eficaz da metodologia no contexto científico.

5.2 Entendimento do problema

A fonte de inspiração para esta pesquisa surge da análise dos registros de incidentes divulgados pela SSPRS, já também explorado em outros trabalhos de análise de dados, como (KREMER, 2023). Visando compreender a questão e explorar a viabilidade de rea-

lizar experimentos nesses dados, ocorreu uma reunião com o comando de inteligência da brigada militar. Durante o encontro, o projeto e a proposta de colaboração foram apresentados. A receptividade foi positiva, sugerindo o foco em uma categoria específica muito popular de delitos: roubo de celulares. A solicitação dos dados foi realizada diretamente à brigada militar do RS, intermediada pelo tenente-coronel Roberto Donato.

5.3 Obtenção dos Dados

A base de dados, obtida através da transparência passiva, foi enviada no formato de planilha *xlsx*. Cada linha na tabela registra algum participante de uma ocorrência de furto de celular, que podem ser Vítima, Só comunicante, Testemunha, Suspeito(a), Indiciado(a), Condutor(a), Acusado(a), Adolescente infrator, Foragido(a), Autor(a).

5.4 Pré-Processamento e Limpeza dos Dados

O tratamento adequado dos dados é um aspecto crucial em qualquer pesquisa científica, e a fase de pré-processamento e limpeza desempenha um papel fundamental na garantia da qualidade e confiabilidade dos resultados obtidos na fase de experimentação. Neste contexto, adotamos uma abordagem alinhada com o *lifecycle* Team Data Science Process (TDSP). A primeira etapa envolve a identificação e tratamento de valores ausentes, empregando métodos de imputação para preencher lacunas e evitar distorções nos resultados.

Além disso, durante o pré-processamento, lidamos com *outliers* que podem impactar negativamente as análises futuras. Técnicas robustas, como a detecção estatística de valores extremos, são aplicadas para garantir que a presença de pontos discrepantes não comprometa a integridade dos dados. Além disso, foi feita uma checagem manual para garantir que não existam valores inseridos de forma errônea.

Na etapa seguinte, concentramos nossos esforços na normalização e padronização dos dados. Essas transformações buscam garantir que diferentes variáveis estejam na mesma escala, facilitando comparações significativas. A discretização controlada é uma técnica adicional utilizada para otimizar a representação dos dados sem perda substancial de informações. Ao final deste processo, teremos um conjunto de dados refinado, pronto para ser usado em análises mais complexas e modelos de AM.

5.5 Análise de Dados

O principal objetivo desta análise é extrair *insights* significativos a partir da formulação de hipóteses que contribuam para a compreensão do cenário de roubo de celulares. Desta maneira, buscamos não apenas identificar padrões, mas também validar suposições que enriqueçam a interpretação dos resultados obtidos nos experimentos.

5.5.1 K-Means

Foi utilizado o algoritmo K-Means para a identificação de padrões e agrupamentos nos dados relacionados a ocorrências de roubo de celulares. Esse método de agrupamento permitiu-nos segmentar os dados, possibilitando uma análise mais aprofundada dentro das características associadas aos grupos presentes no conjunto de dados.

5.5.2 Métodos de Visualização de Dados

Recorremos a diversas técnicas de visualização, como gráficos de barras, setores e de violino, para proporcionar uma representação intuitiva e elucidativa das relações entre variáveis. Essas visualizações contribuem para uma compreensão mais clara dos padrões espaciais e temporais dos incidentes de roubo de celulares.

5.5.3 Métodos Estatísticos, incluindo Qui-Quadrado:

Adotamos métodos estatísticos, notadamente o teste do qui-quadrado (PEARSON, 1900), para avaliar a significância estatística de relações identificadas durante a análise. Isso adiciona uma camada de rigor estatístico às conclusões extraídas, permitindo uma interpretação confiável dos resultados.

Ao incorporar essas técnicas na etapa de análise de dados, buscamos não apenas apresentar resultados, mas também proporcionar uma compreensão abrangente do processo analítico empregado. Essa abordagem visa assegurar a transparência e reprodutibilidade dos resultados, contribuindo para a robustez da metodologia adotada no presente estudo.

6 EXPERIMENTOS E RESULTADOS

Neste capítulo, apresenta-se o desfecho da implementação da metodologia sugerida no capítulo 4, a qual foi aplicada para abordar as ocorrências criminais relacionadas a subtrações e recuperações de veículos no estado do Rio Grande do Sul. Adicionalmente, são discutidas as principais conclusões e *insights* obtidos durante a aplicação do método.

6.1 Entendimento do Problema

A inspiração para este trabalho surgiu da análise dos registros de ocorrência disponibilizados pela SSPRS, especificamente no contexto de roubo de celulares. Dado o crescimento e popularidade deste tipo de delito, o intuito principal é compreender a problemática e explorar os dados utilizando metodologias *data driven*. Após a formulação dessa ideia inicial, promoveu-se uma reunião com o departamento de policiamento inteligente da Brigada Militar do Rio Grande do Sul, na qual a proposta do projeto foi bem acolhida.

6.2 Obtenção dos Dados

Os registros relacionados aos roubos de celulares foram fornecidos no formato de planilha xlsx, em que cada linha na tabela representa a presença de uma pessoa específica na ocorrência. As categorias associadas a cada indivíduo variam entre vítima, só comunicante, testemunha, suspeito(a), indiciado(a), condutor(a), acusado(a), adolescente infrator, foragido(a) ou autor(a). Em virtude da ausência de um dicionário de dados, procedeu-se à criação de um, baseado nas informações disponíveis e detalhado na Tabela 5.1

Tabela 6.1: Dicionário de Dados

Coluna	**Descrição**	**Tipo de Dado**
data_fato	Data em que ocorreu o fato	Numérico Cardinal (temporal)
municipio	Município onde o incidente ocorreu	Textual Categórico (nominal)
crpo	CRPO associado ao ocorrido	Textual Categórico (nominal)
unidade	Unidade policial responsável	Textual Categórico (nominal)
tipo_fato	Categoria geral do fato ocorrido	Textual Categórico (nominal)
sexo	Gênero da pessoa envolvida no fato	Textual Categórico (nominal)
condicao_fisica	Estado físico da pessoa no momento do incidente	Textual Categórico (nominal)
tipo_participacao	Papel ou função da pessoa na ocorrência (vítima, suspeito, etc.)	Textual Categórico (nominal)
medida_protetiva	Existência de medidas protetivas associadas à pessoa	Textual Categórico (nominal)
cor_pele	Cor da pele da pessoa	Textual Categórico (nominal)
idade_minima	Idade mínima da pessoa	Numérico Cardinal
idade_maxima	Idade máxima da pessoa	Numérico Cardinal
idade_aparente	Idade aparente da pessoa	Numérico Cardinal
altura_minima	Altura mínima da pessoa	Numérico Cardinal
altura_maxima	Altura máxima da pessoa	Numérico Cardinal
tipo_altura	faixa de altura (ex: 1,61 a 1,80)	Textual Categórico (nominal)
grau_instrucao	Nivel de escaridade da pessoa	Textual Categórico (ordinal)
estado_civil	Estado civil da pessoa	Textual Categórico (nominal)
tipo_cabelo	Tipo de cabelo da pessoa	Textual Categórico (nominal)
cor_cabelo	Cor do cabelo da pessoa	Textual Categórico (nominal)
cor_olhos	Cor dos olhos da pessoa	Textual Categórico (nominal)
cidade_residencia	Cidade de residência da pessoa	Textual Categórico (nominal)
uf_residencia	Unidade da Federação de residência da pessoa	Textual Categórico (nominal)
bairro_residencia	Bairro de residência da pessoa	Textual Categórico (nominal)
profissao	Profissão da pessoa	Textual Categórico (nominal)
orgao_registro	Órgão de registro associado à pessoa	Textual Categórico (nominal)
ano_registro	Ano do registro associado à pessoa	Numérico Ordinal
quantidade	Quantidade associada à ocorrência	Numérico Cardinal

6.3 Pré-processamento e Limpeza

Nesta fase, realizamos a transição da planilha obtida na fase anterior para o formato de *Dataframes*. Além disso, conduzimos a limpeza dos dados, manipulando valores ausentes e, se necessário, removendo-os caso não sejam considerados significativos. Adicionalmente, implementamos padronizações nos atributos textuais para assegurar consistência dentro da mesma formatação.

No projeto em questão, foi realizado um processo de múltiplas etapas, consistindo em:

1. Transformar todos os atributos em minúsculo e remover pontuações, acentos e espaços.
2. Correção manual de erros de inserção de dados, como endereços iguais escritos de forma diferente. Um exemplo relevante desse processo consistiu na padronização de registros, como a uniformização de "Rubem Berta"para "R. Berta", um dos bairros de maior ocorrência dentro desse conjunto de dados.
3. Exclusão de casos atípicos evidenciados, como situações de criminosos cuja altura apresenta valores mínimos de 20cm e máximos de 999cm.
4. Redução de atributos categóricos com muitos valores diferentes, como os exemplos demográficos de estado civil que foram reduzidos a solteiros, casados e outros.

Todo o processo de pré-processamento e limpeza foi realizado em Jupyter Notebooks, utilizando a biblioteca Pandas na linguagem Python. Essa etapa é fundamentais para preparar os dados de maneira adequada antes de prosseguir para análises mais avançadas.

6.4 Agrupamento de Pessoas Baseado no Tipo de Participação no Crime

A concepção da primeira hipótese surgiu ao explorar as relações entre os participantes no crime de roubo de celulares. Ao analisar os dados de forma abrangente, uma pergunta fundamental emergiu: Existe alguma correlação significativa entre os diferentes papéis desempenhados por indivíduos envolvidos nesses incidentes?

A motivação por trás desta hipótese foi impulsionada pela busca de identificar padrões nas características demográficas, que possam revelar correlações específicas para cada tipo de participação. Além disso, a compreensão dessa dinâmica não só poderia enriquecer a visão dos diferentes perfis dos tipos de participante do delito, mas também

fornecer *insights* valiosos sobre as relações entre os envolvidos.

6.4.1 K-Means

O processo de análise e agrupamento envolveu a aplicação do algoritmo K-Means para explorar padrões nos dados. Inicialmente, optou-se por uma abordagem que priorizasse a investigação nos cinco municípios com o maior número de ocorrências, buscando assim compreender se existem possíveis correlações locais e identificar padrões específicos nessas regiões. A escolha desses municípios estratégicos visou proporcionar uma análise mais aprofundada, considerando a possível variabilidade geográfica nas dinâmicas criminosas.

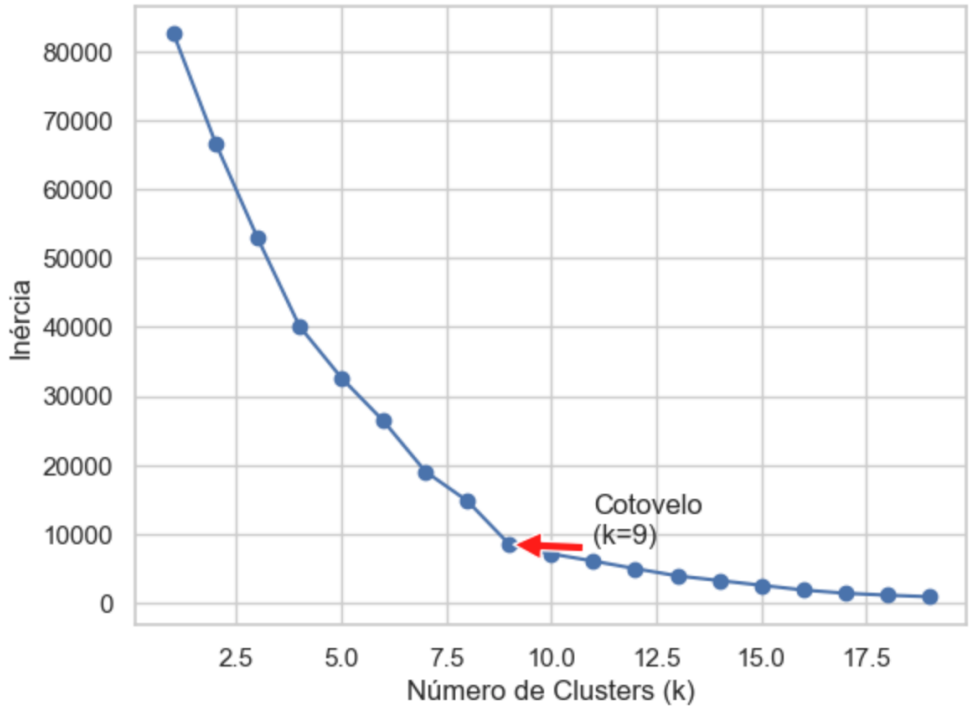
Posteriormente, as categorias "Suspeito(a)", "Indiciado(a)", "Acusado(a)", "Condutor(a)", "Foragido(a)"foram consolidadas sob o rótulo "Criminosos", simplificando assim a análise para concentrar o foco nos participantes envolvidos em atividades criminosas. A filtragem subsequente se concentrou nos dois grupos principais: criminosos e vítimas, permitindo uma investigação mais precisa das dinâmicas entre esses participantes.

No entanto, Ao realizar uma análise mais aprofundada, constatou-se que a categorização por municípios não proporcionava *insights* significativos. Os resultados mais promissores foram alcançados ao combinar as colunas "sexo", "faixa etária", "tipo de participação"e "município", resultando na formação de 11 *clusters*. No entanto, ao examinar as distribuições desses *clusters*, não foi possível identificar uma correlação aparente entre os indivíduos agrupados. Diante desse cenário, optou-se por descartar a categorização por municípios como parte do processo de análise, uma vez que essa abordagem não contribuía de maneira substancial para a compreensão dos padrões criminais identificados. Essa decisão foi baseada na falta de associação clara entre a localização geográfica e as características dos participantes, evidenciando a complexidade do fenômeno em questão.

Diante da falta de *insights* significativos na análise centrada em cinco municípios de maior ocorrência criminal, a abordagem foi ajustada para uma visão estadual mais abrangente. Mantendo as variáveis "sexo", "faixa etária", "tipo de participação"e "município", o algoritmo K-Means foi novamente aplicado, resultando em 7 *clusters*, valor definido pela regra do cotovelo representado na figura 6.1. A decisão de eliminar a categorização por municípios, baseada na falta de associação clara entre os *clusters*, guiou esta expansão. A análise estadual revelou padrões mais robustos, superando as limitações da

abordagem municipal anterior e proporcionando uma compreensão mais abrangente das dinâmicas criminosas. A categorização "Criminosos"foi mantida, concentrando a análise nas interações entre criminosos e vítimas.

Figura 6.1: Regra do cotovelo



Fonte: Elaborado pelo autor

6.4.2 Análise dos clusters

Cluster 1 (1879 ocorrências - 25,3% do dataset):

- Gênero: 90,4% Masculino, 9,6% Feminino.
- Tipo de Participação: 100% Criminoso.
- Educação: 66,6% Fundamental, 26,3% Médio, 7% Superior.
- Estado Civil: 99,9% Solteiro.

Cluster 3 (1509 ocorrências - 20,4% do dataset):

- Gênero: 100% Feminino.
- Tipo de Participação: 100% Vítima.
- Educação: 100% Fundamental.
- Estado Civil: 100% Solteiro.

Cluster 2 (1166 ocorrências - 15,8% do dataset):

- Gênero: 100% Masculino.
- Tipo de Participação: 100% Vítima.
- Educação: 100% Médio.
- Estado Civil: 100% Solteiro.

Cluster 5 (949 ocorrências - 12,9% do dataset):

- Gênero: 100% Feminino.
- Tipo de Participação: 100% Vítima.
- Educação: 100% Médio.
- Estado Civil: 100% Solteiro.

Cluster 0 (685 ocorrências - 9,3% do dataset):

- Gênero: 90,4% Masculino, 9,6% Feminino.
- Tipo de Participação: 100% Criminoso.
- Educação: 66,6% Fundamental, 26,3% Médio, 7% Superior.
- Estado Civil: 99,9% Solteiro.

Cluster 4 (645 ocorrências - 8,7% do dataset):

- Gênero: 50,1% Masculino, 49,9% Feminino.
- Tipo de Participação: 6,98% Criminoso, 93% Vítima.
- Educação: 43,3% Fundamental, 39,7% Médio, 17% Superior.
- Estado Civil: 100% Casado.

Cluster 8 (554 ocorrências - 7,5% do dataset):

- Gênero: 41,5% Masculino, 58,5% Feminino.
- Tipo de Participação: 100% Vítima.
- Educação: 100% Superior.
- Estado Civil: 100% Solteiro.

Cluster 6 (111 ocorrências - 1,5% do dataset):

- Gênero: 41,4% Masculino, 58,6% Feminino.
- Tipo de Participação: 8,11% Criminoso, 91,9% Vítima.
- Educação: 100% Analfabeto.

- Estado Civil: 97,3% Solteiro.

Cluster 7 (3 ocorrências - 0,04% do dataset):

- Gênero: 100% Masculino.
- Tipo de Participação: 33,3% Criminoso, 66,7% Vítima.
- Educação: Sem informações suficientes para análise.
- Estado Civil: 100% Solteiro.

Análise Geral:

- A maioria dos *clusters* é dominada por uma categoria específica de participação (Criminoso ou Vítima).
- Os *clusters* 1 e 0 são semelhantes, ambos predominantemente masculinos, com participação criminal, educação fundamental/média e estado civil solteiro.
- Os *clusters* 3, 2 e 5 são compostos principalmente por mulheres, sendo o 3 totalmente feminino com participação de vítima e educação fundamental/média.
- O *clusters* 4 é diversificado, com uma distribuição mais equitativa de gênero e participação tanto como criminoso quanto como vítima, com uma proporção significativa de casados.
- Os *clusters* 8 e 6 têm uma tendência mais feminina e são caracterizados por participação como vítima, com o cluster 6 sendo notável por uma alta taxa de analfabetismo.
- O *clusters* 7 é pequeno e heterogêneo, com uma mistura de participação criminal e vítima entre homens solteiros.

6.5 IDH e Renda per Capita

Investigar as complexidades da incidência de roubo de celulares envolve a exploração das possíveis correlações entre as características socioeconômicas e a participação em atividades criminosas. A hipótese sugere que fatores como o Índice de Desenvolvimento Humano (IDH) e a renda *per capita* desempenham papéis relevantes na diferenciação entre vítimas e criminosos. Optou-se por enfocar a análise por meio de técnicas de visualização de dados, permitindo uma compreensão mais acessível e intuitiva das disparidades socioeconômicas entre esses grupos. Esta abordagem visa identificar padrões visuais que possam oferecer *insights* sobre as relações entre as variáveis em questão.

Para aprofundar a exploração, foram selecionados os top 10 bairros de residência tanto de vítimas quanto de criminosos em Porto Alegre, sendo essa escolha restrita à capital devido à disponibilidade limitada de dados sobre bairros de outros municípios do Rio Grande do Sul. A fim de capturar as nuances socioeconômicas dessas áreas específicas, optou-se por buscar manualmente informações de Índice de Desenvolvimento Humano (IDH) e renda per capita. Esses dados foram predominantemente obtidos por meio da Procempa, a Empresa de Tecnologia da Informação e Comunicação da Prefeitura de Porto Alegre. A decisão de focar nos top 10 bairros permite uma análise mais detalhada das características socioeconômicas dessas localidades-chave, enquanto a busca manual de dados visa garantir uma representação precisa e abrangente das condições demográficas em questão

Tabela 6.4: Comparação entre os bairros de Porto Alegre com maior número de residentes entre criminosos e vítimas

Bairro	**Quantidade**
Centro	37
Praia de Belas	22
Restinga	12
AP Borges	8
M Deus	8
Lomba do Pinheiro	8
Partenon	7
Bom Jesus	4
Cavalhada	3
Floresta	3

Bairro	**Quantidade**
Lomba do Pinheiro	126
Rubem Berta	95
Restinga	87
Centro	69
Partenon	57
Sarandi	55
Mário Quintana	45
Cascata	34
Cavalhada	29
P.D. Areia	29

A tabela à esquerda representam os criminosos e a direita as vítimas.

6.5.1 IDH

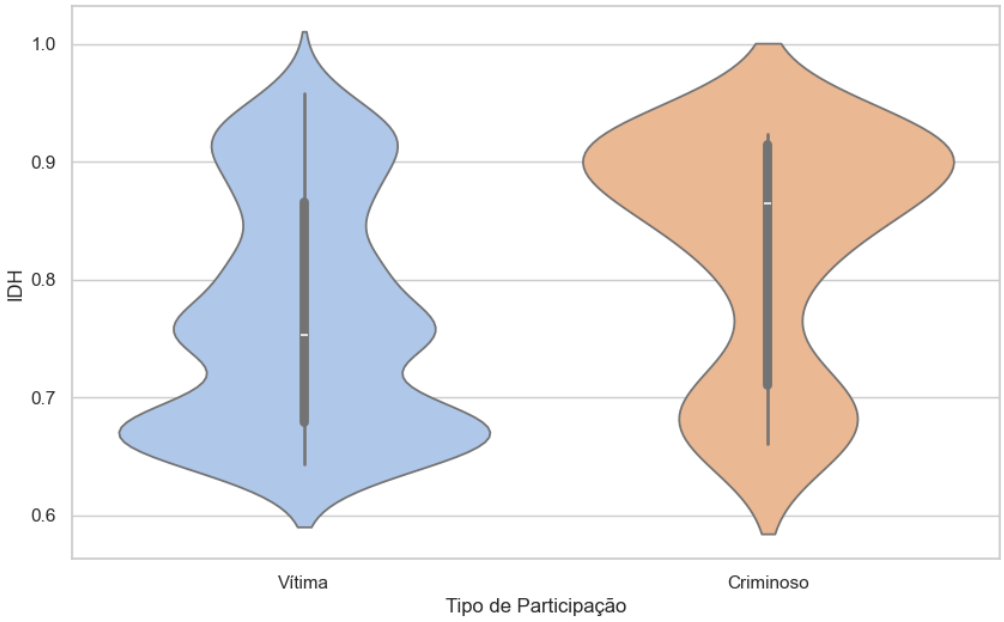
No intuito de visualizar e compreender as disparidades nas características socioeconômicas entre vítimas e criminosos nos top 10 bairros de Porto Alegre, optou-se por representar graficamente a distribuição do Índice de Desenvolvimento Humano (IDH) por meio de gráficos de violino. A escolha desse tipo de gráfico se justifica pela sua capacidade de fornecer uma representação intuitiva da distribuição dos dados, permitindo identificar padrões, assimetrias e variações na densidade das informações.

Os gráficos de violino apresentam vantagens ao comparar distribuições, pois pos-

sibilitam observar não apenas a tendência central, mas também a forma geral da distribuição. Dessa forma, tornam-se uma ferramenta valiosa para identificar visualmente diferenças nas características socioeconômicas entre os grupos analisados. Na Figura 6.2 é apresentado um gráfico de violino em que é feita a comparação entre as distribuições dos bairros mais populares entre vítimas e criminosos e seus respectivos IDHs.

	Ocorrências	Média	Maior Valor	Menor Valor	Desvio Padrão
Vítimas	1103	0.80	0.958	0.643	0.11
Criminosos	86	0.78	0.958	0.643	0.10

Figura 6.2: Distribuição de IDH separado por vítimas e criminosos



Fonte: Elaborado pelo autor

A análise dos dados revela uma clara associação entre os índices de desenvolvimento humano (IDH) e a incidência de roubos de celulares em Porto Alegre. Bairros nobres como Bela Vista, Moinhos de Vento e Três Figueiras, com IDH mais elevado, demonstram uma baixíssima prevalência de crimes. Por outro lado, Lomba do Pinheiro, Rubem Berta e Restinga, apesar de liderarem em número de vítimas, apresentam índices de IDH expressivamente mais baixos. Essa correlação sugere que condições socioeconômicas desfavoráveis podem contribuir para a vulnerabilidade dessas áreas e a suscetibilidade para ocorrências como esta.

No entanto, é importante destacar a notável disparidade entre as informações disponíveis sobre as vítimas e os criminosos, uma discrepância que pode ter implicações

significativas na análise da dinâmica dos roubos de celulares. Enquanto os dados sobre as vítimas oferecem uma visão detalhada dos locais mais propensos a incidentes, a falta de informações proporcionais sobre os criminosos limita nossa capacidade de compreender completamente os fatores subjacentes a esses crimes.

Em suma, a correlação observada entre o Índice de Desenvolvimento Humano (IDH) e as ocorrências criminais destaca a relevância de estratégias integradas que abordem tanto aspectos de segurança quanto desafios socioeconômicos. A associação das vítimas a bairros caracterizados por vulnerabilidades socioeconômicas enfatiza a necessidade de uma análise mais aprofundada dessas áreas. Uma abordagem mais abrangente e detalhada na coleta dessas informações é essencial para garantir uma análise mais precisa e informada sobre os padrões criminais, proporcionando uma base sólida para futuras intervenções e estratégias de segurança.

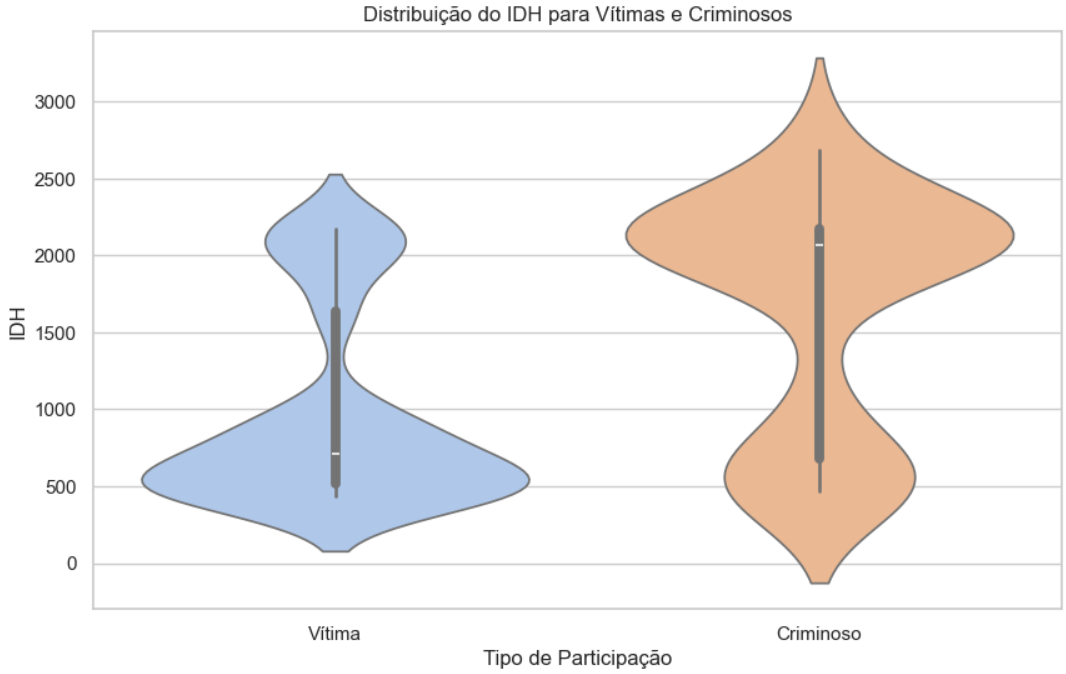
6.5.2 Renda per Capita

Para compreender as disparidades socioeconômicas entre vítimas e criminosos nos top 10 bairros de Porto Alegre, a análise se estende agora para a dimensão da renda per capita dos bairros de residência dos participantes das ocorrências criminais. A renda per capita, um componente essencial do panorama socioeconômico, desempenha um papel significativo na determinação do padrão de vida de cada região especificada.

Tabela 6.5: Descrição das variáveis para vítimas e criminosos.

	Ocorrências	Média	Maior Valor	Menor Valor	Desvio Padrão
Vítimas	1113.00 R$	1481.00 R$	3474.00 R$	435.00 R$	934.00 R$
Criminosos	107.00 R$	1749.00 R$	2000.00 R$	435.00 R$	770.00 R$

Figura 6.3: Distribuição de renda per capita separado por vítimas e criminosos



Fonte: Elaborado pelo autor

A análise revela que tanto as vítimas quanto os criminosos estão concentrados em regiões consideradas vulneráveis, caracterizadas por baixo poder aquisitivo. Bairros como Rubem Berta, Aparício Borges e Lomba do Pinheiro emergem como pontos críticos, indicando uma correlação entre a incidência de crimes e condições socioeconômicas desfavoráveis. Essa constatação ressalta a importância de abordagens integradas que não apenas foquem em aspectos de segurança, mas também busquem elevar as condições de vida e oportunidades nessas comunidades, visando uma abordagem mais abrangente para lidar com os desafios enfrentados por essas regiões vulneráveis. Além disso, vale ressaltar novamente que a análise dos criminosos pode estar sujeita a enviesamentos devido à disparidade no volume de dados em comparação com as vítimas. Portanto, ao interpretar os resultados, é fundamental considerar essas limitações e exercer cautela na generalização das conclusões sobre os criminosos.

6.6 Grau de Escolaridade e Participação em Ocorrências

A investigação das nuances relacionadas ao roubo de celulares envolve a exploração de diversas variáveis, sendo uma delas o grau de escolaridade dos envolvidos. A hipótese em foco sugere que o nível educacional pode desempenhar um papel crucial na

diferenciação entre vítimas e criminosos nesse cenário específico.

Para abordar essa hipótese, optamos por uma metodologia que utiliza o teste estatístico Qui-Quadrado. Essa abordagem tem o objetivo de avaliar a associação entre o grau de escolaridade e a participação em ocorrências de roubo de celulares, proporcionando uma compreensão mais precisa e quantitativa das relações educacionais nesse cenário.

A aplicação do teste Qui-Quadrado permite verificar se existem diferenças estatisticamente significativas nas distribuições de grau de escolaridade entre vítimas e criminosos. Este método estatístico é essencial para avaliar a independência entre as variáveis e oferecer *insights* objetivos sobre a relação proposta.

6.6.1 Teste Qui-Quadrado

Os resultados do teste Qui-Quadrado indicam que não há evidências estatísticas significativas para sustentar a ideia de que existe uma relação estatisticamente significativa entre o grau de escolaridade e o tipo de participação (criminoso ou vítima) nas ocorrências de roubo de celulares no Rio Grande do Sul. Em termos mais simples, não encontramos dados que nos permitam afirmar que o grau de escolaridade está associado de maneira estatisticamente significativa ao envolvimento de indivíduos como criminosos ou vítimas de roubo de celulares.

A "hipótese nula"neste contexto afirma que não há associação entre o grau de escolaridade e o tipo de participação em ocorrências de roubo de celulares. Ao não rejeitar essa hipótese nula, estamos indicando que os dados não fornecem suporte estatístico suficiente para concluir que há uma ligação significativa entre o grau de escolaridade e o tipo de envolvimento em crimes de roubo de celulares. Em suma, os resultados não demonstram uma correlação estatisticamente robusta entre essas variáveis específicas.

6.6.1.1 Tabela de Contingência Normalizada

	Criminoso	**Vítima**
Analfabeto	0.002625	0.016798
Fundamental	0.105774	0.378740
Médio	0.053543	0.329659
Superior	0.014961	0.097900

6.6.1.2 Estatística Qui-Quadrado (0.111)

A estatística Qui-Quadrado, que mede a diferença entre as frequências observadas e esperadas, apresenta um valor baixo (0.111). Isso sugere que as discrepâncias observadas entre as categorias de grau de escolaridade e tipo de participação são mínimas.

6.6.1.3 Valor p (0.8934)

O valor p, elevado a 0.8934, representa a probabilidade de obter resultados semelhantes ou mais extremos por acaso. Quanto mais próximo de 1, menos evidências temos para rejeitar a hipótese nula. Neste contexto, o valor p alto sugere que as diferenças observadas podem ser explicadas pelo acaso.

Esses resultados indicam que, no conjunto de dados analisado, não há uma relação estatisticamente significativa entre o grau de escolaridade e o tipo de participação (criminoso ou vítima) em ocorrências de roubo de celulares. Isso pode indicar que, pelo menos no escopo desta análise, o grau de escolaridade por si só pode não ser um fator determinante significativo para explicar as variações no tipo de participação em roubos de celulares. Outros fatores não considerados neste estudo podem desempenhar um papel mais significativo, ou há a possibilidade de que o fenômeno seja influenciado por uma complexidade de variáveis inter-relacionadas.

Esses resultados ressaltam a importância de uma análise mais abrangente e a consideração de múltiplos fatores ao investigar as dinâmicas associadas a esse tipo específico de crime no Rio Grande do Sul.

7 CONCLUSÃO

Com o intuito de avaliar a viabilidade da aplicação de técnicas de ciência de dados na abordagem do problema de roubo de celulares no estado do Rio Grande do Sul, este estudo delineou sua metodologia e conduziu diferentes experimentos para realizar uma investigação acerca do tema. Os testes foram realizados para investigar incidentes de roubo e furto de celulares que ocorreram nos últimos 5 anos no estado do Rio Grande do Sul.

Para realizar a investigação acerca dos dados foram utilizados algoritmos de aprendizado máquina não supervisionado, como o K-Médias, técnicas de visualização de dados e também métodos estatísticos clássicos. Ao longo do desenvolvimento do trabalho, surgiram hipóteses inicialmente formuladas com a expectativa de revelar padrões específicos relacionados aos casos de roubo de celulares no Rio Grande do Sul. No entanto, durante a execução da análise, tornou-se evidente que a coleta de dados apresentava limitações significativas, comprometendo a qualidade e a confiabilidade das informações disponíveis. Essas limitações resultaram em um cenário em que as hipóteses originalmente propostas não foram plenamente confirmadas, levando a resultados inesperados.

A importância desse cenário reside na conscientização sobre a relevância da qualidade dos dados na condução de análises de ciência de dados. Uma coleta inadequada de dados pode distorcer as conclusões e gerar interpretações equivocadas, impactando diretamente na validade das hipóteses formuladas. Nesse contexto, a necessidade de descartar algumas análises planejadas torna-se crucial para evitar conclusões falhas ou conclusões baseadas em dados inconsistentes.

Dentre os resultados obtidos com os experimentos, destacam-se:

- A avaliação global das características demográficas associadas a esse crime, incluindo sexo, nível educacional e estado civil.
- Agrupamento de vítimas e criminosos com base em dados demográficos.
- Compreensão da problemática social que envolve as vítimas, em relação ao Índice de Desenvolvimento Humano (IDH) e à renda per capita.

No âmbito de futuras pesquisas e trabalhos, é essencial direcionar esforços para aprimorar a coleta de dados, visando permitir análises mais específicas e abrangentes no contexto de segurança pública. Em particular, há uma lacuna notável na obtenção de informações detalhadas sobre as características físicas e dados não catalogados dos

criminosos envolvidos nos casos de roubo de celulares. Melhorar a qualidade e a extensão desses dados pode enriquecer significativamente a compreensão dos perfis criminais, possibilitando uma análise mais refinada das tendências e comportamentos associados a esse delito. A incorporação de dados mais detalhados e abrangentes contribuirá para uma pesquisa mais robusta e aprofundada, permitindo uma abordagem mais precisa e eficaz na compreensão de não só o fenômeno abordado neste trabalho de roubos de celulares, como segurança pública em geral.

REFERÊNCIAS

AUTOR, P.; AUTOR, S. Aprendizado de máquina e o estado atual e tendências. **Scientific Electronic Archives**, 2021. Available from Internet: <https://www.scielo.br/j/ea/a/wXBdv8yHBV9xHz8qG5RCgZd>.

JUPYTER Notebook. Available from Internet: <https://jupyter.org>.

MAHESH, B. **Machine Learning Algorithms Review**. 2020. Available from Internet: <https://learn.microsoft.com/pt-br/azure/architecture/data-science-process/overview>.

MATPLOTLIB. Available from Internet: <https://matplotlib.org/>.

MICROSOFT. **Microsoft Team Data Science Process (TDSP)**. 2018. Available from Internet: <https://learn.microsoft.com/pt-br/azure/architecture/data-science-process/overview>.

PANDAS. Available from Internet: <https://pandas.pydata.org>.

PEARSON. Chi-squared algorithm. 1900.

PRADO, K. H. J. **Data Science Aplicada à Análise Criminal Baseada nos Dados Abertos Governamentais do Brasil**. 2020. <https://ri.ufs.br/bitstream/riufs/14190/2/KLEBER_HENRIQUE_JESUS_PRADO.pdf>.

PYTHON SOFTWARE FOUNDATION. **Python**. 2021. Available from Internet: <https://www.python.org/psf-landing/>.

TUKEY, J. W. **Exploratory Data Analysis**. [S.l.: s.n.], 1977. [S.l.: s.n.].

VARGAS, W. A. L. de. **Data Science & Segurança Pública: Padrões Estatísticos sobre as Ocorrências de Flagrantes em Roubo de Celular na Cidade de São Paulo**. Available from Internet: <https://pesquisa-eaesp.fgv.br/teses-dissertacoes/data-science-seguranca-publica-padroes-estatisticos-sobre-ocorrencias-de>.

WASKOM, M. L. seaborn: statistical data visualization. **Journal of Open Source Software**, v. 6, n. 60, p. 3021, 2021. Available from Internet: <https://doi.org/10.21105/joss.03021>.

WIKILAI, F. S. **Transparência passiva**. 2023. Available from Internet: <https://wikilai.fiquemsabendo.com.br/wiki/Transpar%C3%AAncia_passiva>.

Yellowbrick. **Elbow Method**. 2023. Available from Internet: <https://www.scikit-yb.org/en/latest/api/cluster/elbow.html>.

(PYTHON SOFTWARE FOUNDATION, 2021; MICROSOFT, 2018; MAHESH, 2020; AUTOR; AUTOR, 2021; PEARSON, 1900; JUPYTER..., ; PANDAS, ; MATPLOTLIB, ; PRADO, 2020; VARGAS, ; TUKEY, 1977; WASKOM, 2021; WIKILAI, 2023; Yellowbrick, 2023)