Detecção de Padrões em Subtração de Celulares no Estado de São Paulo

Carlos Eduardo Ferreira, Jose Emilio de Lucena Junior
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo (USP)
03828-000, São Paulo, SP, Brasil

Abstract

Smartphones are now essential in people's daily lives for a wide range of activities, from the most mundane to entertainment and even essential equipment for the work of millions of Brazilians. Thus, the lack of these can have significant impacts on the population's quality of life and well-being. In this study, we analyzed data on smartphone thefts in the State of São Paulo and explored the use of Machine Learning techniques, more specifically Random Forest Regression, to predict thefts in a given population.

Index terms— data crime, statistical learning, Random Forest Regression

1 Introdução

Em maio de 2024, o Brasil contava com 258 milhões de *smartphones* em uso - 1,2 aparelho por habitante. Dentre os dispositivos digitais, 52% eram celulares inteligentes e 48% eram computadores: *notebook*, *desktop* e *tablet*. Para cada TV vendida, três celulares eram comercializados, tendência observada no país e no mundo (Meirelles, 2024).

Hoje os telefones celulares são amplamente utilizados pela população como principal ferramenta de trabalho, além do uso como ferramenta para estudo (Tokarnia, 2020), entretenimento e até como método de acesso a banco, meios de pagamento e identificação digital.

A subtração de telefones celulares - por furto, roubo, perda e outros - é um problema crônico de segurança pública, especialmente em grandes centros urbanos. Em São Paulo, o Estado mais populoso do Brasil, com 21,6% da população brasileira (IBGE, 2024), foram registrados mais de 318 mil boletins de ocorrência (BOs) envolvendo celulares subtraídos no ano de 2024 (SSP/SP, 2024).

Esse número ressalta a importância de entender padrões dessas ocorrências – quando e onde ocorrem com mais frequência, e quais fatores podem influenciar seu desfecho. Identificar tais padrões pode auxiliar as forças de segurança a alocar recursos de forma mais eficaz e a implementar estratégias preventivas nos locais e horários de maior risco.

Nosso trabalho explora dados de subtração de celulares, mais especificamente ocorridos no Estado de São Paulo, e aplica técnicas de aprendizado de máquina para prever, baseado em características extraídas da base, a quantidade de furtos numa dada região, em situação de normalidade ou em eventos com grande aglomeração de pessoas.

2 Trabalhos relacionados

O estudo de dados criminais e seu uso na predição ou prevenção não é algo novo e é um tema amplamente estudado envolvendo vários aspectos metodológicos, éticos, legais e sociais. Do ponto de vista prático esses estudos visam o uso potencial e aplicações como em predição de crimes, prevenção e análise criminológica.

Estudos de princípios teóricos de criminologia podem ajudar a fornecer bases no uso e influência de indicadores ou ajudar a entender a mecânica básica de crimes como por exemplo a Teoria da Atividade Rotineira (Cohen and Felson, 1979). Essa teoria acredita que a ocorrência da maioria dos crimes, especialmente os predatórios, necessita da convergência de três elementos, incluindo infratores motivados, alvos adequados e falta de capacidade de defesa no tempo e no espaço.

A Teoria da Escolha Racional (Gudjonsson, 1988) sustenta que as escolhas do infrator em termos de localização, objetivos e métodos podem ser explicadas pelo equilíbrio racional entre esforço, risco e recompensa.

A Teoria dos Padrões de Criminalidade (UN–Habitat, 2007) integra a teoria das atividades rotineiras e a teoria da escolha racional, que explica mais detalhadamente a distribuição espacial dos eventos criminais. As pessoas formam um

"mapa cognitivo" e um "espaço de atividades" por meio das atividades diárias.

De acordo com a Teoria Econômica do Crime (Becker, 1968), o agente racional/econômico decide cometer ou não um crime por meio de uma ponderação entre os benefícios e os custos dessa atividade, conforme o seguinte modelo:

$$g > p(f + \ell t) \tag{1}$$

onde:

g ganho ao praticar a atividade criminosa;

p probabilidade percebida de captura (sensação de segurança);

f multa;

ℓ desutilidade por unidade de tempo;

t tempo de cumprimento da pena.

Esse modelo considera que qualquer cidadão pode cometer um delito, como uma atividade econômica qualquer, quando o ganho (g) decorrente dessa ação for maior que o custo de punição $p(f + \ell t)$. Segundo essa teoria, o cidadão não nasce criminoso e pode agir como tal quando essas condições forem satisfeitas (Lucena Jr., 2014).

Em anos recentes o uso de métodos estatísticos e aprendizado de máquina para desenvolver modelos bem-sucedidos de previsão de crimes tem sido um tópico de pesquisa significativo. Por exemplo, Alves et al. (2018) utilizam métricas urbanas, dados sócio-econômicos e aprendizado estatístico para a predição de crimes. Já Ilgun and Dener (2025), utilizando dados criminais históricos, avaliam o desempenho de algoritmos de predição para o tipo de crime. Bediroglu and Colak (2024) combinam dados criminais, dados de sistemas geolocalizados (GIS) e técnicas de aprendizado de máquina.

3 Metodologia

3.1 Base de dados e Análise exploratória

Nossa base de dados são as ocorrências de subtração de celulares - furto, roubo, perda e outros - que aconteceram dentro do Estado de São Paulo e a nossa amostra são as que efetivamente foram registradas/notificadas em Boletins de Ocorrência nas delegacias, físicas ou eletrônicas, durante o ano de 2024. A fonte também apresenta dados para o primeiro bimestre de 2025, que podem ser incorporados ao dataset de 2024 ou utilizados como conjunto de dados de validação.

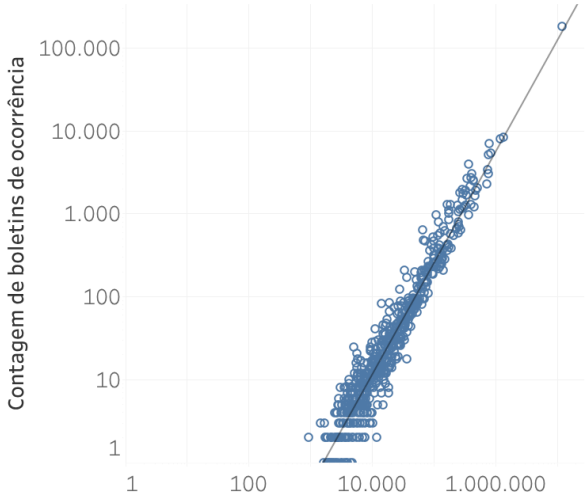

Figure 1: População municipal x boletins de ocorrência (2024).

As bases consideradas foram extraídas do site da Secretaria da Segurança Pública do Estado de SP (2025), onde, até a data de submissão desse artigo, estavam disponíveis dados de subtração de celulares de 2017 até o primeiro bimestre de 2025.

Esses conjuntos de dados foram considerados porque apresentam as notificações formais das ocorrências de subtração, com detalhamento dessas ocorrências. A base contém informações de natureza, local, data, horário, conduta, entre outros.

Como não temos a informação das ocorrências não notificadas, não sabemos o tamanho da população. Nossa amostra são as ocorrências que foram registradas, por meio de 318.265 boletins de ocorrência em 2024, ou instâncias do nosso conjunto de dados.

Os dados de cada ano são fornecidos separadamente e para esse trabalho compomos as bases para a análise exploratória.

A análise exploratória dos dados foi realizada utilizando Tableau Public (2025) e já neste momento tentamos entender como as features se relacionam e se distribuem. Na Figura 1 podemos verificar, por exemplo, a relação direta entre quantitativo populacional residente e ocorrências de furtos, com $R^2 \approx 0,9332$ e $P < 0.0001$.

Quando olhamos a Figura 2 o box-plot das ocorrências em um dia, separados por ano, podemos facilmente observar casos de outliers e uma queda na média de ocorrências a partir de 2020, período marcado pelo início da pandemia de covid. Na Figura 3 podemos ver os picos (outliers) e, pesquisando os dias ocorridos, verificamos que correspondem aos dias de Carnaval, pré-Carnaval,

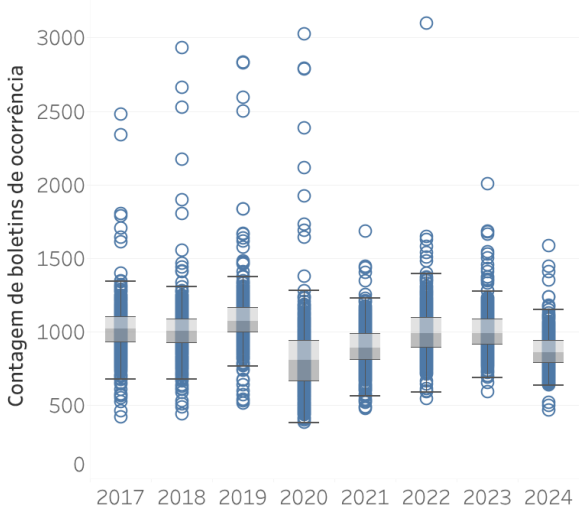

Figure 2: Gráfico de caixa - boletins de ocorrência em um dia, por ano (2017-2024).

Parada do Orgulho LGBT e outros eventos onde há aglomeração de pessoas. Estes eventos atraem milhares de pessoas todos os anos e geram alta concentração de pessoas em uma área. Em contrapartida, percebemos uma redução geral de subtração de celulares no período de férias escolares de verão, quando paulistanos deixam a capital para festas de fim de ano (Agência Brasil, 2019). Nesse período houve aumento de subtração nas cidades litorâneas, em especial na virada, onde pessoas vão para as ruas para comemorar o primeiro dia do ano. Isso nos levou a questionar se não apenas a população mas também a densidade populacional poderia afetar o total de subtrações. Veremos mais adiante que esse pressuposto não é verdadeiro quando olhamos para densidade populacional de cidades.

A Figura 4 facilita a visualização de que as datas dos eventos com maior subtração de celulares - Carnaval, pré-Carnaval, Parada do Orgulho LGBT - não coincidem ao longo dos anos, bem como o efeito das práticas de isolamento, durante os anos de pandemia.

3.2 Predição de subtração de celulares

Para o uso na criação do modelo preditivo, utilizamos os dados agregados por cidade do ano de 2024, incluindo dados de densidade populacional, que não havia na base de dados original, mas que durante a exploração acreditamos ser uma variável geográfica que pudesse ser de interesse nesse estudo. Além disso, incluímos informações sobre população, área e o número total de furtos registrados para cada município.

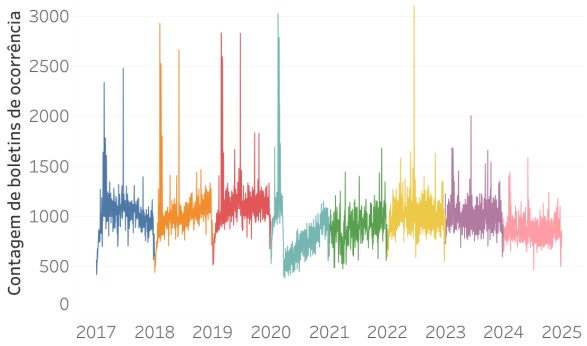

Figure 3: Gráfico de linha - boletins de ocorrência em um dia, por ano (2017-2024).

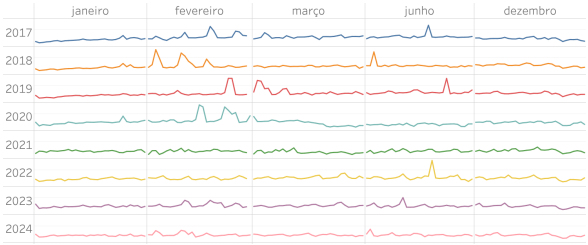

Figure 4: Gráfico de linhas (jan, fev, mar, jun e dez) - boletins de ocorrência em um dia, por ano (2017-2024).

Os dados de densidade populacional foram obtidos por meio de bases do Instituto Brasileiro de Geografia e Estatística (IBGE) (2025). O objetivo principal desta análise foi explorar as características do conjunto de dados, entender a distribuição das variáveis numéricas, agora agrupadas, e investigar as relações entre o número total de subtração de celulares e as variáveis demográficas e geográficas (População, Área, Densidade).

3.3 Pré-processamento dos Dados

O conjunto de dados agregado contém informações de 645 cidades. As estatísticas descritivas revelaram uma grande variação nos valores das variáveis numéricas, especialmente em População e Total, indicando a presença de cidades com características muito distintas (desde cidades pequenas até grandes metrópoles).

- População: Média de 71.276 habitantes, mas com desvio padrão muito alto (482.441) e valor máximo de 11.9 milhões, indicando forte assimetria à direita.
- Área: Média de 384.8 km², com variação considerável.
- Densidade: Média de 0.34 hab/km², com máximo de 13.9 hab/km², esse último indicando

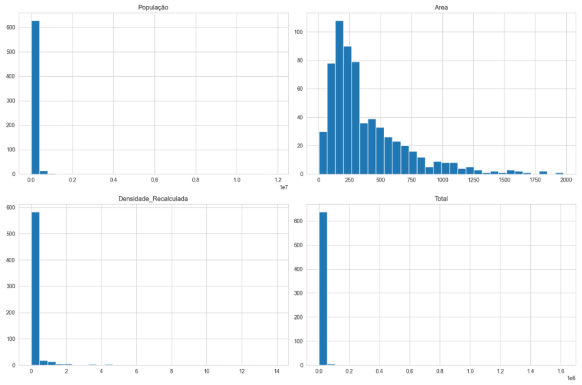

Figure 5: Distribuição das variáveis numéricas nos dados agregados por cidade.

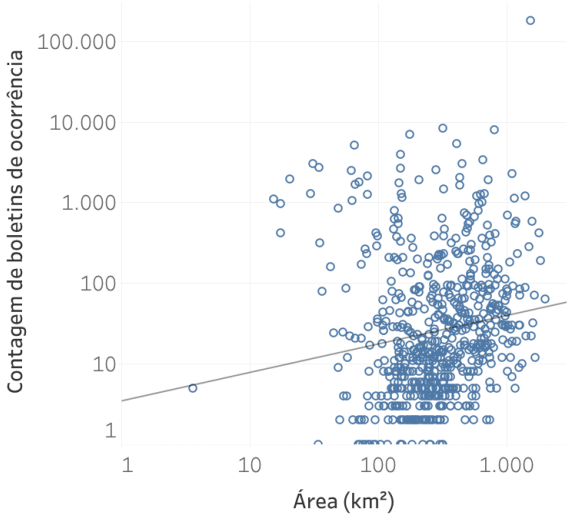

Figure 6: Relação entre total de subtração de celulares e área

uma grande cidade, densamente populada.

- Total (subtração de celulares): Média de 4.487, com desvio padrão altíssimo (64.056) e máximo de 1.6 milhão, também indicando forte assimetria e presença de outliers (provavelmente a capital)

Os gráficos de dispersão (Figura 6, 7 e 1) mostraram as seguintes relações:

- Total vs. Area (Figura 6): Não foi observada uma relação linear clara. Cidades com áreas muito diferentes podem ter totais de furtos semelhantes, e vice-versa. A maioria das cidades se concentra em áreas menores.

- Total vs. Densidade (Figura 7): Uma leve tendência positiva pode ser sugerida, mas a relação é muito menos pronunciada do que

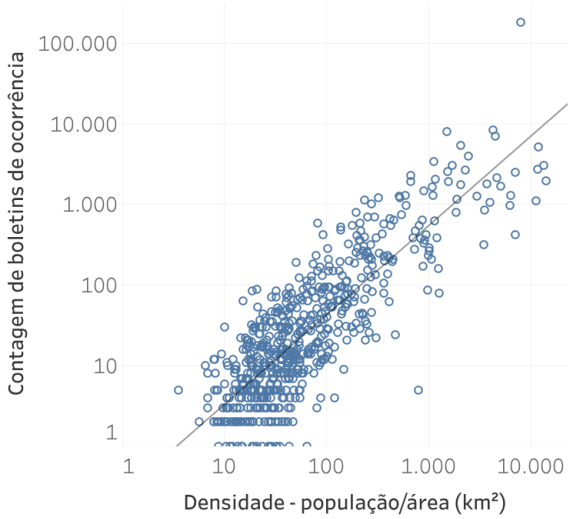

Figure 7: Relação entre total de subtração de celulares e densidade.

com a população. Há muita dispersão nos dados.

- Total vs. População (Figura 1): Uma clara tendência positiva. Cidades com maior população tendem a ter um número total de furtos significativamente maior. A relação parece ser de lei de potência para a maioria dos pontos, mas com alguns pontos (cidades muito grandes) se destacando.

A matriz de correlação (Figura 8) quantificou as relações lineares observadas:

- Total e População: Correlação de Pearson muito forte e positiva (0.99). Isso confirma que a população é o fator mais fortemente correlacionado linearmente com o número total de furtos neste dataset.

- Total e Área: Correlação muito fraca e ligeiramente positiva (0.14).

- Total e Densidade Recalculada: Correlação fraca e positiva (0.28).

- População e Densidade Recalculada: Correlação fraca e positiva (0.34).

- População e Área: Correlação muito fraca e ligeiramente positiva (0.16).

- Área e Densidade Recalculada: Correlação fraca e negativa (-0.14).

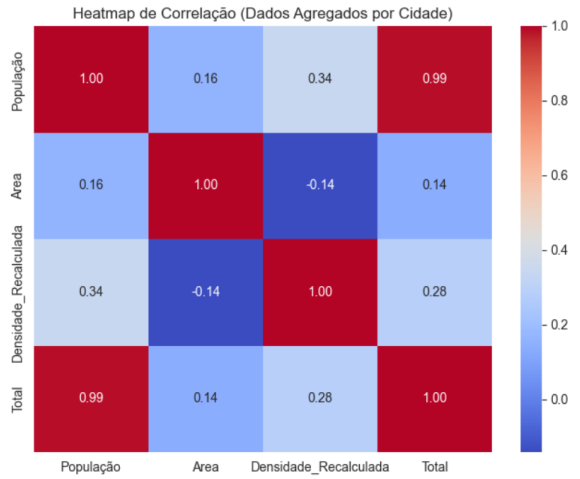

Figure 8: Heatmap da Matriz de Correlação entre as variáveis numéricas.

3.4 Modelo de Regressão

O objetivo foi construir um modelo capaz de estimar o total de subtrações com base em características demográficas e geográficas das cidades, especificamente população, área e densidade populacional.

Para o treinamento do modelo, o conjunto de dados pré-processado (645 amostras) foi dividido em três subconjuntos para treinamento e avaliação do modelo, utilizando uma semente aleatória (random state=42) para garantir a reprodutibilidade:

- Conjunto de Treinamento (60%): 387 amostras, utilizadas para treinar o modelo de regressão.
- Conjunto de Validação (20%): 129 amostras, utilizadas para avaliação intermediária e potencial ajuste de hiperparâmetros (embora não realizado nesta análise inicial).
- Conjunto de Teste (20%): 129 amostras, utilizadas exclusivamente para a avaliação final da performance do modelo treinado.

Foi selecionado o algoritmo RandomForestRegressor da biblioteca scikit-learn. Este modelo é um ensemble de árvores de decisão, conhecido por sua robustez, capacidade de capturar relações não lineares complexas e por fornecer uma métrica de importância das features.

O modelo foi treinado utilizando as seguintes features: população, área e densidade recalculada, tendo total como variável alvo. Foram utilizados os hiperparâmetros padrão do scikit-learn, incluindo n estimators=100.

Conjunto	R^2	**MAE**	**MSE**
Conjunto de Validação	0.0678	12 588.2153	18 798 519 157
Conjunto de Teste	0.8846	447.2605	2 944 045.05

Table 1: Métricas de desempenho obtidas nos conjuntos de validação e teste.

3.5 Métricas de Avaliação

A performance do modelo foi avaliada utilizando as seguintes métricas de regressão padrão:

- R² (Coeficiente de Determinação): Indica a proporção da variância na variável dependente (Subtrações) que é previsível a partir das variáveis independentes (features). Varia de menos infinito a 1, onde 1 representa um ajuste perfeito.
- MAE (Mean Absolute Error - Erro Absoluto Médio): Mede a média das diferenças absolutas entre os valores previstos e os valores reais. É expresso na mesma unidade da variável alvo.
- MSE (Mean Squared Error - Erro Quadrático Médio): Mede a média dos quadrados das diferenças entre os valores previstos e os reais. Penaliza erros maiores mais fortemente.

4 Resultados

O modelo RandomForestRegressor foi treinado no conjunto de treinamento e avaliado nos conjuntos de validação e teste. Os experimentos foram executados utilizando código python, executando em notebook jupyter sob um computador MacBook Pro, usando chipset Apple M3 Pro com 18 GB de memória. O código e dataset para reprodução dos experimentos podem ser encontrados no repositório deste projeto[1] As métricas de performance obtidas foram de acordo com a Tabela 1.

Observa-se uma diferença significativa na performance entre os conjuntos de validação e teste. O R^2 no conjunto de teste (0.8846) sugere que o modelo, com as features selecionadas, consegue explicar aproximadamente 88.5% da variância no total de furtos nesse conjunto específico. No entanto, o R² muito baixo no conjunto de validação (0.0678) indica que a performance do modelo pode ser instável e variar consideravelmente dependendo da amostra de dados utilizada para avaliação.

[1] https://github.com/eduprivate/experimentacao_am_sp_crime_prediction.

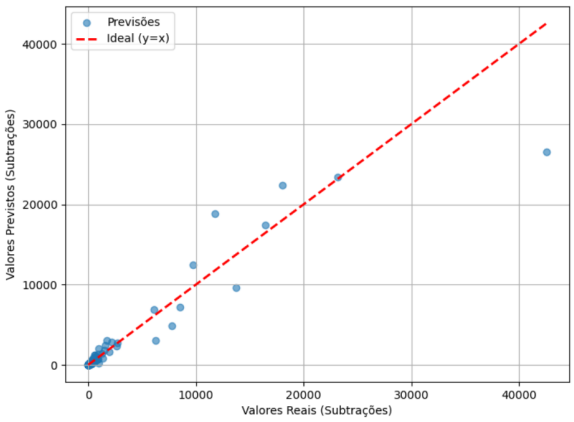

Figure 9: Comparação entre valores reais e previstos de furtos no conjunto de teste.

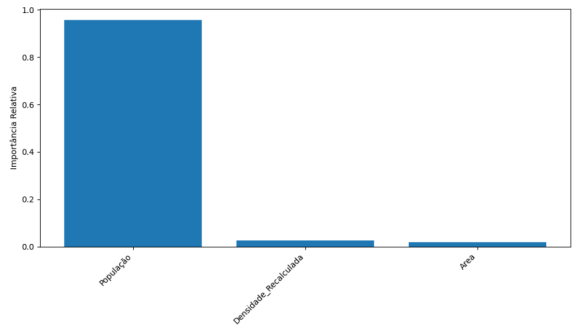

Figure 10: Importância relativa das features calculada pelo modelo RandomForestRegressor.

O gráfico de dispersão (Figura 9) compara os valores reais do total de furtos com os valores previstos pelo modelo no conjunto de testes. Idealmente, os pontos deveriam se alinhar próximos à linha diagonal (y=x). Observa-se que, para a maioria dos pontos com menor número de furtos, há uma boa aderência à linha ideal.

No entanto, para valores mais altos (cidades com muitos furtos), a dispersão aumenta, indicando maior erro de previsão nesses casos.

O gráfico de importância das features (Figura 10) mostra a contribuição relativa de cada variável de entrada (população, área, densidade recalculada) para as previsões do modelo RandomForestRegressor.

Conforme o gráfico, a feature população demonstrou ser, de longe, a mais importante para o modelo, seguida por densidade recalculada e, por último, área, que teve uma importância relativa consideravelmente menor.

Dado os resultados na validação recomenda-se a utilização de técnicas como validação cruzada (Cross-Validation) para obter uma estimativa mais robusta e confiável da performance de generalização do modelo, mitigando a dependência de uma única divisão treino-validação-teste.

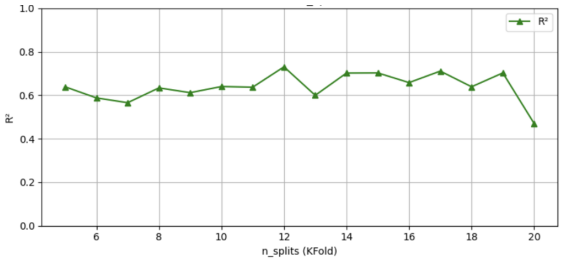

Figure 11: Influência de n splits no R².

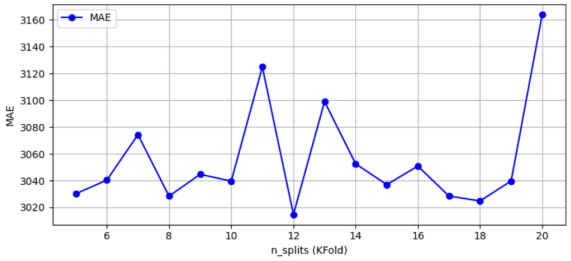

Figure 12: Influência de n splits no MAE.

4.1 Cross-Validation

Para o teste e validação usando Cross-Validation utilizando a ferramenta KFold da biblioteca scikit-learn e testamos a performance das métricas variando o valor de 5 a 20 do hiperparâmetro n splits. As Figuras 11, 12 e 13 demonstram os resultados.

Analisando os gráficos vemos que a escolha de n splits = 12 minimiza os erros MSE e MAE e obtém um R² máximo, sendo portanto, uma escolha razoável para esse conjunto de dados.

4.2 Discussão

O modelo RandomForestRegressor treinado apresentou um desempenho notavelmente bom no conjunto de teste ($R^2 \approx 0.88$), indicando que as features selecionadas, principalmente a População, têm um forte poder preditivo sobre o número total de furtos agregados por cidade, conforme os dados

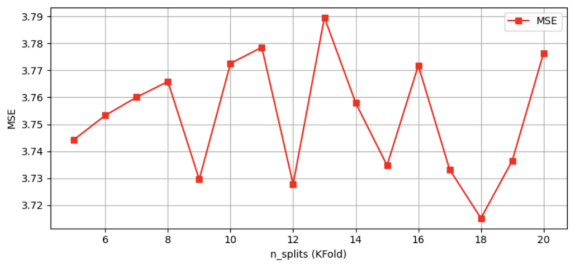

Figure 13: Influência de n splits no MSE.

e a divisão realizada.

A alta importância atribuída à População pelo modelo está alinhada com os resultados da análise exploratória anterior, que já apontava uma correlação linear muito forte (≈ 0.99) entre essas duas variáveis. Isso reforça a ideia de que o volume populacional é o principal direcionador do número absoluto de furtos em nível municipal.

A discrepância acentuada entre as métricas do conjunto de validação ($R^2 \approx 0.07$) e do conjunto de teste ($R^2 \approx 0.88$) é um ponto crítico que exige atenção. Essa diferença pode ser devida a vários fatores:

- Sensibilidade à Divisão dos Dados: A divisão específica gerada pelo random state=42 pode ter resultado em conjuntos de validação e teste com características muito diferentes, ou com a presença de outliers influentes em um conjunto mas não no outro.

- Overfitting Potencial: Embora o RandomForest seja geralmente robusto, a performance muito superior no teste em comparação com a validação poderia (embora menos comum neste padrão) indicar algum grau de ajuste excessivo às particularidades do conjunto de treino que, por acaso, se generalizou bem para este conjunto de teste específico, mas não para o de validação.

- Variância do Modelo: Modelos baseados em árvores podem ter alguma variância. A performance pode variar dependendo dos dados específicos em cada subconjunto.

Devido a esse cenário, utilizamos a técnica de validação cruzada (Cross-Validation) para obter uma estimativa mais robusta e confiável da performance de generalização do modelo, mitigando a dependência de uma única divisão treino-validação-teste.

4. Conclusão

O modelo RandomForestRegressor treinado apresentou um desempenho notavelmente bom no conjunto de teste ($R^2 \approx 0.88$), indicando que as features selecionadas, principalmente a População, têm um forte poder preditivo sobre o número total de furtos agregados por cidade, conforme os dados e a divisão realizada.

A alta importância atribuída à População pelo modelo está alinhada com os resultados da análise exploratória anterior, que já apontava uma correlação linear muito forte (≈ 0.99) entre essas duas variáveis. Isso reforça a ideia de que o volume populacional é o principal direcionador do número absoluto de furtos em nível municipal.

A discrepância acentuada entre as métricas do conjunto de validação ($R^2 \approx 0.07$) e do conjunto de teste ($R^2 \approx 0.88$) é um ponto crítico que exigiu atenção e nos levou a aplicar a técnica de validação cruzada, que nos deu resultados melhores que os testes anteriores.

Apesar dos resultados, os experimentos foram realizados utilizando apenas três características: Área, População e Densidade Populacional. Acreditamos, dada a literatura da área, que esse estudo pode ser enriquecido acrescentando outras características relevantes que podem influenciar os índices de furtos, por exemplo:

- densidade de ruas por cidade

- quantidade de pontos de interesse (shoppings, prédios, etc)

- índices macroeconômicos e de desenvolvimento

- planejamento urbano

References

Agência Brasil. 2019. Paulistanos já deixam a capital para festas de fim de ano.

Luiz G. A. Alves, Haroldo V. Ribeiro, and Francisco A. Rodrigues. 2018. Crime prediction through urban metrics and statistical learning. *Physica A: Statistical Mechanics and its Applications*, 505:435–443.

Gary S. Becker. 1968. Crime and punishment: An economic approach. *Journal of Political Economy*, 76(2):169–217.

Gamze Bediroglu and Husniye Ebru Colak. 2024. Predicting and analyzing crime—environmental design relationship via gis-based machine learning approach. *Transactions in GIS*, 28(5):1377–1399.

Lawrence E. Cohen and Marcus Felson. 1979. Social change and crime rate trends: A routine activity approach. *American Sociological Review*, 44(4):588–608.

Gisli H. Gudjonsson. 1988. The reasoning criminal: Rational choice perspectives on offending. *Behaviour Research and Therapy*, 26(3):246–287.

IBGE. 2024. Estimativas da população residente no brasil e unidades da federação com data de referência em 1º de julho de 2024. Relatório técnico, Instituto Brasileiro de Geografia e Estatística (IBGE), Rio de Janeiro. Diretoria de Pesquisas (DPE) – Coordenação de População e Indicadores Sociais (COPIS).

Esen G. Ilgun and Murat Dener. 2025. Exploratory data analysis, time series analysis, crime type prediction, and trend forecasting in crime data using machine learning, deep learning, and statistical methods. *Neural Computing and Applications*. Advance online publication.

Instituto Brasileiro de Geografia e Estatística (IBGE). 2025. Cidades e estados. Portal Cidades@ do IBGE.

José Emilio de Lucena Jr. 2014. Modelo de ising aplicado ao estudo da criminalidade. Dissertação de mestrado em ciências, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, Brasil. Versão corrigida, Programa de Pós-Graduação em Modelagem de Sistemas Complexos.

Fernando S. Meirelles. 2024. 35ª pesquisa anual do uso de tecnologia de informação nas empresas: Resultados da pesquisa. Relatório técnico, FGVcia – Centro de Tecnologia de Informação Aplicada, Escola de Administração de Empresas de São Paulo, Fundação Getulio Vargas, São Paulo. Coordenação por meio de Prof. Fernando S. Meirelles.

Secretaria da Segurança Pública do Estado de SP. 2025. Consultas — estatísticas criminais. Portal oficial da SSP-SP.

SSP/SP. 2024. Base de dados de celulares subtraídos – 2024. `https://www.ssp.sp.gov.br/estatistica/consultas`. Disponível por meio do portal da SSP-SP. Acesso em: 04 maio 2025.

Tableau Public. 2025. Tableau public. Plataforma gratuita de visualização e compartilhamento de dados na Web.

Mariana Tokarnia. 2020. Celular é a principal ferramenta de estudo e trabalho na pandemia. Dados da 3ª edição do Painel TIC Covid-19.

UN–Habitat. 2007. *Enhancing Urban Safety and Security: Global Report on Human Settlements 2007*. United Nations Human Settlements Programme, Nairobi, Kenya. Global Report on Human Settlements.