Saul Sousa da Rocha
Orientador: Glauber Dias Gonçalves

Relação entre crimes e conteúdo gerado por usuários na web: um estudo de caso baseado em pontos de interesse e tweets georreferenciados

Picos - PI
31 de julho de 2023

Saul Sousa da Rocha
Orientador: Glauber Dias Gonçalves

Relação entre crimes e conteúdo gerado por usuários na web: um estudo de caso baseado em pontos de interesse e tweets georreferenciados

Trabalho de conclusão de curso apresentado na Univerdade Federal do Piauí como parte dos requisitos necessários para a obtenção do grau de bacharel em Sistemas de Informação.

Universidade Federal do Piauí
Campus Senador Heuvídio Nunes de Barros
Bacharelado em Sistemas de Informação

Picos - PI
31 de julho de 2023

Agradecimentos

Agradeço a Deus, que me concedeu força, sabedoria e perseverança ao longo desta jornada acadêmica.

Agradeço de coração aos meus pais, Maria Juscilene de Macedo Sousa e Francisco Irene da Rocha, pelo amor incondicional, apoio constante e por serem minha fonte de inspiração. Sem o seu encorajamento e dedicação, eu não estaria realizando essa conquista.

Agradeço ao meu irmão, Samuel Sousa da Rocha, por seu apoio inabalável, incentivo e amizade. Você sempre acreditou em mim e me motivou a nunca desistir dos meus sonhos e sempre será minha inspiração.

Agradeço aos meus amigos e familiares que estiveram ao meu lado durante toda a trajetória acadêmica seja por ter dado carona de minha cidade natal até a cidade em que moro ou por seus encorajamentos, palavras de ânimo, carinho e momentos de descontração que me ajudaram a manter o equilíbrio e a perspectiva.

Gostaria também de fazer um agradecimento especial a Maria Francinete de Sousa Sá, conhecida como Neneta. Embora não esteja mais entre nós, seu amor, gentiliza e exemplo de vida continuam a me inspirar. Com certeza é um dos pilares que me fizeram ser quem sou hoje. Sinto-me abençoado por ter tido a oportunidade de conhecê-la e carregarei sua memória em meu coração.

Por fim, expresso minha sincera gratidão ao meu orientador, Glauber Dias Gonçalves, pela sua orientação valiosa, paciência e dedicação ao longo deste projeto. Suas orientações e insights foram fundamentais para o desenvolvimento e sucesso deste trabalho.

A todos que contribuíram direta ou indiretamente para a realização deste TCC, meu muito obrigado. Seu apoio e incentivo foram essenciais em cada etapa deste processo.

A espada do destino tem dois gumes. Um deles é você.

Andrzej Sapkowski

Resumo

A quantidade de dados disponíveis na Internet bate recordes a cada ano, o que inclui conteúdos diversos, em especial dos tipos texto, imagem e vídeo. Atualmente uma parte relevante desses conteúdos são gerados pelos próprios usuários dos diversos serviços disponíveis na Internet como redes sociais virtuais, mapeamento de vias públicas e comentários em portais de notícias. Tal conteúdo pode refletir características e problemas de regiões urbanas. A área de pesquisa que estuda tais problemas é conhecida como computação urbana. Controlar os índices de criminalidade é um dos desafios enfrentados nos grandes centros urbanos do mundo. Fontes alternativas de dados que explorem as características do espaço urbano e de seus habitantes podem subsidiar a análise e compreensão dos índices de criminalidade nas cidades. Neste trabalho de conclusão do curso, investigamos esse problema explorando atributos extraídos de conteúdos gerados por usuários, especificamente pontos de interesse e tweets georreferenciados, para avaliar o potencial desse tipo de dado na previsão de taxas de criminalidade por região da cidade. Nosso foco foi a cidade de São Paulo, que é a maior da América Latina e possui altos índices de criminalidade. Nesse sentido, construímos um conjunto de dados com atributos de POI e o comportamento dos usuários do Twitter relacionados aos índices oficiais de criminalidade por regiões da cidade. Analisamos essa relação com modelos de regressão baseados em aprendizagem de máquina e seus respectivos desempenhos. Nossos resultados mostram evidências de que apenas esses atributos podem predizer razoavelmente as taxas anuais de criminalidade, atingindo erros relativos abaixo de 39% das taxas oficiais e coeficientes de determinação do modelo R2 acima de 10%.

Palavras-chaves: Crime urbano, Conteúdo gerado por usuários, Pontos de interesse, Aprendizagem de Máquina.

Abstract

The amount of data available on the Internet beats records a each year, which includes different content, especially text, image and video. Currently, a relevant part of this content is generated by the users themselves. of the various services available on the Internet such as virtual social networks, mapping on public roads and comments on news portals. Such content may reflect characteristics and problems of urban regions. The research area that studies such problems is known as urban computing. Controlling crime rates is one of the challenges faced in big urban centers worldwide. Alternative data sources that explore characteristics of the urban space and its inhabitants could support the analysis and understanding of crime rates in cities. We investigate this issue by exploiting attributes extracted from user-generated content, specifically points of interest and georeferenced tweets, to assess the potential of this type of data in predicting crime rates by city region. Our focus was the city of São Paulo, which is the largest in Latin America and has high crime rates. In this sense, we built a dataset with attributes of POI and the behavior of Twitter users related to official crime rates by city regions. We analyzed this relation with regression models based on machine learning and their respective performances. Our results show evidence that these attributes alone can reasonably predict annual crime rates, reaching relative errors below 39% of the official rates and model determinations coefficients R2 above 10%.

Lista de ilustrações

Figura 1 – Diagrama da Floresta Aleatória 17
Figura 2 – Gráfico de dispersão apresentando o limite de decisão da máquina de vetores de suporte linear (linha tracejada) 17
Figura 3 – Ilustração do modelo KNN 18
Figura 4 – Categorias de crimes analisadas na cidade de São Paulo entre abril 2022 a maio 2023: (a) Ocorrências por categoria, (b-c) dez regiões (identificador) com ocorrências mais frequentes de (b) furto e roubo, (c) lesão corporal, estupro e homicídio. 22
Figura 5 – Regiões (distritos policiais) em São Paulo (SSP-SP, 2021). 23
Figura 6 – Visão geral dos dados coletados: (a) Pontos de Interesse (POI) mais frequentes e (b) regiões que acumulam o maior volume de POI, (c) regiões que acumulam maior volume *tweets* georreferenciados e seus usuários. 25
Figura 7 – Área de localização dos tweets 26
Figura 8 – Características de usuários independente da região onde postaram *tweets*: (a) anos de atividade no Twitter, (b) número de tweets postados por ano, e (c) número de pessoas seguindo e seguidores. 27
Figura 9 – Média aritmética de atributos de usuários nas regiões monitoradas: (a) anos ativos no Twitter, (b) número de tweets postados por ano, e (c) número de pessoas seguindo e seguidores. 27
Figura 10 – Correlação entre valores preditos e reais considerando o melhor modelo para cada categoria de crime e coeficientes de correlações de Pearson entre parênteses. 31

Lista de tabelas

Tabela 1 – Resumo dos trabalhos da literatura, dentre os vários discutidos nessa seção, que são mais relacionados ao propósito desse projeto 20

Tabela 2 – Desempenho dos métodos de regressão para predição da taxa de crimes entre maio 2022 a abril 2023: Floresta Aleatória (FA), *Support Vector Regression* (SVR),*Gradient Boosting Regressor* (GBR) e *K-Nearest Neighbors* (KNN). 31

Tabela 3 – Relação de quatro atributos mais importantes para predição da taxa anual de crimes. 32

Lista de abreviaturas e siglas

POI	Ponto de Interesse
UGC	Conteúdo Gerado por Usuários
GPS	Global Positioning System
FA	Floresta Aleatória
SVR	Support Vector Regressor
GBR	Gradient Boosting Regressor
KNN	K-Nearest Neighbors
API	Application Programming Interface
OSM	Open Street Maps
MRE	Mean Relative Error
RAE	Mean Absolute Error
R2	Coeficiente de Determinação
CDF	Função de Distribuição Acumulada

Sumário

1 Introdução 10
1.1 Objetivos 12

2 Referencial Teórico 13
2.1 Computação Urbana 13
2.2 Modelos de Predição 14
2.2.1 Minimização de Erros 15
2.2.2 Aprendizagem de Máquina 15
2.2.3 Regressão via Aprendizagem de Máquina 16

3 Trabalhos Relacionados 19

4 Metodologia 21
4.1 Índices de Crimes Oficiais 21
4.2 Pontos de Interesse (POI) 23
4.3 Tweets Georreferenciados 24
4.4 Configurações 28

5 Resultados 30
5.1 Análise de Desempenho 30
5.2 Atributos Relevantes 32

6 Conclusão 33

7 Publicações 34

Referências 35

1 Introdução

Redução e controle de taxas de criminalidade são desafios enfrentados nos grandes centros urbanos do mundo. Esta é uma questão ainda mais grave nos países em desenvolvimento. No Brasil, por exemplo, foram registradas 40.804 mortes por crimes violentos em 2022, contra 41.160 em 2021, 356 mortes a menos (NEV-USP, 2022). A taxa de mortes decorrentes de crimes violentos por 100 mil habitantes varia de 1,41 a 1,69 no cenário nacional, enquanto nos estados do Nordeste essa taxa chega a mais de 3 mortes.

A pesquisa em computação urbana tem ganhado foco nos últimos anos em razão do crescimento das fontes públicas de dados na web que contém informações de tempo e espaço (SILVA et al., 2019). Da mesma forma a evolução de técnicas de computação para extrair esses dados estão contribuindo para o surgimento de novos avanços científicos na área. Não somente estes fatores, mas também em razão do surgimento de diversos serviços para as cidades como mobilidade urbana, alimentação e monitoramento dos diversos tipos problemas cotidianos dos locais urbanizados como poluição, doenças infecciosas, corrupção, e em especial criminalidade.

Fontes alternativas de dados que retratam características do espaço urbano e de seus habitantes podem contribuir para a análise e compreensão de problemas das cidades como taxas de criminalidade. Conteúdo gerado por usuários na web (UGC, do inglês *user-generated content*) é uma fonte crescente de informações que reflete a interação das pessoas com o espaço na vida cotidiana como várias pesquisas já evidenciaram (CHENG et al., 2011; AHMED; HONG; SMOLA, 2013; KOTZIAS; LAPPAS; GUNOPULOS, 2016). O volume de dados produzidos nas cidades vem aumentando à medida que os dispositivos móveis dotados de GPS (*global positioning system*) se tornam mais integrados à vida cotidiana.

Dois tipos de UGC que se destacam por capturar a interação das pessoas com espaço via GPS são pontos de interesse (POI) e mensagens georeferenciadas no *Twitter* (*tweets georreferenciados*). POI são disponibilizados em serviços de mapeamento urbano como *Open Street Maps (OSM)*, *FourSquare* e *Google Maps* contendo informações sobre um local da cidade como categorização por atividades desenvolvidas, popularidade e comentários (WEISBURD; GROFF; YANG, 2012; YUAN; ZHENG; XIE, 2012; WANG et al., 2021). Similarmente, *tweets georreferenciados* mostram resultados promissores em descrever a complexa estrutura social e espacial das cidades (TUCKER et al., 2021; IRANMANESH; ATUN, 2020). Isso inclui, em especial, os padrões comportamentais dos cidadãos e possibilidades de identificar suas atividades e espaços frequentados. Logo, POI e *tweets* georreferenciados também são potencialmente úteis para estudos sobre crimes, visto que características de um local e padrões de comportamentos das pessoas que nele convivem, podem indicar ocorrências de alguns tipos de crimes.

A comunidade científica de computação, vem explorando UGC disponíveis publicamente via serviços Web para predição de crimes. Em (WANG et al., 2019) e (BELESIOTIS; PAPADAKIS; SKOUTAS, 2018) foi mostrado que POI e metadados de fotos combinados com dados demográficos oficiais de Nova York e Londres possibilitaram a predição das taxas criminais por regiões dessas cidades. Em (HUANG et al., 2018), dados de crimes oficiais da cidade de Nova York foram incrementados com POI para prever se uma categoria de crime acontecerá numa região num tempo futuro. Em (CASTRO; RODRIGUES; BRANDAO, 2020) também foi observado que dados não oficiais, extraídos do serviço Web brasileiro "Onde Fui Roubado", adicionados às fontes oficiais melhora significativamente a predição de índices de criminalidade.

Todos esses trabalhos são baseados em fontes de dados oficiais e usam UGC, especialmente POI, apenas como um incremento de informação. Contudo, há ainda questões sobre o potencial de UGC a serem investigadas. Especificamente, até que ponto o uso desse tipo de dado *unicamente* pode refletir os índices de criminalidade das cidades? Isso envolve uma questão mais ampla que é entender quão representativo o espaço físico é o cyberespaço da web associado a cidades de países em desenvolvimento da América Latina como o Brasil, regiões essas ainda pouco exploradas em estudos nesse contexto. A investigação dessa questão é importante porque em uma eventual falta ou atraso na coleta de dados oficiais, UGC poderia oferecer indicações ou estimativas aproximadas de índices de crimes por região da cidade para orientar governos. No entanto, é necessário conhecer o nível de erro desses dados para explicar o contexto específico de taxas de crimes.

Nesse trabalho investigamos essa questão com a utilização atributos extraídos de POI do *OSM* e *tweets* georreferenciados para avaliar o potencial desse tipo de dado na predição de índices de crimes por regiões da cidade. Nosso foco foi a cidade de São Paulo, que é a maior da América Latina e concentra portanto altos índices de criminalidade. O estado de São Paulo é um dos poucos que disponibiliza publicamente índices de crimes por categoria e região (distrito policial) de todas as suas cidades mensalmente (SSP-SP, 2021). Adicionalmente, São Paulo concentra um vasto número de POI e tweets georefenciados devido a sua importância econômica a nível internacional, se tornando a cidade com condições propícias para essa investigação. Conduzimos essa investigação baseada em oitenta e dois distritos policiais, unidades de espaço, onde relacionamos as ocorrências de crimes com atributos de UGC em cada unidade. Para quantificar essa relação utilizamos modelos de aprendizagem de máquina para regressão e análises de erros desses ao inferir taxa anual de categorias de crimes mais frequentes na cidade.

Nossos resultados indicam que atributos extraídos de UGC *unicamente* podem predizer razoavelmente taxas de algumas categorias de crimes. Por exemplo, o melhor modelo de regressão avaliado (floresta aleatória) pode predizer furtos com erro médio de 29% relativo à taxa anual oficial, sendo que UGC do tipo POI, em especial pontos de taxi e estacionamentos, são os atributos mais relevantes para tal predição. Outras categorias de

crimes mais comuns em regiões periféricas como homicídios, ainda podem ser explicados via atributos de UGC apesar da menor quantidade de dados para a construção de modelos, o que leva a erros médios de até 52% em relação à taxa oficial. Contudo, um total de quatro categorias de crimes em São Paulo (estupro, lesão corporal, roubos e homicídios) podem ser preditos com os modelos, considerando coeficientes de determinação (R2) superiores a 10% e média de erros relativos (MRE) inferiores 39% da taxa oficial.

1.1 Objetivos

Este trabalho tem por objetivo investigar o potencial de conteúdo gerado por usuário na Internet, em particular nos serviços *Twitter* e *Open Street Maps*, para predizer índices de crimes por regiões urbanas com níveis de acurácia e precisão adequados. Os objetivos específicos compreendem:

1. Desenvolver métodos de coleta de dados da Internet, especificamente, conteúdo gerado por usuários em serviços web como redes sociais virtuais;

2. Obter um conjunto de dados UGC georreferenciados utilizados para extrair atributos de POI e comportamento de usuário em redes sociais virtuais, i.e., atributos do cyberspaço web, que estão associados a taxas de crimes oficiais em regiões do espaço físico real na cidade de São Paulo.

3. Realizar análises para identificar o potencial desses atributos na predição (i.e., explicação) de taxas de criminalidade, assim como identificar os atributos mais relevantes para tal predição, explorando regiões da cidade em baixa granularidade e cinco categorias de crime.

2 Referencial Teórico

Nesta Seção são descritos os conceitos fundamentais para o entendimento deste trabalho.

2.1 Computação Urbana

A computação urbana é a aquisição, integração e análise de um grande volume de dados. Esses dados podem ser encontrados em vários formatos e são gerados por uma variedade de fontes em áreas urbanas. Sensores, dispositivos, veículos, software e até mesmo humanos podem ser usados como fontes. Os dados extraídos dessas fontes são utilizados com o objetivo de melhorar a qualidade de vida de quem mora em regiões metropolitanas. Essa melhoria é atribuída ao tratamento de questões urbanas como poluição do ar, aumento do uso de energia, escassez de água, congestionamentos de trânsito, entre outros (ZHENG et al., 2014).

Em um sentido mais amplo, a computação urbana visa auxiliar no entendimento e previsão de fenômenos urbanos, além de contribuir para o planejamento do futuro das cidades. Dessa forma, pode ser descrito como uma área de estudos interdisciplinar decorrente de paralelos entre a informática e campos tradicionais como economia, sociologia e transporte no contexto dos espaços urbanos. A computação urbana está relacionada às áreas de redes de computadores, redes veiculares, sensoriais, distribuídas, sistemas cooperativos, interação humano-computador, inteligência artificial e redes sociais no campo da ciência da computação (SILVA et al., 2019).

Nesse trabalho foram utilizadas técnicas para monitoramento e coleta de conteúdo gerado por usuários (UGC) baseadas nos fundamentos de computação urbana. Especificamente, foi utilizado técnicas extração de (i) pontos de interesse (POI) de vias urbanas e (ii) comportamento de usuários em redes sociais virtuais associados a um espaço físico (região) da cidade.

Inicialmente, abordamos os Pontos de Interesse (POIs) gerados por pessoas e disponibilizados em serviços Web. Em seguida, analisamos os perfis de usuários da rede social virtual *Twitter*, coletando comentários postados pelos usuários e extraindo suas características de perfil. Tanto as informações dos POIs quanto dos perfis de usuários serão utilizadas para aprimorar a acurácia e especificidade dos modelos de predição de crimes por regiões de uma cidade. Esses assuntos estão detalhados nas seções 4.2 e 4.3.

Tweets georreferenciados são um tipo de UGC com informações de localização que mostram resultados promissores em relacionar estruturas sociais e espaciais das cidades. No estudo realizado em (LEE; WAKAMIYA; SUMIYA, 2013), os autores propõem mo-

delos para identificar padrões comportamentais dos cidadãos no espaço urbano por meio de *tweets* assim como os seus tipos de espaços tanto no domínio privado como no público.

Em (IRANMANESH; ATUN, 2020) foi analisado a situação oposta, os autores propõem uma abordagem para caracterizar o espaço real de uma cidade e seus habitantes para predizer o volume de *tweets* georreferenciados ou não que podem ser gerados em regiões específicas. Por sua vez, os autores em (KOTZIAS; LAPPAS; GUNOPULOS, 2016) utilizaram *tweets* georreferenciados para desenvolver e avaliar um arcabouço de localização de usuários baseado em conteúdo de *tweets* postados e laços sociais dos usuários.

Dentre os tipos de UGC com informações geográficas destaca-se pontos de interesse (POI) providos pelos serviços de mapeamento urbano. Em (YUAN; ZHENG; XIE, 2012) os autores mostraram que o uso de informações categóricas de POI são úteis para traçar o perfil de atividades que caracterizam bairros. Mais recentemente, em (WANG et al., 2021) foi proposto um arcabouço de métodos para identificar as características funcionais de uma determinada região em uma cidade com base em POI dessas áreas, utilizando o serviço OSM para a extração de POI. Desde então essa área vem evoluindo e conteúdos gerados por usuários (UGC) via serviços de localização baseados em redes sociais é um tipo de dado que contribuiu para consolidar análises de crimes em pequena granularidade por regiões da cidade (SILVA et al., 2019).

Os autores em (MUELLER et al., 2017) investigaram se *checkins* de usuários em regiões de POI podem ser usados para avaliar preferências de gênero por locais em diferentes regiões urbanas no mundo físico. Em (SILVA et al., 2017) é apresentada uma técnica para identificar POI e, com base neles, reconhecer pontos turísticos em uma região.

O objetivo desse trabalho é investigar como os dados de POIs e perfis de usuários podem contribuir para melhorar a precisão dos modelos de predição de crimes, permitindo um entendimento mais abrangente das regiões mais propensas a ocorrências criminais. Ao considerar as características dos POIs e dos usuários do *Twitter*, esperamos aperfeiçoar as previsões e, consequentemente, fornecer informações valiosas para o planejamento e implementação de estratégias de segurança mais eficazes.

2.2 Modelos de Predição

Neste trabalho, os modelos de predição utilizados são baseados em regressão, que pode ser realizada através da minimização de erros ou da aplicação de técnicas de aprendizado de máquina. Esses modelos estatísticos buscam estimar as relações entre uma variável dependente (também conhecida como variável-alvo) e uma ou mais variáveis independentes (também chamadas de variáveis preditoras) (YAN; SU, 2009). É importante destacar que os métodos de regressão evidenciam correlações entre uma variável dependente e um conjunto de variáveis independentes em um determinado conjunto de dados. No entanto, esses métodos não abordam questões de causalidade entre essas correlações, ou seja, se

uma variável causa diretamente a outra.

A seguir são descritas as técnicas de regressão via minimização de erros e via aprendizagem de máquina. Focaremos nessa segunda técnica, que embora mais complexa geralmente provê melhor desempenho de predição para diferentes tipos de dados.

2.2.1 Minimização de Erros

A análise de regressão é normalmente utilizada por duas razões conceituais distintas. A primeira é que usa-se frequentemente para predição, ou seja, encontrar relações entre variáveis independentes e dependentes. A segunda que em algumas situações, a análise de regressão pode ser usada para previsão, identificar tendências futuras quando se observa relações entre variáveis que têm padrões temporais ou ciclos. Regressão é um dos métodos estatísticos mais comuns usados em diversas áreas científicas como medicina, biologia, economia, engenharia, sociologia, geologia, etc (YAN; SU, 2009).

Existem três tipos de regressão. O primeiro é o linear simples, usado para modelagem da relação linear entre duas varáveis. O segundo tipo é o de regressão multilinear, na qual possui uma varável dependente e uma ou mais variáveis independentes. E o por último, o modelo de regressão não linear. O mesmo assume que a relação entre a variável dependente e variáveis independentes não é linear nos parâmetros de regressão (YAN; SU, 2009). Este modelo é mais complexo que o linear em termos de estimativa de parâmetros do modelo.

2.2.2 Aprendizagem de Máquina

A aprendizagem de máquina é o estudo de sistemas de computação que podem se aprimorar automaticamente com base na experiência e nos dados (MITCHELL, 1997). Concentra-se no emprego de algoritmos para descobrir padrões em dados frequentemente densos e complexos (BZDOK, 2017). As técnicas de aprendizagem de máquina são utilizadas em geral de duas formas: regressão ou classificação. A primeira consiste em predizer valores de uma variável dependente em função de variáveis ou características dependentes, conforme acima discutido. A segunda forma consiste em identificar a qual classe uma instância de dado pertence, ou seja, classificação de dados considerando duas ou mais classes.

Os métodos da aprendizagem de máquina podem ser organizados em supervisionado, semi-supervisionado ou não supervisionado. Os métodos de aprendizado supervisionado é caracterizado pelo uso de conjuntos de dados rotulados para treinar algoritmos para identificar adequadamente os dados ou prever resultados para regressão ou classificação. Métodos não supervisionado, analisa e agrupa conjuntos de dados (*clusterização*) não rotulados usando métodos algoritmos como *K-Means*. Métodos semi-supervisionado fornecem um meio-termo entre aprendizado supervisionado e não supervisionado. Em suma esse método é utilizado para orientar a categorização e a extração de recursos de um

conjunto de dados maior e não rotulado durante o treinamento usando um conjunto de dados rotulado menor[1].

Nesse trabalho *planejamos* o uso de aprendizagem de máquina para regressão com a abordagem de aprendizagem supervisionada. Nesse sentido avaliaremos, a princípio, os métodos floresta aleatória (FA), *support vector regression* (SVR), *Gradient Boosting Regressor* (GBR) e *K-Nearest Neighbors* (KNN). Para fazer uso dos modelos acima mencionados, utilizaremos a biblioteca *Scikit Learning Python*. Pedregosa et al. (2011a) conceitua Scikit Learning como "Uma biblioteca de aprendizado de máquina de código aberto que oferece suporte ao aprendizado supervisionado e não supervisionado. Ele também fornece várias ferramentas para ajuste de modelos, pré-processamento de dados, seleção de modelos, avaliação de modelos e muitas outras utilidades".

2.2.3 Regressão via Aprendizagem de Máquina

Agora abordaremos o uso de técnicas de aprendizagem de máquina aplicadas à tarefa de regressão mencionadas acima. Serão explicados cada um desses métodos, suas características distintas e seus respectivos benefícios em lidar com diferentes cenários de regressão.

O primeiro método é chamado Floresta aleatória, também conhecida como floresta de decisões aleatórias, é um método de aprendizado de conjunto utilizado para tarefas de classificação, regressão e outras análises [2]. Esse método opera criando várias árvores de decisão durante o processo de treinamento. Para tarefas de regressão, a saída é a média ou a previsão média das árvores individuais como mostra a figura 1.

O objetivo principal das florestas de decisão aleatórias é combater o fenômeno conhecido como *overfitting* (ou superajuste). O *overfitting* ocorre quando as árvores de decisão se adaptam muito bem ao conjunto de treinamento específico em que foram construídas, tornando-se excessivamente especializadas nesses dados. Como resultado, essas árvores podem não generalizar adequadamente para novos dados desconhecidos, levando a previsões menos precisas em situações reais.

Ao criar várias árvores de decisão durante o processo de treinamento e combiná-las para obter uma previsão final, as florestas aleatórias reduzem significativamente o risco de sobreajuste. Esse método incorpora aleatoriedade tanto na seleção de subconjuntos de dados usados para treinar cada árvore, como também nas variáveis consideradas em cada divisão do nó durante a construção das árvores. Essas escolhas aleatórias ajudam a diversificar as árvores e tornar o modelo mais robusto, melhorando a generalização e, consequentemente, a precisão das previsões em dados não observados.

O segundo modelo corresponde a uma variação das *Support Vector Machines* (SVMs), que são modelos de aprendizado supervisionado com algoritmos associados utilizados para

[1] Machine Learning; IBM. Disponível em: https://www.ibm.com/cloud/learn/machine-learning
[2] Random Forest; Wikipedia. Disponível em: https://www.wikipedia.org/wiki/Random-forest

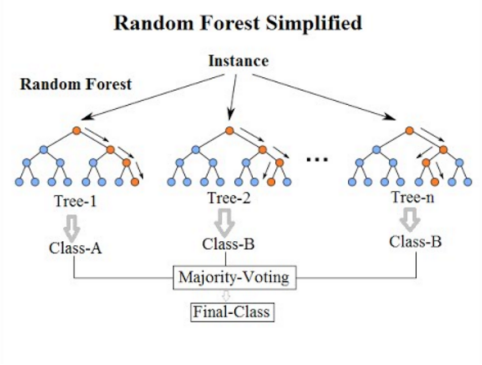


Figura 1 – Diagrama da Floresta Aleatória

classificação e análise de regressão. Essa variante é conhecida como *Support Vector Regression* (SVR).

Diferente dos algoritmos de regressão tradicionais, o SVR é uma adaptação do método de Support Vector Machine (SVM) para problemas de regressão. Ele opera mapeando os dados de entrada em um espaço de maior dimensão, buscando encontrar um hiperplano que se ajuste melhor aos pontos de dados, com o objetivo de maximizar a margem entre esses pontos e o hiperplano, como mostra a figura 2. Essa abordagem torna o SVR particularmente valioso em situações em que os dados não apresentam uma relação linear simples, permitindo que ele lide com problemas mais complexos e se beneficie de uma maior flexibilidade em suas previsões [3].

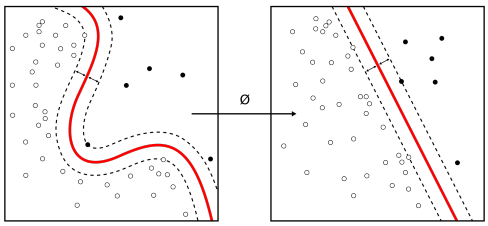


Figura 2 – Gráfico de dispersão apresentando o limite de decisão da máquina de vetores de suporte linear (linha tracejada)

O terceiro modelo a ser abordado é o GBR (Gradient Boosting Regression - Regressão por Impulsionamento de Gradiente). O GBR é uma técnica de aprendizado de máquina baseada em árvores de decisão que também pertence à família dos métodos de aprendizado de conjunto. O processo de treinamento do GBR envolve a criação iterativa de árvores de decisão, onde cada nova árvore é ajustada para corrigir os erros cometidos pelas árvores anteriores. Isso é feito através do conceito de gradiente descendente, onde os erros residuais são minimizados a cada iteração. O GBR é conhecido por ser um método poderoso, capaz de lidar com relações complexas entre as variáveis e produzir previsões precisas[4].

[3] Suport Vector Machine; Wikipedia. Disponível em: https://www.wikipedia.org/wiki/Support-vector-machine

[4] Gradient Boosting; Wikipedia. Disponível em: https://www.wikipedia.org/wiki/Gradient-boosting

Por último, temos o KNN (K-Nearest Neighbors - K-Vizinhos Mais Próximos), que é um método de aprendizado supervisionado aplicável tanto para classificação quanto para regressão. A essência do KNN é fundamentada na premissa de que pontos de dados semelhantes possuem rótulos ou valores similares. Ao receber um novo ponto de dados, o algoritmo busca pelos "k"pontos mais próximos a ele no conjunto de treinamento, como ilustrado na Figura 3, e, em seguida, realiza a previsão com base nos rótulos ou valores desses pontos vizinhos. Para problemas de regressão, a previsão é geralmente calculada como a média ou a mediana dos valores dos vizinhos selecionados [5].

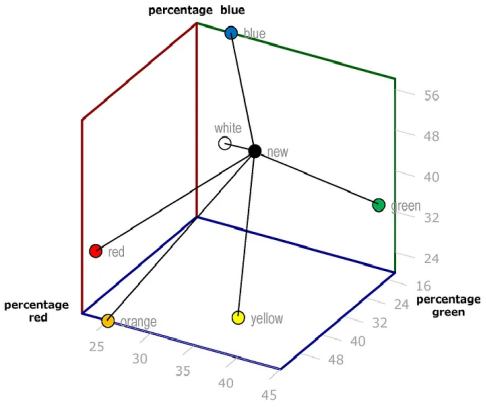


Figura 3 – Ilustração do modelo KNN

Essa abordagem de aprendizado é conhecida por sua facilidade de implementação e interpretação, o que o torna uma escolha popular em problemas de regressão, especialmente quando o relacionamento entre as variáveis não pode ser adequadamente modelado por métodos lineares ou técnicas mais complexas. Contudo, é importante ressaltar que o desempenho do KNN pode ser afetado em conjuntos de dados com muitas instâncias ou dimensões, devido ao alto custo computacional envolvido no cálculo das distâncias. Assim, ao utilizar o KNN, é crucial considerar o valor adequado de "k"e o tamanho do conjunto de dados para obter resultados mais precisos e eficientes em tarefas de regressão.

[5] K-Nearest Neighbors; Medium. Disponível em: https://www.medium.com/brasil-ai/knn-k-nearest-neighbors-1-e140c82e9c4e

3 Trabalhos Relacionados

Há alguns anos, pesquisadores das áreas de ciências sociais e urbanismo têm investigado a relação entre crimes, características das populações e espaço geográfico (censo demográfico, dados de mobilidade urbana, estatísticas sociais e outros). Estudos clássicos nessa área como (MASI et al., 2007) investigam o quanto questões raciais e o grau de violência influenciam nos resultados de gravidez, enquanto em (BECKER; KASSOUF, 2017) foi analisado se o gasto público do governo em educação impacta na redução da taxa de homicídios.

Mais recentemente, em (ADORNO; NERY, 2019; NERY; SOUZA; ADORNO, 2019) foram analisadas a trajetória de crimes na cidade de São Paulo ao longo dos anos e a sua distribuição geográfica, confrontando teorias que definem de forma estática bairros violentos e não-violentos, regiões centrais e periferia. Já em (WANG et al., 2020), foi abordada a influência de fatores demográficos como geografia, economia, educação, habitação, urbanização e estrutura populacional nas taxas de crimes.

Há também uma linha de especialistas na área de criminologia cujo foco é analisar crimes por regiões de uma cidade, i.e., unidades geográficas menores e específicas. Em (WEISBURD; GROFF; YANG, 2012) foi realizado um trabalho baseado no histórico de 16 anos de crimes nas cidades de Seattle e Washington nos EUA, se tornando uma referência seminal sobre esforços de pesquisa em criminologia com análises de crimes por regiões. Já em (SOUSA; FEITOSA; GONÇALVES, 2021a), investiga a relação entre índices de criminalidade e características refletidas em pontos de interesse (POIs) que as pessoas registraram em um serviço Web de mapeamento para a cidade de São Paulo.

Mais relacionado ao nosso trabalho estão os estudos que mesclam fontes de dados oficiais, i.e., censo e outros dados populacionais providos por órgãos de governo, com atributos UGC para predição de crimes em regiões urbanas. Nesse sentido, os trabalhos em (WANG et al., 2017; BELESIOTIS; PAPADAKIS; SKOUTAS, 2018; HUANG et al., 2018) investigaram o quanto a adição de POI, coletados em serviços de mapeamento (e.g., Google Map, Open Street Map, Foursquare) incrementam dados oficiais (e.g, dados demográficos de senso) e fluxos urbanos como táxis e ônibus para melhorar a predição das taxas criminais nas cidades de Londres, Nova York e Chicago. Já os autores em (CASTRO; RODRIGUES; BRANDAO, 2020) observaram o quanto atributos UGC extraídos do serviço web "Onde fui Roubado"adicionados a dados oficiais melhorava a predição de índices de criminalidade em regiões da cidade de Rio de Janeiro no Brasil, onde esse serviço é mais popular.

Outros trabalhos focam em mesclar fontes de dados oficiais explorando *tweets* georreferenciados. Em (TUCKER et al., 2021), atributos adicionais para predição de crimes foram extraídos a partir da contagem de usuários caracterizados como moradores, turis-

tas ou passageiros de uma região onde postaram *tweets* georreferenciados na cidade de Boston. Os autores em (VOMFELL; HÄRDLE; LESSMANN, 2018) analisaram predição de crimes violentos e crimes contra propriedades em intervalos de semana na cidade de Nova York com modelos de regressão e aprendizagem de máquina utilizando atributos de *tweets* georreferenciados com dados do censo, fluxos de taxis e POI. Outra pesquisa foi realizada em (WILLIAMS; BURNAP; SLOAN, 2017), na qual modelos de regressão linear foram utilizados para avaliar o impacto de atributos extraídos de *tweets* postados em Londres sob atributos de censo demográficos para predição de crimes organizados em nove categorias em regiões da cidade.

Em contraponto a todos esses trabalhos, nesse trabalho investigamos o quanto atributos extraídos de UGC *unicamente* podem predizer, i.e., explicar, taxas criminais por regiões das cidades. Nossas análises focam em UGC georreferenciado e estatísticas de crimes que coletamos em São Paulo, a maior cidade da America Latina, e para nosso conhecimento, ainda não estudada em trabalhos de predição de crimes com UGC.

Na tabela 3 selecionamos os trabalhos relacionados a esse trabalho. Nas cinco colunas apresentadas, resumimos as principais técnicas, se naquele trabalho foram utilizados dados de pontos de interesse, dados georreferenciado em redes sociais virtuais e se a predição é baseada apenas em UGC comparados à proposta desse projeto.

Tabela 1 – Resumo dos trabalhos da literatura, dentre os vários discutidos nessa seção, que são mais relacionados ao propósito desse projeto

Trabalho Relacionado	**Análises**	**POIs**	**Tweets Georreferenciados**	**Predição Baseada em UGC**
Belesiotis, Papadakis e Skoutas (2018)	Regressão e Caracterização.	Sim	Não	Não
Becker e Kassouf (2017)	Caracterização.	Não	Não	Não
Adorno e Nery (2019)	Caracterização.	Não	Não	Não
Wang et al. (2020)	Regressão e Caracterização.	Não	Não	Não
Weisburd, Groff e Yang (2012)	Regressão.	Não	Sim	Não
Tucker et al. (2021)	Analise de variância e Classificação.	Não	Sim	Não
Wang et al. (2017)	Regressão e Caracterização.	Sim	Não	Não
Belesiotis, Papadakis e Skoutas (2018)	Regressão e Caracterização.	Sim	Não	Não
Huang et al. (2018)	Regressão e Caracterização.	Sim	Não	Não
Sousa, Feitosa e Gonçalves (2021b)	Regressão e Caracterização.	Sim	Não	Não
Vomfell, Härdle e Lessmann (2018)	Regressão.	Sim	Sim	Não
Williams, Burnap e Sloan (2017)	Regressão.	Não	Sim	Não
Sousa, Feitosa e Gonçalves (2021a)	Regressão e Caracterização.	Sim	Não	Sim
Este Trabalho	Regressão e Análise Estatística.	Sim	Sim	Sim

4 Metodologia

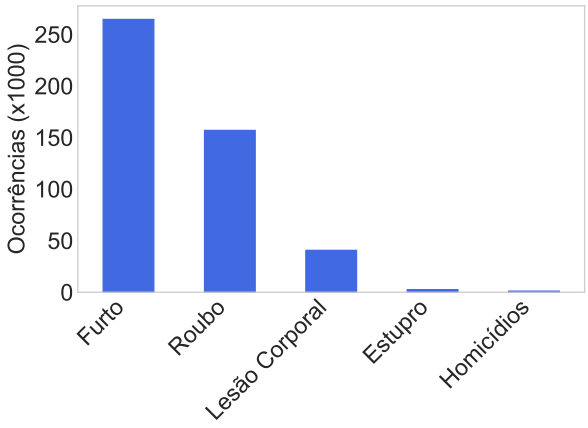
Nesta seção descrevemos as bases de dados e a metodologia de processamento desses dados para o uso em modelos de predição. Primeiramente, descrevemos os dados sobre índices de criminalidade que foram extraídos de fontes de dados oficiais. A seguir, descrevemos a metodologia para extração de características de usuários e POI.

4.1 Índices de Crimes Oficiais

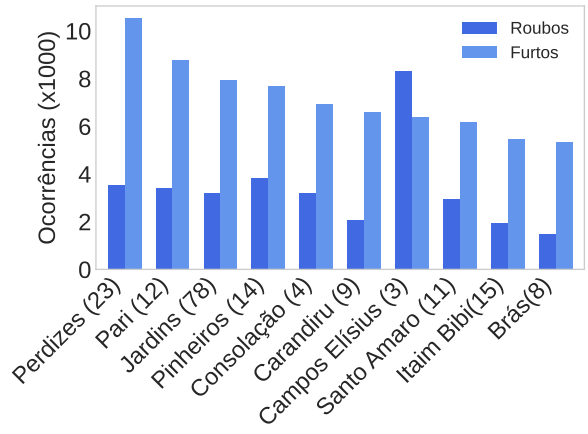
Os dados de índices criminais foram extraídos da secretaria de Segurança Pública do Estado de São Paulo (SSP-SP, 2021). Esses dados são divulgados mensalmente organizados em vinte e três categorias de crimes contendo a contagem de ocorrências registradas por regiões do estado desde 2001. A área geográfica de cada região consiste na delimitação dos *distritos policiais* definidos em lei pelo governo do estado (São Paulo, 2015), como mostra a Figura 4.1. Essas regiões são áreas delimitadas por ações estratégicas de segurança pública e não seguem propriamente as definições de bairros conhecidas dessas cidades.

O foco de nosso estudo foi nos distritos policiais da cidade de São Paulo, a capital do estado, por se tratar de regiões com os maiores índices criminais. São Paulo atualmente contem noventa e três distritos policiais e coletamos dados de oitenta e dois desses entre maio de 2022 a abril de 2023. Utilizamos esse período para compatibilizar com a coleta de conteúdos gerados por usuários na Web, descritos a seguir. Para facilitar a análise de crimes, reduzimos a complexidade do estudo ao agrupar as 21 categorias de crimes em 5 categorias mais representativas. Isso nos permitiu lidar de maneira mais eficiente com os dados coletados dos 82 distritos policiais de São Paulo entre maio de 2022 e abril de 2023, período alinhado com a coleta de conteúdo gerado por usuários na Web. A Figura 4-a mostra as categorias de crimes mais frequentes em nossa base de dados, considerando a soma de cada categoria no referido período:

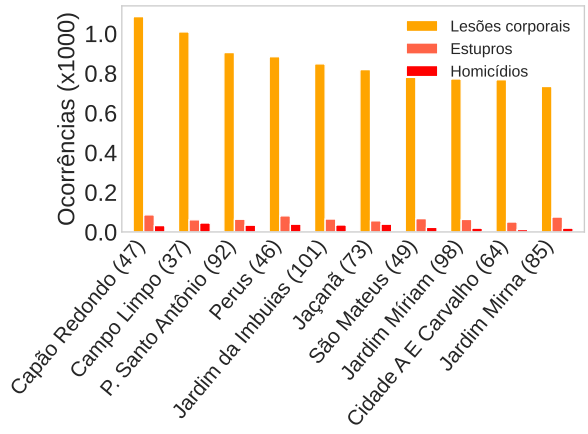
- Furto: total de furto e furto de veículos.
- Roubo: total de roubo e roubo de veículos.
- Homicídios: homicídio doloso, homicídio culposo por acidente de trânsito e homicídio culposo.
- Estupro: estupro e estupro de vulnerável.
- Lesão Corporal: lesão corporal seguida de morte, lesão corporal dolosa, lesão corporal culposa por acidente de trânsito e lesão corporal culposa.



(a)



(b)



(c)

Figura 4 – Categorias de crimes analisadas na cidade de São Paulo entre abril 2022 a maio 2023: (a) Ocorrências por categoria, (b-c) dez regiões (identificador) com ocorrências mais frequentes de (b) furto e roubo, (c) lesão corporal, estupro e homicídio.

A Figura 4-a mostra as ocorrências. Como pode-se observar furtos e roubos são os crimes mais frequentes nesse período com cerca de 264.503 e 156.895 ocorrências, respectivamente. Não obstante, homicídios, ocupa a quinta posição com um acumulado de 980

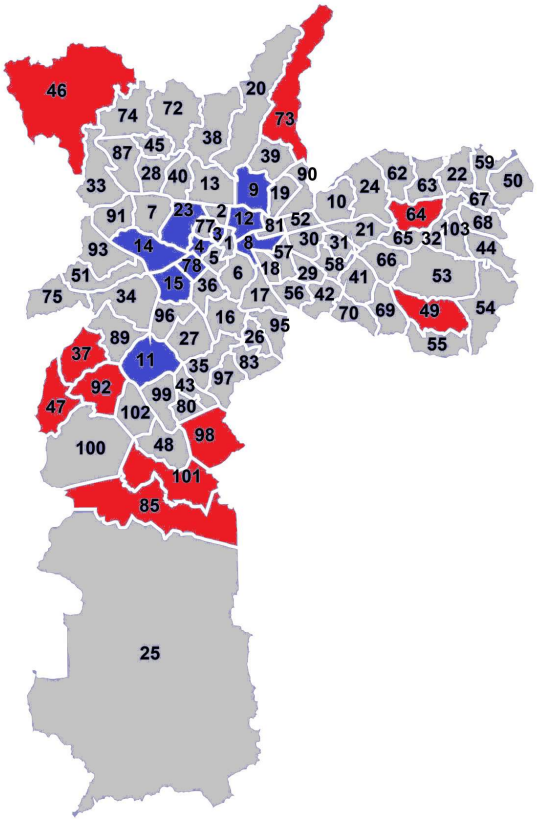


Figura 5 – Regiões (distritos policiais) em São Paulo (SSP-SP, 2021).

ocorrências nesse período em que abordamos.

A Figura 4 b mostra as regiões mais violentas para os dois crimes mais frequentes (furto e roubo). Pode-se observar que furto e roubo são mais (Figura 4-b) frequentes nas regiões centrais da cidade (ver Figura 4.1), também regiões de maior poder aquisitivo da população (e.g., Perdizes, Campos Elísios, Pari, Jardins e Consolação) com taxas entre 7 mil e 11 mil ocorrências para furtos, ao passo que roubos alcançam taxas entre 2 a 8 mil ocorrências. Contudo, à medida que crimes aumentam o grau de violência também cresce, como é o caso de lesão corporal, estupros e homicídios (Figura 4-c), eles ainda prevalecem mais altos nas regiões centrais em direção às regiões periféricas. Por exemplo, em regiões como Perus (46) e Jaçanã (73) a taxa de homícios no referido período foi igual a 39. Essas regiões têm histórico de altos índices de homicídios devido a contrastes sociais e em especial ao tráfego de drogas (ADORNO; NERY, 2019).

4.2 Pontos de Interesse (POI)

POI são tipos de dados mantidos por serviços Web de mapeamento urbano que provê informações sobre locais da cidade com precisão de coordenadas geográficas (YUAN; ZHENG; XIE, 2012). A informação principal de um ponto de interesse é a sua categoria,

que representa atividades econômicas ou culturais do local, inferida pelo serviço a partir de sensoriamento participativo e redes sociais baseadas em localização (SILVA et al., 2019). Nesse trabalho exploramos POI do serviço OSM obtidos da API Osmosis (RAMM; TOPF; CHILTON, 2011) que oferece ferramentas de programação para coletar POI por regiões delimitadas por coordenadas geográficas.

Utilizamos essa API para mapear todos os POI dos oitenta e dois distritos policiais da cidade de São Paulo. As descrições sobre delimitações de cada região foram obtidas no caderno do Estado de São Paulo (São Paulo, 2015). Contudo, tais descrições não contém as coordenadas geográficas necessárias para o mapeamento via API do OSM. Para obtê-las seguimos as descrições de delimitações, i.e. nome de ruas, pontos de referências, direções, capturando as coordenadas via Google Maps manualmente[1].

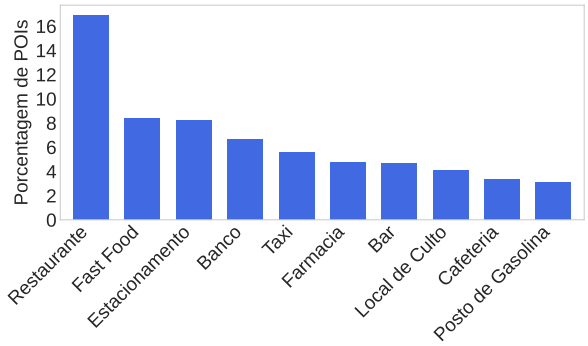
Dado esses procedimentos, coletamos 8.675 POI cadastrados no OSM em abril de 2023 na cidade de São Paulo, organizados em 108 categorias. A Figura 6-a mostra as dez categorias mais frequentes em nossa base de dados. É possível observar algumas tendências em relação aos pontos de interesse. Notavelmente, restaurante é a categoria mais frequente (16,88%), e isso ocorre provavelmente devido ao uso massivo do OSM por usuários com forte demanda por locais para alimentação. Em segundo lugar, temos *fast food* com uma porcentagem relevante (8,41%) indicando regiões onde as pessoas têm demanda por refeições rápidas. Há ainda outras categorias representativas não relacionadas diretamente a alimentação como estacionamentos (8,18%), bancos (6,63%) e taxis (5,57%). Avaliaremos o potencial para predição de crimes de todas as categorias de POI, pois esperamos que algumas dessas tenham informações preditivas para crimes, ainda que estejam em menor porcentagem.

A Figura 6-b mostra as dez regiões (distritos policiais) da cidade de São Paulo que acumulam maior volume de POI. Como esperado, a maioria dessas regiões fazem parte ou são vizinhas do centro de São Paulo onde há maior atividade econômica e por conseguinte movimentação diária de pessoas, observando que as regiões Pinheiros e Jardins tem as maiores quantidades de POI. A categoria restaurante é majoritária nessas dez regiões sendo um atributo importante para prever furtos. Desconsiderando essa categoria, para fins informativos, destacam-se Consolação e Pinheiros com predominância das categorias teatro (12,96%) e bar (23,22%) respectivamente, i.e., regiões com volumes representativos de atividades culturais. Por sua vez, a região Jardins tem a maior predominância de bancos e estacionamentos.

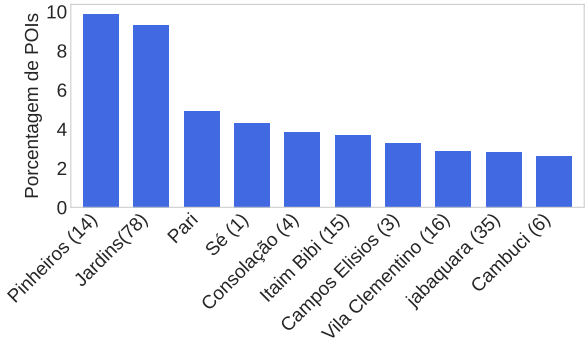
4.3 Tweets Georreferenciados

Caracterizamos usuários da rede social virtual Twitter que postaram mensagens de texto (tweets) georreferenciados nas delimitações geográficas das oitenta e duas áreas

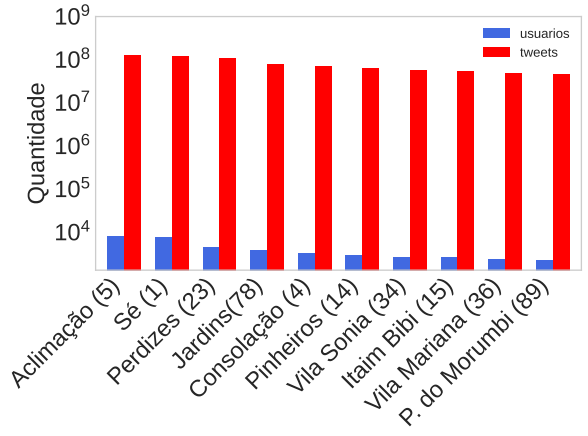
[1] https://www.google.com.br/maps



(a)



(b)



(c)

Figura 6 – Visão geral dos dados coletados: (a) Pontos de Interesse (POI) mais frequentes e (b) regiões que acumulam o maior volume de POI, (c) regiões que acumulam maior volume *tweets* georreferenciados e seus usuários.

(distritos policiais) de São Paulo. É importante mencionar que optamos pelo Twitter devido aos seus recursos para pesquisas com esse tipo de postagem. A caracterização foi realizada em duas etapas. Primeiramente, monitoramos as referidas áreas para a coleta de *tweets*. Em seguida, processamos informações dos *tweets* para extrair características dos usuários que realizaram as postagens.

Na primeira etapa, monitoramos cada área utilizando a API do Twitter versão 2 com

a biblioteca *Tweepy* na linguagem Python[2]. Essa biblioteca funciona como um facilitador para acessar a API, tornando possível buscar *tweets* por uma área retangular delimitada por quatro coordenadas geográficas. A figura 7 mostra um exemplo de busca de *tweets* em uma determinada área na região de São Paulo.

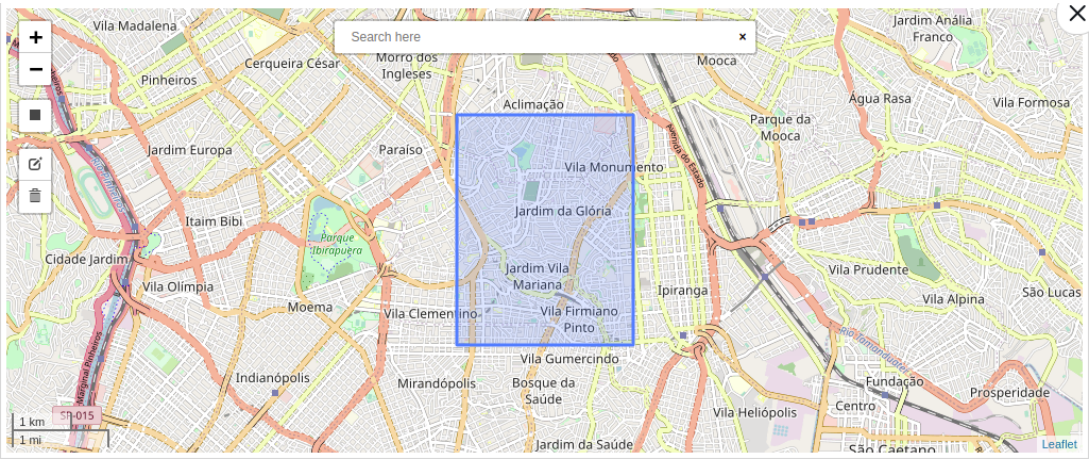


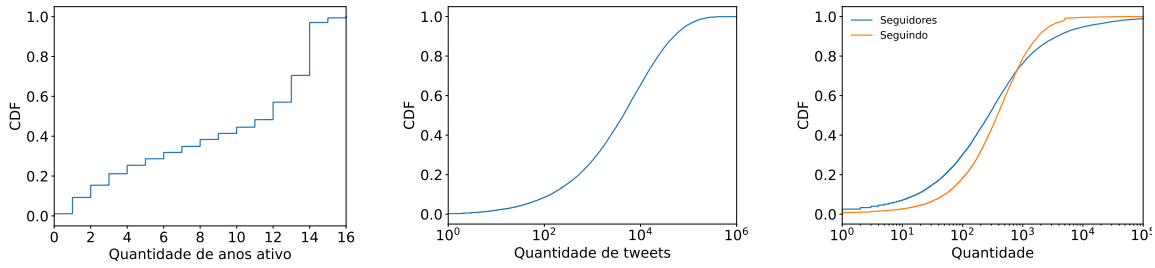
Figura 7 – Área de localização dos tweets

Dessa forma, mapeamos cada região de modo que ela esteja inteiramente contida numa área retangular, mas buscando reduzir ao máximo sobreposições entre essas áreas. Coletamos *tweets* postados por região diariamente de maio de 2022 a abril de 2023, somando um total de oito meses completos de coleta[3]. Importante mencionar que a API do Twitter retorna *tweets* georreferenciados de até sete dias anteriores. Logo, o monitoramento diário (uma requisição por dia por região) foi o melhor compromisso encontrado para aumentar as chances de coletar *tweets georreferenciados* distintos por região e não ultrapassar o limite de dados requisitados da API. Ao fim dessa etapa, obtivemos um total de 278.058 *tweets* e 80.453 usuários distintos. A Figura 6-c mostra as dez áreas que acumulam maior volume de *tweets* georreferenciados e usuários distintos que os postaram. A maioria dessas regiões fazem parte ou são vizinhas do centro de São Paulo.

A próxima etapa consiste em extrair características dos usuários que postaram os *tweets* coletados por região. A caracterização focou em padrões de comportamento dos usuários ao invés de conteúdo dos tweets. Isso porque não foi encontrado variabilidade dos conteúdos por regiões em termos de sentimento (polaridade e intensidade) e tópicos, mesmo utilizando ferramentas estado da arte como BERT (DEVLIN et al., 2019), e exploraremos essa questão em trabalhos futuros. Dessa forma, calculamos atributos quantitativos a partir de dados públicos nos perfis dos usuários: tweets postados, seguidores, pessoas seguidas, e anos ativo desde criação da conta. Adicionalmente, coletamos dois atributos binários, i.e., sim ou não, sobre o perfil dos usuários: perfil privado e conta

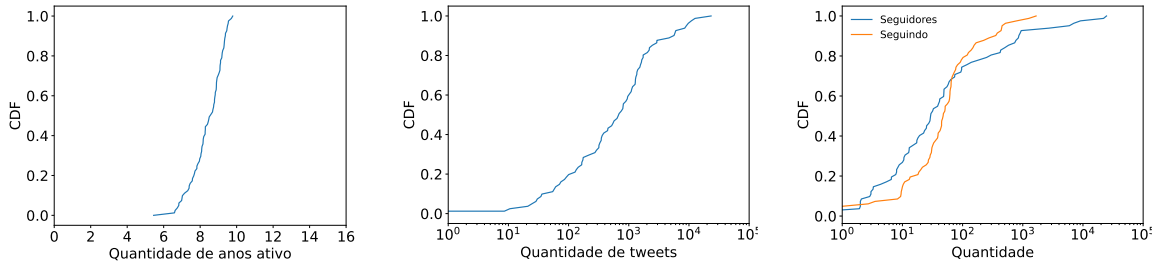
[2] https://docs.tweepy.org/en/stable/

[3] Não foi possível coletar todos os meses consecutivamente devido limites de requisições de dados da licença de pesquisa com a API com Twitter.



(a) (b) (c)

Figura 8 – Características de usuários independente da região onde postaram *tweets*: (a) anos de atividade no Twitter, (b) número de tweets postados por ano, e (c) número de pessoas seguindo e seguidores.



(a) (b) (c)

Figura 9 – Média aritmética de atributos de usuários nas regiões monitoradas: (a) anos ativos no Twitter, (b) número de tweets postados por ano, e (c) número de pessoas seguindo e seguidores.

verificada pelo usuário. Extraímos, assim, um total de seis atributos de tweets georreferenciados referentes a padrões de comportamento de usuários.

A Figura 8 mostra funções de distribuições acumuladas (CDF) para atributos de comportamento dos usuários independente da área onde eles postam *tweets*. Os anos ativos dos usuários no Twitter são bem distribuídos entre menos de um ano até dezessete anos com uma leve concentração 42.2% dos usuários entre 13 e 15 anos de atividade (Fig. 8-a). Contudo, a quantidade de tweets postado por usuários é bem desbalanceada com um grupo pequeno (5% dos usuários) que já postaram mais de cem mil tweets (10^5) durante sua atividade na rede. Em termos de taxa tweet/ano por usuário (Fig. 8-b) observa-se ainda esse desbalanceamento de mais de 10 mil tweets/ano para 5% dos usuários, ao passo que outros 60% postam menos de 1000 tweets/ano (Fig. 8-b). A quantidade de pessoas seguindo e seguidoras por usuário também é muito desbalanceada, a vasta maioria (80% dos usuários) tem bem menos de 1000 pessoas seguindo ou seguidoras e novamente observa-se um grupo pequeno (5% dos usuários) onde esse número ultrapassa 10000 pessoas (Fig. 8-c). Quanto aos atributos binários temos 4.1% dos usuários com perfil protegido e 1.5% de usuários com perfil verificado.

Agora observamos o comportamento dos usuários agregado por área, afim de utilizá-los como atributos para predição de crimes. Nesse sentido, aplicamos média aritmética dos atributos de usuários por região, considerando a associação de usuários às regiões

onde ele postou ao menos um *tweet*. Assumimos, desse modo, que usuários do Twitter estão associados a uma área ou mais por terem realizado alguma atividade antes, durante ou após a postagem no local, seja como um residente ou transeunte (e.g., turistas, passageiros de transporte público ou privado). Em concordância com pesquisas recentes nessa questão (TUCKER et al., 2021; IRANMANESH; ATUN, 2020), entendemos cada *tweet* georreferenciado como uma interação relevante do usuário com o espaço urbano, pois é necessário habilitar o georreferenciamento manualmente por esse não ser padrão no aplicativo do Twitter para dispositivo móvel.

A Figura 9 mostra CDF das médias de atributos de usuários nas regiões monitoradas. Observa-se que essas distribuições seguem ligeiramente as tendências mostradas na Figura 8, com exceção para os anos de atividade cujas médias por região se concentram em valores entre 5 e 10 anos. Por sua vez, nos atributos binários temos 9% das regiões com nenhum perfil verificado, alcançando quase 3% na região com a maior marca de verificações, ao passo que a proteção de perfis mostra ter maior adesão dos usuários e a vasta maioria das regiões (90%) teve entre 3-6% de seus usuários com perfis protegidos.

Nesta seção apresentamos nossos resultados sobre as bases de dados de crimes, características de usuários do *Twitter* e POI apresentadas na seção anterior. Primeiramente, descrevemos as configurações utilizadas nos experimentos, considerando métodos de regressão e métricas de desempenho para predição. A seguir, discutimos os resultados alcançados em termos de desempenho de diferentes métodos e importância dos atributos utilizados na predição.

4.4 Configurações

O nosso objetivo nesses experimentos é investigar o potencial de POI e as características de usuários do Twitter para explicar taxas de ocorrências de crimes por categoria e por região da cidade. Para isso, propomos modelos de regressão em que o total de ocorrências para uma determinada categoria de crime por região da cidade seja a variável a ser predita (y). Por sua vez, POI e características de usuários serão utilizados como atributos para essa predição. Portanto, definimos uma matriz X das cento e sete categorias de POI e seis características de usuários (total de 113 colunas) para as regiões da cidade (82 linhas).

Contudo, precisamos construir um modelo para predizer a categoria de crime em uma região alvo a, desconsiderando essa região nos valores a serem preditos y e na matriz de atributos X para fins de avaliação do modelo. Nesse sentido, adotamos a metodologia *leave out one* que consiste em predizer o total de crimes para uma região utilizando dados das outras regiões. Mais formalmente construímos modelos com o formato:

$$\hat{y}_a = M(y, X)_{\setminus a}, \tag{4.1}$$

onde $\hat{y}_a$ é a taxa de uma categoria de crime para uma região alvo a ser predita, a função

M representa diferentes métodos de regressão a serem utilizados para o treinamento do modelo. Por sua vez, a região alvo a será excluída das taxas de crimes y e atributos X a serem utilizados no treinamento do modelo. Em outras palavras, retiramos a região alvo dos dados para realizar a sua predição, ao passo que as outras regiões foram utilizadas para treinar o modelo.

Todas as predições para cada categoria crime foram utilizados para avaliar o desempenho do modelo. O desempenho dos modelos de predição foi avaliado através das métricas média do erro absoluto (MAE), média do erro relativo (MRE) e o índice R2. Especificamente essas métricas foram calculadas da seguinte forma:

$$MAE = \frac{\sum_i |y_i - \hat{y}_i|}{n}; MRE = \frac{\sum_i |y_i - \hat{y}_i|}{\sum_i y_i}; R2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{4.2}$$

onde y_i e $\hat{y}_i$ representam os valores real e predito para a taxa de crime na i-ésima região, ao passo que $\bar{y}$ representa a média da taxa de crime considerando todas as n regiões para a construção do modelo, i.e., oitenta e dois distritos policiais da cidade de São Paulo.

Nossos experimentos foram realizados com as implementações de métodos de regressão da biblioteca *scikit-learn* da linguagem *python* (PEDREGOSA et al., 2011b). Utilizamos regressão linear, aprendizagem de máquina e redes neurais artificiais para construção do modelo M mostrado na Equação 4.1. Reportamos os métodos de aprendizagem de máquina que obtiveram os resultados mais satisfatórios para os dados coletados. Redes neurais, em especial, requer um maior volume de dados para treinamento e avançaremos com esse tipo de modelo à medida que coletamos mais dados em trabalhos futuros. Os métodos de aprendizagem de máquina utilizados são floresta aleatória (FA) e *support vector regression* (SVR), *Gradient Boosting Regressor* (GBR) e *K-Nearest Neighbors* (KNN) (DRUCKER et al., 1997; BREIMAN, 2001). FA combina um conjunto de preditores de árvores de modo que cada árvore depende dos valores de um vetor aleatório da amostragem com a mesma distribuição. SVR, busca predizer um valor real após traçar duas retas paralelas, chamadas de limites. O modelo ainda traça uma reta linear entre as duas outras retas afim de ajustar seus valores. Já o GBR é baseado na técnica *boosting* que combina várias árvores de decisão fracas para construir um modelo preditivo sequencialmente. Por fim, o KNN que é um algoritmo que busca prever valores de saída com base nos k vizinhos mais próximos no espaço de atributos.

5 Resultados

Nesta seção, será mostrado o desempenho dos diferentes métodos utizados para a predição e os atributos mais relevantes para os métodos de regressão.

5.1 Análise de Desempenho

A Tabela 2 mostra o desempenho dos métodos de regressão para as cinco categorias de crimes. Analisamos primeiramente o desempenho dos quatro métodos de regressão utilizados. Floresta Aleatória obteve o melhor desempenho, apresentando os maiores índices R2 e menores erros absolutos e relativos para quatro das cinco categorias de crimes analisadas. Por sua vez, SVR foi o melhor método para predizer roubos com ligeira superioridade sobre FA. Desconsiderando essa categoria de crime, GBR foi o segundo método com melhor desempenho, enquanto KNN apresentou os piores resultados. Observa-se, portanto, que, de forma geral, à medida que o R2 aumenta, os erros diminuem, sendo que esses erros são menores para os modelos com o método Floresta Aleatória.

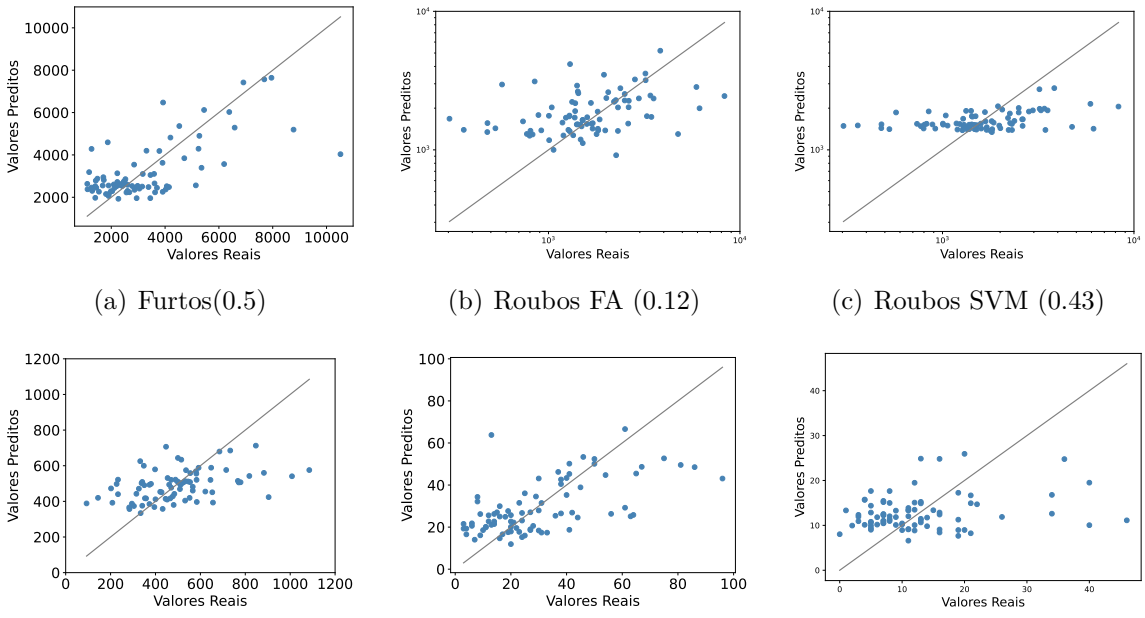
Agora focamos nos resultados da regressão com floresta aleatória, que obteve melhor desempenho, para analisar as categorias de crimes. Observa-se que as categorias de crimes com melhor desempenho do modelo são Furtos, Lesão corporal e Estuproscom índices de determinação R2 de 0,48, 0,39, 0,14 respectivamente e média de erros relativos (MRE) inferiores 39% da taxa oficial. Entendemos que esses são crimes que ocorrem com mais frequência em regiões no centro geográfico da cidade, em especial Roubos, onde há maior circulação de pessoas e portanto maior disponibilidade de UGC na web. Nesse caso, os modelos de aprendizagem de máquina conseguem extrair mais informações dos atributos para a predição.

As categorias de crimes Lesão corporal e Estupros, além de serem frequentes no centro também ocorrem em regiões como Itaquera, Parrelheiros e São Mateus, que são periféricas geograficamente e com menos UGC coletado. Embora essas regiões estejam dentre as dez com maiores taxas de Lesão corporal e Estupros, não observa-se aumentos drásticos dos erros para esses crimes, possivelmente, porque o modelo captura padrões preditivos dos atributos nessas regiões semelhantes aos das regiões centrais. Por sua vez, os crimes de Roubos e Vítimas de homicídios não se concentram em regiões do centro há maior disponibilidade de UGC. Conjecturamos que isso impacta na qualidade dos atributos para inferências do modelo nessas categorias (melhores R2 alcançados entre 0,05 e 0,09 com métodos RF e SVM). Contudo, elas ainda podem ser preditas via atributos de UGC com erros médios de até 52.7% em relação à taxa oficial.

Os gráficos de pontos da Figura 10 fornecem uma visão mais detalhada sobre os erros absolutos médios mostrados na Tabela 2 nos crimes Furtos, Lesão corporal e Estupros. É notável a tendência de correlação linear entre valores reais e preditos alcançando índices de correlação de Pearson superiores a 0,4 ligeiramente proporcionais aos índices R2 do melhor modelo (RF). Nota-se ainda melhores desempenhos (maior proximidade dos pontos à linha diagonal de referência) para os valores baixos, médios e altos de taxas de crimes nas categorias Estupros, Lesão corporal e Furtos respectivamente. Os crimes de Roubos e Homicídios, alcançam índice de correlação de Pearson de 0,12 e 0,24 respectivamente. Observamos que as predições nessas duas categorias tendem a se concentrar em valores médios ao passo os valores reais variam entre baixo e alto em algumas regiões. Ainda assim, é possível explorar UGC para predizer os crimes na maioria de regiões. Por exemplo, SVR alcançou índice de correlação de 0,43, capturando padrões entre as taxas de Roubos e os atributos de UGC nas regiões onde categoria de crime tem valores medianos.

Tabela 2 – Desempenho dos métodos de regressão para predição da taxa de crimes entre maio 2022 a abril 2023: Floresta Aleatória (FA), *Support Vector Regression* (SVR),*Gradient Boosting Regressor* (GBR) e *K-Nearest Neighbors* (KNN).

Crimes	RAE				MAE				R2			
	FA	SVR	GBR	KNN	FA	SVR	GBR	KNN	FA	SVR	GBR	KNN
Furtos	**0,29**	0,34	0,30	0,38	**936,06**	1120,36	975,53	1239,31	**0,48**	0,24	0,45	0,02
Roubos	0,41	**0,39**	0,49	0,45	794,89	**747,09**	943,57	871,28	0,05	**0,09**	-0,18	-0,12
Lesão corporal	**0,25**	0,28	0,27	0,28	**126,99**	140,92	135,51	143,10	**0,14**	0,04	0,11	-0,01
Estupros	**0,39**	0,66	0,41	0,47	**11,79**	19,99	12,37	14,07	**0,39**	-0,43	0,37	0,20
Homicídios	0,52	0,75	0,59	0,55	6,43	9,06	7,15	6,60	0,06	-0,81	-0,18	-0,11



(a) Furtos(0.5) (b) Roubos FA (0.12) (c) Roubos SVM (0.43)

(d) Lesão corporal(0.41) (e) Estupros(0.61) (f) Homicídios(0.24)

Figura 10 – Correlação entre valores preditos e reais considerando o melhor modelo para cada categoria de crime e coeficientes de correlações de Pearson entre parênteses.

5.2 Atributos Relevantes

Finalmente, analisamos a importância dos atributos UGC para predição de crimes por região, baseados no melhor método de regressão obtido, i.e., floresta aleatória (RF). Uma vantagem relevante de RF é estimar essa importância dentro de seu próprio algoritmo via médias e o desvios padrões da redução de impurezas acumuladas nas árvores aleatórias geradas (BREIMAN, 2001). A Tabela 3 mostra cada uma das categorias de crimes seguido dos quatro atributos mais importantes para predição com suas respectivas porcentagens de relevância indicadas pelo método FA entre parênteses.

Tabela 3 – Relação de quatro atributos mais importantes para predição da taxa anual de crimes.

Crime	Ordem de importância dos atributos com sua respectiva porcentagem (%) estimada pelo método FA			
	1a.	2a.	3a.	4a.
Estupros	Twitter Anos Ativos (51,3%)	Clínica (4,9%)	Local de Culto (4,1%)	Twitter Verificado (3,5%)
Homicídios	Twitter Anos Ativos (16,4%)	Twitter Seguindo (9,3%)	Local de Culto (9,3%)	Tribunal. (6,0%)
Lesão corporal	Twitter Anos Ativos (21,0%)	Local de Culto (13,3%)	Escolas (5,3%)	Twitter Seguindo (4,9%)
Roubos	Telefone (17,0%)	Twitter Anos Ativos (9,7%)	Cafeteria (8,6%)	Pontos de Táxi (6,1%)
Furtos	Pontos de Táxi (28,9%)	Estacionamentos (15,4%)	Restaurantes (9,4%)	Fast Food (6,9%)

Do total de cento e treze atributos UGC utilizados, treze aparecem entre as quatro mais importantes para predição dos crimes apresentados, onde três características de comportamento de usuários e os demais são POI. Quantidade de anos ativos no *Twitter* é a categoria predominante aparecendo entre a primeira e a segunda ordens de importâncias para quatro categorias de crime. Por outro lado, atributos UGC do tipo POI relacionados a locais de estacionamento e alimentação são os quatro mais importantes para predizer crimes de *Furtos*. Isso indica que a quantidade desses POI por região podem capturar o seu nível e o tipo de atividade econômica que, por conseguinte, estão associadas às ocorrências de furtos oficialmente registradas.

É importante destacar que na ausência de dados oficiais é possível predizer razoavelmente algumas categorias de crimes com dados extraídos unicamente da Web, no entanto, que explicações sobre a importância de alguns atributos POI para determinados crimes não são triviais e requer a análise de especialistas experientes em criminologia e urbanismo. Por exemplo, Clínica e Local de Culto são categorias de POI importantes para predizer crimes de Lesão corporal e Estupros, mas explicações intuitivas para esse relacionamento podem ser tornar complexas. Possivelmente, há outras características espaciais associadas à frequência desses POI em algumas regiões que levam a relações indiretas com tais crimes. De modo semelhante, a presença de pessoas que têm algum nível de atividade na web (e.g., Twitter) em regiões da cidade, registrado via tweets georreferenciados, podem indicar taxas de crimes contra pessoas (especificamente Lesão corporal, Estupros, Vítimas de homicídios) nessas regiões. Novamente, essas são evidências empírica que podem ser tratados por sociólogos e demais especialistas em comportamento humano. Essas questões sobre a causalidade dos atributos mencionados serão investigadas em futuras extensões desse trabalho.

6 Conclusão

Nesse trabalho investigamos o potencial de atributos extraídos de UGC na web, especificamente POI e *tweets* georreferenciados para predizer taxas anuais de ocorrências de crimes por categoria, tomando como um importante caso de estudo regiões da cidade de São Paulo. Os esforços de pesquisa em computação sobre nesse tema tipicamente utilizam UGC como um incremento à dados oficiais sobre demografia e censo urbano para predizer crimes. O nosso desafio nesse trabalho é predizer crimes unicamente baseado em atributos extraídos de UGC na web como um recurso adicional na ausência ou atraso de dados oficiais. Nesse sentido, construímos um conjunto de dados com atributos de POI e comportamento de usuários do Twitter relacionados a taxas de crimes oficiais por regiões da cidade. Analisamos essa relação com modelos de regressão baseados em aprendizagem de máquina e seus respectivos desempenhos via média de erros absolutos, média de erros relativos e o índice R2.

Nossos resultados evidenciam que o uso de atributos extraídos de UGC na web unicamente podem predizer razoavelmente algumas categorias de crimes. O melhor modelo de regressão avaliado obteve um erro médio de 29% relativo à taxa anual oficial e R2 de 0,48, utilizando UGC do tipo POI, em especial pontos de taxi e estacionamentos, como atributos mais relevantes para predição de furtos. Outras quatro categorias de crimes (estupro, lesão corporal, roubos e homicídios) foram preditos com média de erros relativos inferiores 39% da taxa oficial e índices R2 superiores a 10%, utilizando atributos de comportamento de usuários no Twitter como os mais relevantes para predição.

Esse artigo abre oportunidades trabalhos futuros. Dentre essas consideramos as mais imediatas, e mencionadas nas seções anteriores, a continuidade do monitoramento e coleta de dados para explorar modelos baseados em rede neurais para regressão. Adicionalmente, há oportunidades para desenvolvimento de modelos de processamento de linguagem natural para extração de atributos de conteúdos de *tweets*, e análises de causalidade de atributos.

7 Publicações

Abaixo são listadas as publicações científicas do autor deste trabalho:

- Feitosa, M. F., Rocha, S., Gonçalves, G. D., Ferreira, C. H., & Almeida, J. M. (2022, November). Sentiment Analysis on Twitter Repercussion of Police Operations. In Proceedings of the Brazilian Symposium on Multimedia and the Web (pp. 84-88).
- Rocha, S.; Goncalves, G. D. . Avaliação de Ferramentas Computacionais para Análise de Sentimento de Pessoas sobre Violência Urbana. 2022. (Apresentação de Trabalho/Seminário).

Referências

ADORNO, S.; NERY, M. B. Crime e violências em são paulo: retrospectiva teórico-metodológica, avanços, limites e perspectivas futuras. *Cadernos Metrópole*, SciELO Brasil, v. 21, n. 44, p. 169–194, 2019. Citado 3 vezes nas páginas 19, 20 e 23.

AHMED, A.; HONG, L.; SMOLA, A. J. Hierarchical geographical modeling of user locations from social media posts. In: *Proceedings of the 22nd international conference on World Wide Web*. [S.l.: s.n.], 2013. p. 25–36. Citado na página 10.

BECKER, K. L.; KASSOUF, A. L. Uma análise do efeito dos gastos públicos em educação sobre a criminalidade no brasil. *Economia e Sociedade*, SciELO Brasil, v. 26, n. 1, p. 215–242, 2017. Citado 2 vezes nas páginas 19 e 20.

BELESIOTIS, A.; PAPADAKIS, G.; SKOUTAS, D. Analyzing and predicting spatial crime distribution using crowdsourced and open data. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, ACM New York, NY, USA, v. 3, n. 4, p. 1–31, 2018. Citado 3 vezes nas páginas 11, 19 e 20.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado 2 vezes nas páginas 29 e 32.

BZDOK, D. *Classical statistics and statistical learning in imaging neuroscience. Front. Neurosci. 11, 543.* 2017. Citado na página 15.

CASTRO, U. R.; RODRIGUES, M. W.; BRANDAO, W. C. Predicting crime by exploiting supervised learning on heterogeneous data. In: *ICEIS (1)*. [S.l.: s.n.], 2020. p. 524–531. Citado 2 vezes nas páginas 11 e 19.

CHENG, Z. et al. Exploring millions of footprints in location sharing services. In: *Proceedings of the International AAAI Conference on Web and Social Media*. [S.l.: s.n.], 2011. v. 5, n. 1, p. 81–88. Citado na página 10.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, v. 1, 2019. Disponível em: <https://doi.org/10.48550/arXiv.1810.04805>. Citado na página 26.

DRUCKER, H. et al. Support vector regression machines. *Advances in neural information processing systems*, Morgan Kaufmann Publishers, v. 9, p. 155–161, 1997. Citado na página 29.

HUANG, C. et al. Deepcrime: Attentive hierarchical recurrent networks for crime prediction. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. [S.l.: s.n.], 2018. p. 1423–1432. Citado 3 vezes nas páginas 11, 19 e 20.

IRANMANESH, A.; ATUN, R. A. Reading the urban socio-spatial network through space syntax and geo-tagged twitter data. *Journal of Urban Design*, Taylor & Francis, v. 25, n. 6, p. 738–757, 2020. Citado 3 vezes nas páginas 10, 14 e 28.

KOTZIAS, D.; LAPPAS, T.; GUNOPULOS, D. Home is where your friends are: Utilizing the social graph to locate twitter users in a city. *Information Systems*, Elsevier, v. 57, p. 77–87, 2016. Citado 2 vezes nas páginas 10 e 14.

LEE, R.; WAKAMIYA, S.; SUMIYA, K. Urban area characterization based on crowd behavioral lifelogs over twitter. *Personal and ubiquitous computing*, Springer, v. 17, p. 605–620, 2013. Citado na página 13.

MASI, C. M. et al. Neighborhood economic disadvantage, violent crime, group density, and pregnancy outcomes in a diverse, urban population. *Social science & medicine*, Elsevier, v. 65, n. 12, p. 2440–2457, 2007. Citado na página 19.

MITCHELL, T. *Machine Learning*. McGraw-Hill, 1997. (McGraw-Hill International Editions). ISBN 9780071154673. Disponível em: <https://books.google.com.br-/books?id=EoYBngEACAAJ>. Citado na página 15.

MUELLER, W. et al. Gender matters! analyzing global cultural gender preferences for venues using social sensing. *EPJ Data Science*, Springer, v. 6, n. 1, p. 5, 2017. Citado na página 14.

NERY, M. B.; SOUZA, A. A. L. d.; ADORNO, S. Os padrões urbano-demográficos da capital paulista. *Estudos Avançados*, SciELO Brasil, v. 33, n. 97, p. 5–36, 2019. Citado na página 19.

NEV-USP. *Monitor da violência*. 2022. Disponível em: https://nev.prp.usp.br/projetos/projetos-especiais/monitor-da-violencia/. Acesso em 07 de jun. 2023. Citado na página 10.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 16.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, JMLR. org, v. 12, p. 2825–2830, 2011. Citado na página 29.

RAMM, F.; TOPF, J.; CHILTON, S. *OpenStreetMap: using and enhancing the free map of the world*. [S.l.]: UIT Cambridge Cambridge, 2011. Citado na página 24.

SILVA, T. H. et al. Uma fotografia do instagram: Caracterização e aplicação. *Revista Brasileira de Redes de Computadores e Sistemas Distribuídos*, 2017. Citado na página 14.

SILVA, T. H. et al. Urban computing leveraging location-based social network data: a survey. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 52, n. 1, p. 1–39, 2019. Citado 4 vezes nas páginas 10, 13, 14 e 24.

SOUSA, D. d. S.; FEITOSA, M. P. F.; GONÇALVES, G. D. Relações entre crimes e o espaço urbano: Um estudo de caso baseado em pontos de interesses extraídos da web. In: SBC. *Anais do V Workshop de Computação Urbana*. [S.l.], 2021. p. 196–208. Citado 2 vezes nas páginas 19 e 20.

SOUSA, D. da S.; FEITOSA, M. P. F.; GONÇALVES, G. D. Relações entre crimes e o espaço urbano: Um estudo de caso baseado em pontos de interesses extraídos da web. In: SBC. *Anais do V Workshop de Computação Urbana*. [S.l.], 2021. p. 196–208. Citado na página 20.

SSP-SP. *Dados Estatísticos do Estado de São Paulo*. 2021. Disponível em: http://www.ssp.sp.gov.br/estatistica/pesquisa.aspx. Acesso em 10 de mai. 2021. Citado 4 vezes nas páginas 6, 11, 21 e 23.

São Paulo. *Diário Oficial Do Estado De São Paulo*. 2015. Disponível em: https://www.imprensaoficial.com.br. Acesso em 07 de jul. 2021. Citado 2 vezes nas páginas 21 e 24.

TUCKER, R. et al. Who 'tweets' where and when, and how does it help understand crime rates at places? measuring the presence of tourists and commuters in ambient populations. *Journal of Quantitative Criminology*, Springer, v. 37, n. 2, p. 333–359, 2021. Citado 4 vezes nas páginas 10, 19, 20 e 28.

VOMFELL, L.; HÄRDLE, W. K.; LESSMANN, S. Improving crime count forecasts using twitter and taxi data. *Decision Support Systems*, Elsevier, v. 113, p. 73–85, 2018. Citado na página 20.

WANG, H. et al. Learning task-specific city region partition. In: *The World Wide Web Conference*. [S.l.: s.n.], 2019. p. 3300–3306. Citado na página 11.

WANG, H. et al. Non-stationary model for crime rate inference using modern urban data. *IEEE transactions on big data*, IEEE, v. 5, n. 2, p. 180–194, 2017. Citado 2 vezes nas páginas 19 e 20.

WANG, J. et al. Crime risk analysis through big data algorithm with urban metrics. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 545, p. 123627, 2020. Citado 2 vezes nas páginas 19 e 20.

WANG, Z. et al. Identification and analysis of urban functional area in hangzhou based on osm and poi data. *Plos one*, Public Library of Science San Francisco, CA USA, v. 16, n. 5, p. e0251988, 2021. Citado 2 vezes nas páginas 10 e 14.

WEISBURD, D.; GROFF, E. R.; YANG, S.-M. *The criminology of place: Street segments and our understanding of the crime problem*. [S.l.]: Oxford University Press, 2012. Citado 3 vezes nas páginas 10, 19 e 20.

WILLIAMS, M. L.; BURNAP, P.; SLOAN, L. Crime sensing with big data: The affordances and limitations of using open-source communications to estimate crime patterns. *The British Journal of Criminology*, Oxford University Press, v. 57, n. 2, p. 320–340, 2017. Citado na página 20.

YAN, X.; SU, X. *Linear Regression Analysis: Theory And Computing*. World Scientific Publishing Company, 2009. ISBN 9789814470087. Disponível em: <https://books-.google.com.br/books?id=afzFCgAAQBAJ>. Citado na página 15.

YUAN, J.; ZHENG, Y.; XIE, X. Discovering regions of different functions in a city using human mobility and pois. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2012. p. 186–194. Citado 3 vezes nas páginas 10, 14 e 23.

ZHENG, Y. et al. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, ACM New York, NY, USA, v. 5, n. 3, p. 1–55, 2014. Citado na página 13.