PinRisk- Estudo Aplicado na Utilização de Machine Learning no Alerta da Possibilidade de Roubo e Furto de Smartphone.

Guilherme Angelus Felix Araujo; Pedro Henrique Rezende Oliveira; Thiago Alves do Nascimento
Orientador: Wenderson Alexandre de Sousa Silva

Resumo: O presente artigo acadêmico tem como objetivo apresentar um modelo de previsão de roubos de *smartphones* baseado em *machine learning* em áreas urbanas utilizando como exemplo a cidade de São Paulo. O modelo é alimentado com dados dos boletins de ocorrência disponíveis no portal da transparência da Secretaria de Segurança Pública do Estado de São Paulo para identificar padrões que indicam roubo. O modelo realiza a regressão através da técnica de *Random Forest* (Floresta Aleatória) prevendo a probabilidade de que em uma determinada área ocorra o roubo de um *smartphone*. Os resultados mostram que o modelo é capaz de prever o roubo podendo alcançar uma compreensão dos dados de até 95%.

Palavras-chave: Smartphone, Roubo, Modelo, Áreas Urbanas.

PinRisk - Applied Study on the Use of Machine Learning in Alerting the Possibility of Theft and Robbery of Smartphones.

Abstract: The present academic article aims to present a smartphone theft prediction model based on machine learning in urban areas, using the city of São Paulo as an example. The model is fed with data from occurrence bulletins available on the transparency portal of the Public Security Secretariat of the State of São Paulo to identify patterns indicating theft. The model performs regression using the Random Forest technique, predicting the probability of a smartphone theft occurring in a specific area. The results show that the model is capable of predicting theft with a data comprehension accuracy of up to 95%.

Keywords: Smartphone, Theft, Model, Urban Areas.

1. INTRODUÇÃO

O roubo e furto de aparelhos celulares é um problema crescente atualmente em grandes cidades, principalmente em São Paulo, a maior cidade da América Latina conforme o portal (capital.sp.gov.br). Tratando-se de celular, quando o usuário tem medo de o atender na rua, em transporte público ou até mesmo em lugares fechados, o celular deixa de ser móvel e passa a ser portátil, utilizando somente em lugares em que se sente seguro como por exemplo dentro de estabelecimentos comerciais, sua residência e seu trabalho. Portanto, é notável que a falta de segurança pública afeta diretamente o propósito central do serviço de telefonia móvel e toda a sua infraestrutura, além de diminuir a qualidade de vida.

Para auxílio da população, deve-se comunicar a sua prestadora de telefonia referente ao roubo, furto, extravios de dispositivo, transferência titular do dispositivo, informações cadastrais, conforme portal (ANATEL) Lei nº 9.472, de 16 de julho de 1997 Art. 4º, VII, “a” da Resolução nº 632/2014. Acreditamos que só essas leis não têm surtido o efeito necessário para atender o cidadão paulistano e inibir os criminosos nessa grande metrópole cheio de diversidade restaurantes, bares, teatros, cinemas, etc.

Os alvos preferidos são as pessoas que andam distraídas pela rua, falando ao celular, tirando selfies ou com o aparelho no bolso de trás (JN, 2016). Os aparelhos roubados vão parar no mercado clandestino (SP1, 2017).

As vítimas são orientadas a informar a Identidade Internacional de Equipamentos Móveis (IMEI) do aparelho no boletim de ocorrência, para que as operadoras de telefonia celular bloqueiem os aparelhos. Porém, investigações acerca dos destinos dos aparelhos celulares realizado pelo Departamento de Operações Policiais Estratégicas (DOPE) apuraram que grande parte dos aparelhos seriam levados para Dacar, em Senegal, na África, onde não têm o controle do IMEI (BAND, 2020b). Portanto, muitos aparelhos não retornam às vítimas.

Além de os aparelhos não retornarem às vítimas, há outros danos causados que vão além da perda do aparelho celular, eles podem incluir a utilização dos dados das vítimas que vão desde aplicar golpes em familiares e amigos, até realizar transferências de dinheiro (SP2, 2020). Agora com tecnologias que facilitam a transações, como o PIX, também viraram alvo dos criminosos que estão desbloqueando os aplicativos bancários “fazendo o pix” que levam a prejuízos imensuráveis ao cidadão metropolitano (UOL, 2022).

Os dados analisados neste artigo são as ocorrências de roubo e furto de aparelhos celulares na cidade de São Paulo disponibilizadas pela Secretaria de Segurança Pública do Governo do Estado de São Paulo, em conjunto com os dados do GEOSAMPA, ferramenta que disponibiliza os dados geográficos da cidade de São Paulo, fazendo uso dos dados em dois modelos de regressão, Árvore de Decisão e *Random Forest*.

1.1 Justificativa

A motivação chave para o projeto foi entender qual o modelo de inteligência artificial que se adaptaria aos dados de roubo e furto na cidade de São Paulo do ano 2017 a 2022 conforme o portal (TRANSPARÊNCIASP) e em conjunto aos dados do GEOSAMPA.

Através desse estudo gerar previsões dos meses onde são registrados maiores números de delitos para que assim a Segurança Pública da Cidade de São Paulo venha atuar com maior contingente de servidores, e para que empresas do segmento tivessem indicadores para tomada de decisão estratégicas referente a melhores locais de locações, entre outros.

Mediante a esses indicadores, empresas de tecnologia poderiam criar aplicativos que possam identificar os locais com mais incidência dessa natureza delito através da latitude x longitude com parceria das operadoras de telefonia para ajudar a segurança pública em conjunto com a população.

1.2 Objetivo geral

O objetivo principal deste estudo é analisar os dados de roubos e furtos de *smartphones* na cidade de São Paulo, treinar e avaliar dois modelos de aprendizado de máquina, Árvore de Decisão e *Random Forest* e determinar o melhor modelo para se adaptar ao volume de dados disponíveis. Com a ajuda do Geosampa, também será possível determinar os limites geográficos de cada distrito, permitindo uma análise mais detalhada da distribuição dos crimes na cidade. Esperamos que este estudo contribua para a compreensão e prevenção de tais ocorrências na cidade de São Paulo.

1.3 Objetivos específicos

- Estudo para prever pontos críticos de roubo e furto de celulares na cidade de São Paulo
- Treinar um modelo que interprete corretamente os dados.

2. REVISÃO BIBLIOGRÁFICA

Nesta seção, serão apresentadas as principais opiniões de especialistas, conclusões de estudos prévios e conceitos realizados por outras instituições referentes aos temas centrais utilizados como embasamento deste trabalho, que abrange o tema principal “Utilização de inteligência artificial na predição de roubos e furtos”, expondo informações sobre a sua utilização e relatando também informações dos principais métodos envolvidos no uso de inteligência artificial que auxiliam na análise desejada para este estudo.

A teoria dos geradores e atratores de crimes tem sido amplamente explorada no campo da criminologia ambiental, visando compreender a relação entre o ambiente físico e as atividades criminosas. Essa abordagem sugere que os crimes não ocorrem aleatoriamente no espaço e tempo, podendo ser estudados e analisados para identificar padrões e comportamentos criminais. Bernasco e Block (2011) destacam a importância dos geradores de crimes como locais de fácil acesso ao público, onde a presença de grandes grupos de pessoas cria oportunidades para a prática de delitos. Centros comerciais, escolas secundárias e estações de transporte público são exemplos típicos de geradores de crimes. Esses lugares concentram pessoas e possuem características que podem atrair criminosos, como a disponibilidade de bens de valor e a presença de potenciais vítimas distraídas.

Por outro lado, os atratores de crimes são locais que, embora não atraiam necessariamente grandes grupos de pessoas, apresentam características específicas que os tornam propícios para que infratores motivados encontrem vítimas ou alvos atraentes e mal protegidos. Becos escuros e isolados são exemplos de atratores de crimes, oferecendo oportunidades para agressões ou roubos sem serem facilmente observados por outras pessoas. A análise espacial dos dados criminais, combinada com técnicas estatísticas e algoritmos de aprendizado de máquina, tem se mostrado promissora para identificar áreas de maior concentração de crimes e direcionar medidas preventivas de forma mais eficiente.

No entanto, é importante ressaltar que a previsão de crimes ainda é um campo em desenvolvimento, e desafios como questões éticas, privacidade dos dados e enviesamento dos algoritmos precisam ser abordados para garantir o uso responsável dessas técnicas na área da segurança pública.

2.1 Previsão e séries temporais

Uma previsão é a predição de algum evento futuro. A previsão em séries temporais busca identificar e modelar os padrões presentes nos dados históricos para realizar estimativas precisas das futuras observações. Existem várias abordagens e técnicas disponíveis para realizar essa tarefa, como suavização exponencial, modelos autorregressivos (AR), modelos de média móvel (MA), modelos autorregressivos de média móvel (ARMA) e modelos autorregressivos integrados de média móvel (ARIMA), entre outros (Montgomery et al., 2015).

2.2 Aprendizado de máquina

Aprendizado de Máquina é uma área de Inteligência Artificial (AI) cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado, bem como a construção de sistemas capazes de adquirir conhecimento de forma automática. Um sistema de aprendizado de máquina é um programa de computador que toma decisões com base em experiências acumuladas por meio da solução bem-sucedida de problemas anteriores. (MONARD; BARANAUSKAS, 2003).

Segundo Mitchell (1997), Aprendizado de Máquina pode ser definido como a capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência. Desta forma, no Aprendizado de Máquina, os computadores são programados para aprender com a experiência anterior.

Monard e Baranauskas (2003, p. 90) afirmam que o Aprendizado de Máquina é uma ferramenta poderosa para a aquisição automática de conhecimento. Porém, deve-se observar que não existe uma única técnica que apresente o melhor desempenho para todos os problemas. Sendo assim, “é importante compreender o poder e a limitação dos diversos algoritmos de Aprendizado de Máquina utilizando alguma metodologia que permita avaliar os conceitos induzidos por esses algoritmos em determinados problemas”.

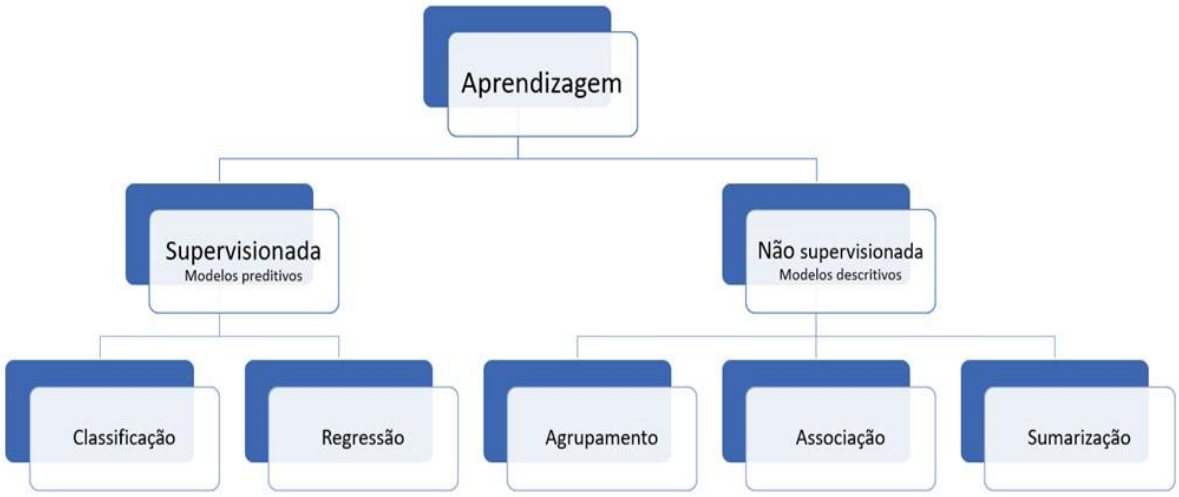
No contexto mencionado, o Aprendizado de Máquina compartilha uma semelhança com o aprendizado humano, pois ambos são processos baseados em experiências. Assim como os seres humanos, as máquinas aprendem ao reconhecer padrões em dados. A quantidade de dados disponíveis para a máquina é fundamental, pois quanto maior a quantidade de dados fornecidos, mais precisa será a capacidade da máquina em identificar e reconhecer esses padrões.

Em outras palavras, alimentar a máquina com uma maior quantidade de dados resultará em um aprendizado mais preciso e eficiente, pois a máquina terá mais exemplos para aprender e generalizar a partir deles. Portanto, a quantidade de dados disponíveis desempenha um papel crucial no processo de aprendizado da máquina.

Para os autores Muller e Guido (2017), as aplicações de métodos de *Machine Learning*, nos últimos anos, se tornaram presentes na vida cotidiana. Desde as recomendações automáticas de quais filmes assistir, que comida pedir ou quais produtos comprar, rádio online personalizado, reconhecimento de amigos por meio de fotos etc. Isso se tornou possível devido a muitos sites e dispositivos modernos possuírem algoritmos de Aprendizado de Máquinas em seus núcleos. Sites complexos como o Facebook, Amazon e Netflix e aplicações comerciais contêm modelos de Aprendizado de Máquinas implementados. Além disso, o Aprendizado de Máquina teve uma tremenda influência nos sistemas de Recuperação da Informação e na forma como a pesquisa é feita atualmente.

Por consequência, os algoritmos de Aprendizado de Máquina são amplamente utilizados em diversas tarefas, que podem ser organizadas de acordo com o método de aprendizado a ser adotado. Com base nos critérios escolhidos para a realização de cada atividade, os métodos de aprendizado podem ser divididos em aprendizado supervisionado e aprendizado não supervisionado conforme ilustrado na figura 1.

Figura 1 – Hierarquia do Aprendizado



Fonte: ADAPTAÇÃO DE MONARD E BARANAUSKAS (2003)

2.3 Aprendizado supervisionado

Segundo Matos (2015), o termo aprendizado supervisionado é usado sempre que o programa é “treinado” sobre um conjunto de elementos pré-estabelecidos. “Baseado no treinamento com os dados pré-definidos, o programa pode tomar decisões precisas quando recebe novos dados”. Aprendizado supervisionado “refere-se à capacidade que determinados algoritmos têm de aprender a partir de exemplos.” (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

Ou seja, no aprendizado supervisionado, a máquina é alimentada com uma amostragem de dados, onde cada exemplo é rotulado de acordo com suas características dominantes. Essa abordagem permite que o algoritmo aprenda a partir dessas amostras, assimilando a relação entre as características dos dados e suas respectivas categorias ou rótulos. Podemos visualizar esse processo como se a máquina estivesse recebendo orientações de um professor, que fornece informações explícitas sobre as respostas corretas. É por isso que esse tipo de aprendizado é chamado de "supervisionado", pois a máquina recebe uma supervisão externa durante o processo de aprendizado. Essa orientação permite que a máquina generalize e faça previsões precisas em relação a novos dados, utilizando as informações aprendidas durante o treinamento com as amostras rotuladas.

Compreende-se, assim, que nesse método é fornecido ao algoritmo de aprendizado, ou indutor, um conjunto de exemplos de treinamento para os quais o rótulo da classe associada é conhecido (CHEESEMAN; STUTZ, 1996).

O aprendizado supervisionado compreende a abstração de um modelo de conhecimento a partir dos dados apresentados na forma de pares ordenados (entrada, saída desejada). Por entrada entende-se o conjunto de valores das variáveis (atributos) de entrada do algoritmo para um determinado caso. Tais variáveis são denominadas atributos previsores. A saída desejada corresponde ao valor de uma variável (denominada atributo-alvo) que se espera que o algoritmo possa produzir sempre que receber os valores especificados em entrada. (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

Para Markov e Larose (2007), a partir de uma coleção de objetos identificados com uma classe, o sistema de aprendizagem cria um mapeamento entre os elementos e as classes que pode ser usado para classificar novos objetos que ainda não foram rotulados. Como a rotulagem (classificação) do conjunto inicial (treinamento) é feita por um agente externo ao sistema, essa configuração é chamada de aprendizagem supervisionada. Assim, os algoritmos desse modelo são treinados usando exemplos rotulados, tendo um conjunto de dados de entrada e as possíveis saídas desejadas. Ou seja, o algoritmo de aprendizagem recebe um conjunto de dados de entrada junto com as saídas corretas correspondentes. Com base em uma coleção de exemplos, o programa faz previsões, buscando padrões nos rótulos. Ao encontrar o melhor padrão possível, ele utilizará essa referência para fazer predições para dados de testes sem rótulo.

O aprendizado supervisionado requer uma função de aprendizado dos dados de treinamento fornecidos como entrada. No caso da classificação de textos, os dados de treinamento são compostos de pares do tipo documento-classe indicando as classes corretas para os documentos dados, de acordo com especialistas humanos. Esses dados de treinamento são então usados para aprender uma função de classificação, a qual pode ser usada para fazer predições de classes para dados ainda não vistos ou novos. A abordagem só funciona se a função aprendida for tal que ela possa pegar dados que não foram vistos antes e prever classes para eles com alta precisão. (BAEZ-AYATES; RIBEIRONETO, 2013).

Esse modelo de aprendizagem se divide em duas subcategorias: Regressão e Classificação.

2.4 Regressão

De acordo com Michie et al. (1994), “a regressão compreende a busca por uma função que mapeie os registros de um banco de dados em um intervalo de valores reais. Esta tarefa é similar à tarefa de classificação, com a diferença de que o atributo-alvo assume valores numéricos.” (citado por GOLDSCHMIDT; PASSOS; BEZERRA, 2015, p. 25). A técnica de regressão é comumente usada para modelar relações complexas entre elementos de dados fazendo estimativa de uma variável a partir da outra.

O modelo de regressão busca fazer previsões, porém seus resultados são derivados de experiências ou conhecimentos anteriores. É uma subcategoria de aprendizagem supervisionada usada quando o valor que está sendo previsto difere de um “sim ou não” e segue um espectro contínuo. Por exemplo, dada uma imagem de um homem ou de uma mulher, precisa-se prever sua idade com base nas informações da imagem.

Dois modelos de regressão amplamente utilizados em aprendizado de máquina supervisionado e que ambos foram utilizados neste trabalho, são respectivamente a Árvore de Decisão e a Floresta Aleatória (*Random Forest*).

2.5 Classificação

De acordo com Gonçalves (2013), desde os tempos antigos, dos primeiros dias da Grande Biblioteca de Alexandria, por volta de 300 a.C., os bibliotecários tinham que lidar com armazenamento de documentos para recuperação e leitura futura. Com o passar do tempo, o tamanho das coleções cresceu e o problema tornou-se cada vez mais difícil. Procurar por um livro em particular dentre centenas de livros tornou-se uma tarefa tediosa, demorada e

impráticável. Para aliviar o problema, os bibliotecários começaram a rotular os documentos. Isso forneceu metainformação ao seu conteúdo, permitindo assim que os livros fossem organizados com uma visão que permitia busca rápida e recuperação. Uma das primeiras abordagens para rotular documentos foi atribuir um identificador único para cada documento. Isso resolvia o problema sempre que o usuário soubesse dos identificadores dos livros que eles queriam, mas não resolvia o problema mais genérico de encontrar documentos sobre um assunto ou tópico específico. Nesse caso, a solução natural é agrupar os documentos por tópicos comuns e nomear cada grupo com um ou mais rótulos significativos. Cada grupo rotulado é o que chamamos de uma classe, isto é, um conjunto no qual podemos inserir documentos cujo conteúdo pode ser descrito pelo seu rótulo.

Esse processo de inserção dos documentos em classes é comumente conhecido como classificação de textos. Assim como esse método foi utilizado para classificar textos, ele pode ser aplicado também, para classificação em modelos estatísticos.

Desta forma, com a finalidade de aprimorar os recursos para organização da informação, estudos foram realizados e algoritmos foram desenvolvidos para a indução automática de sistemas capazes de lidar com problemas de classificação.

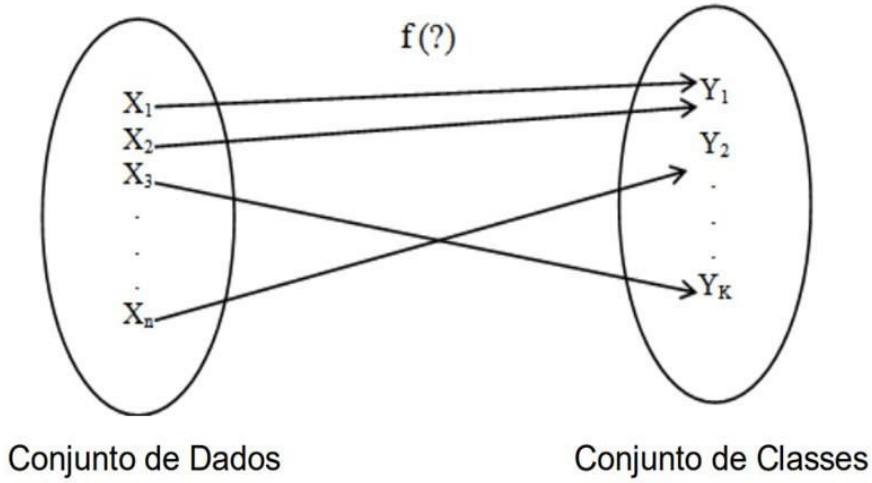
Segundo Jacob (2004), os sistemas de classificação são usados para organizar o conhecimento. A classificação como processo envolve a atribuição ordenada e sistemática de cada entidade a uma e apenas uma classe dentro de um sistema de classes mutuamente exclusivas (i.e., não sobrepostas). Este processo é legal e sistemático: lícito porque é realizado de acordo com um conjunto estabelecido de princípios que governam a estrutura de classes e as relações entre elas; e sistemática porque exige a aplicação consistente desses fundamentos dentro da estrutura de uma ordenação prescrita da realidade.

O esquema em si é artificial e arbitrário: artificial porque é uma ferramenta criada com o propósito expresso de estabelecer uma organização significativa; e arbitrário, porque os critérios usados para definir classes no esquema refletem uma perspectiva única do domínio, excluindo todas as outras perspectivas (JACOB, 2004).

Em resumo, na tarefa de classificação, um dos grupos possui somente um atributo, que corresponde ao atributo-alvo, ou seja, a propriedade pela qual se deve fazer a predição de um valor. Nesse caso, o atributo é categórico (domínio composto por categorias/classes) e o outro conjunto contém os atributos a serem utilizados na predição do valor, denominados previsores ou de predição.

De acordo com Goldschmidt, Passos e Bezerra (2015), a tarefa de classificação pode ser compreendida como “a busca por uma função que permita associar corretamente cada registro Xi de um conjunto de dados a um único rótulo categórico Yi, denominado classe. Uma vez identificada, essa função pode ser aplicada a novos registros de forma a prever as classes em que tais registros se enquadram”. A Figura 2 ilustra as associações entre registros de dados e suas respectivas classes.

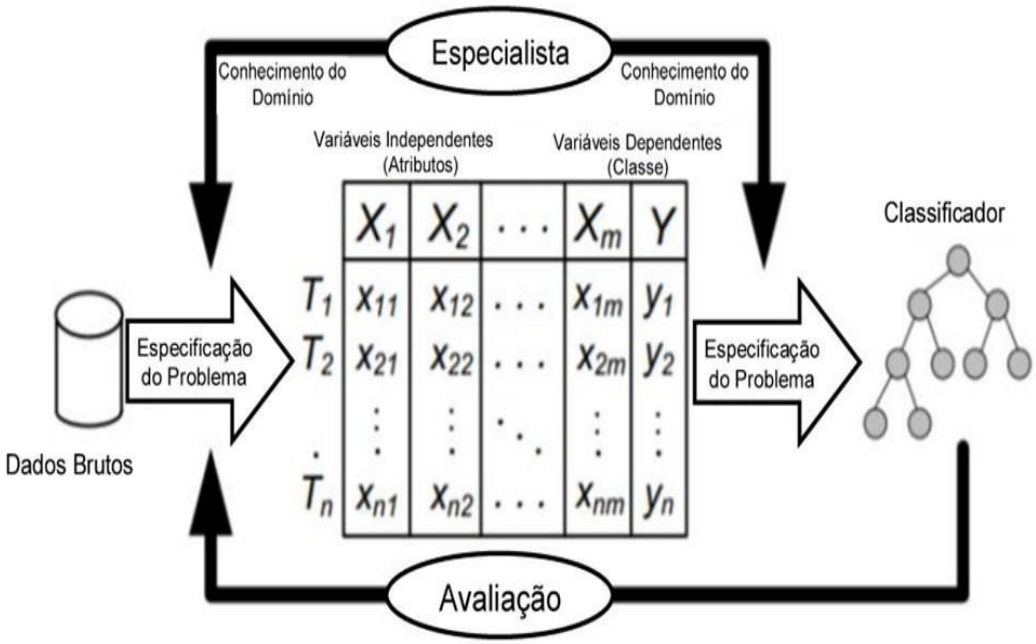
Figura 2 – Associações entre registros de dados e classes



Fonte: GOLDSCHMIDT, PASSOS E BEZERRA (2015)

Em suma, na classificação, o algoritmo de aprendizado constrói um classificador capaz de determinar corretamente a classe de novos exemplos ainda não etiquetados, dado um conjunto de classe e um conjunto de exemplos de treinamento. Na maioria dos casos esse processo pode usar o conhecimento de um domínio para fornecer alguma informação previamente conhecida como entrada ao indutor. E, após induzido, o classificador é normalmente avaliado e o processo de classificação pode ser repetido, se necessário. (MONARD; BARANAUSKAS, 2003). A figura a seguir ilustra esse processo

Figura 3 – Processo de Classificação



Fonte: MONARD; BARANAUSKAS, 2003.

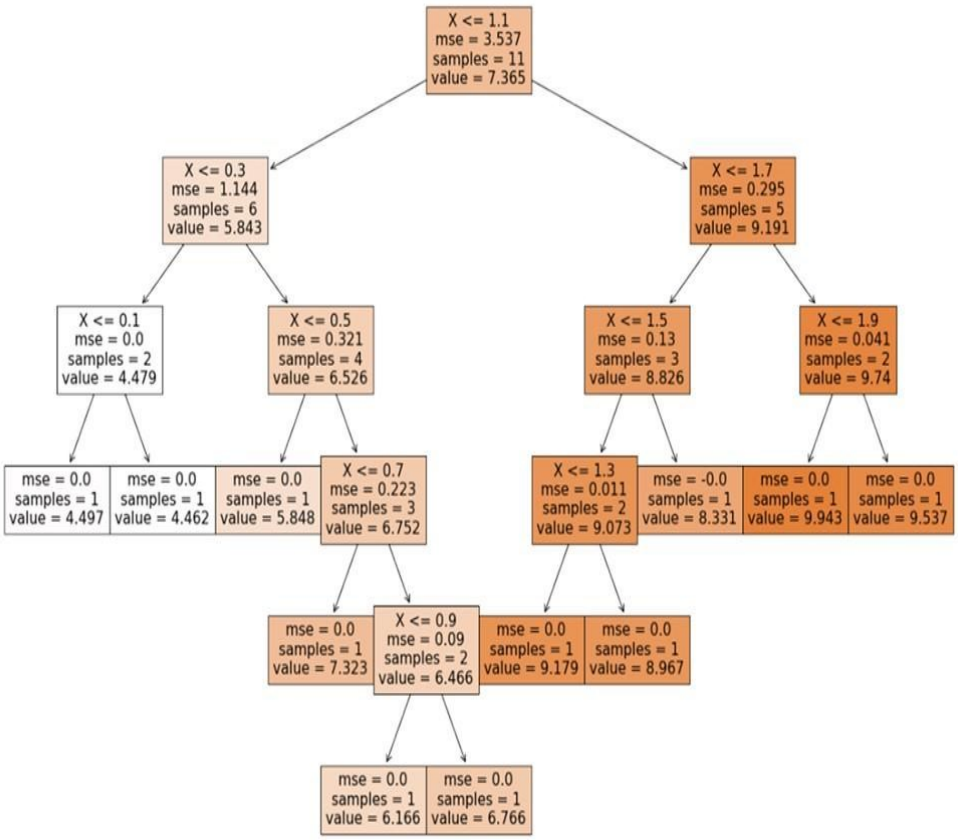
Assim, “após a etapa de treinamento, tem-se um classificador que deve ser capaz de predizer corretamente o rótulo de novos exemplos, que ainda não foram rotulados.” (REZENDE, 2005 apud BORGES, 2012).

De acordo com Monard e Baranauskas (2003), normalmente, um conjunto de exemplos é dividido em dois subconjuntos: A coleção de treinamento é utilizada para o aprendizado do conceito e a de testes usada para medir o grau de efetividade do conceito aprendido. “Esses conjuntos são normalmente disjuntos para assegurar que as medidas obtidas, utilizando o conjunto de testes, sejam de um conjunto diferente do usado para realizar o aprendizado, tornando a medida estatisticamente válida.” (MONARD, BARANAUSKAS, 2003).

2.6 Árvore de decisão

A técnica de árvore de decisão é um modelo que recebe como entrada um conjunto de atributos e produz uma única "decisão" ou valor de saída (Russell e Norvig, 2009). Esses valores de entrada e saída podem ser tanto discretos quanto contínuos, e a árvore de decisão alcança sua decisão por meio de uma sequência de testes. Em uma árvore de decisão, cada nó interno corresponde a um teste em um atributo de entrada e as ramificações são definidas com base nos possíveis valores desse atributo. Por fim, cada nó de folha na árvore especifica o valor que deve ser retornado pela função.

Figura 6 – Exemplo Geral de uma Árvore de Decisão



Fonte: PREVISÃO DE PONTOS CRÍTICOS DE CRIME (2021)

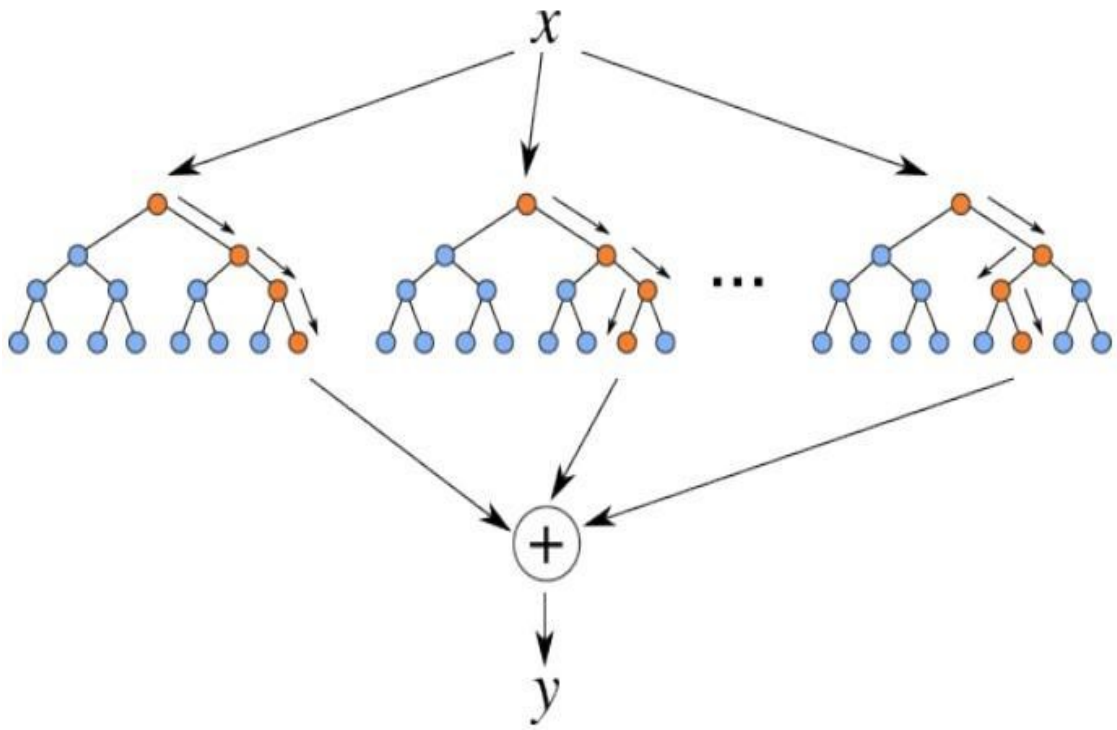
2.7 Floresta aleatória

Floresta Aleatória é um algoritmo de aprendizado de máquina que combina várias árvores de decisão para realizar previsões ou classificações (Breiman, 2001). É uma técnica poderosa e amplamente utilizada que oferece maior robustez e precisão em comparação com uma única árvore de decisão.

Neste algoritmo, várias árvores de decisão são criadas usando diferentes subconjuntos aleatórios dos dados de treinamento e recursos (variáveis de entrada). Cada árvore de decisão individual na floresta opera de forma independente, realizando testes em atributos de entrada e fornecendo uma decisão ou valor de saída. A decisão final da floresta aleatória é determinada pela combinação das decisões de todas as árvores, seja por meio de votação majoritária (no caso de classificação) ou média (no caso de regressão).

A floresta aleatória possui várias vantagens, como redução do *overfitting* (ajuste excessivo aos dados de treinamento), capacidade de lidar com dados de entrada com alta dimensionalidade e robustez em relação a valores ausentes ou ruidosos nos dados. Além disso, a floresta aleatória permite a avaliação da importância relativa dos atributos de entrada na tomada de decisão.

Figura 7 – Exemplo Geral de uma Floresta Aleatória



Fonte: TELOKEN (2016)

Na imagem acima, temos um exemplo de floresta aleatória onde, partindo do elemento X, uma entrada de base de dados, são geradas diversas árvores de decisão onde cada árvore é treinada com uma amostra aleatória dos dados de entrada. A decisão final da floresta aleatória é baseada na combinação dos resultados de todas as árvores. Essa combinação pode ser feita de diferentes maneiras, dependendo do problema em questão.

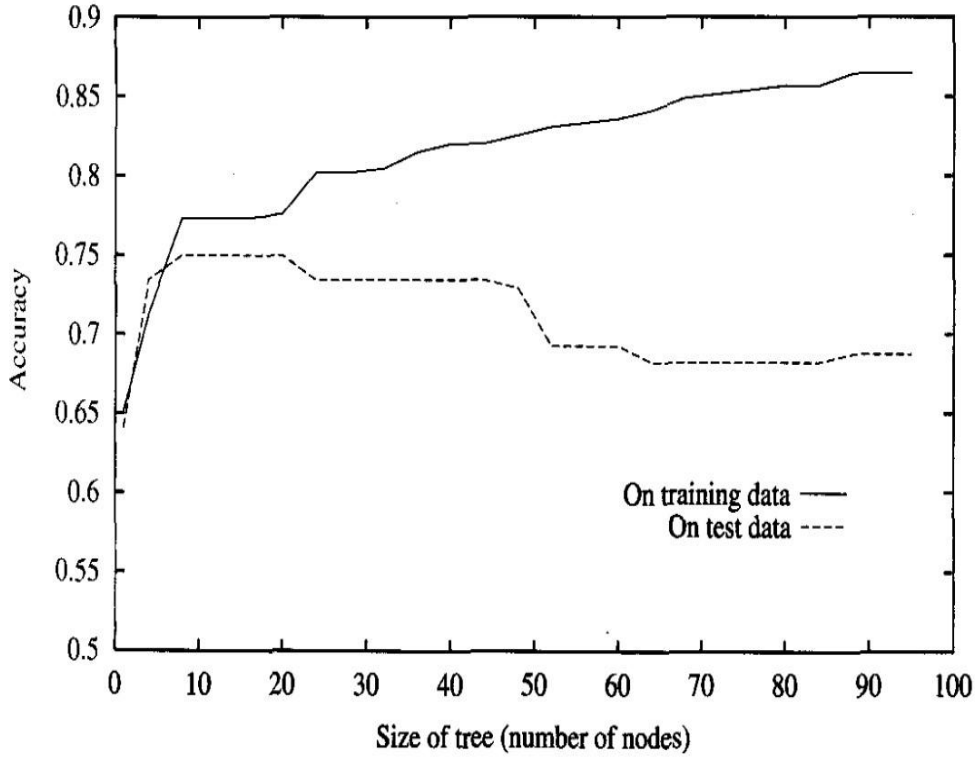
2.8 Overfitting

Overfitting é considerado um grande problema dos modelos de Aprendizado de Máquina (AM), devido a facilidade em que este problema pode ocorrer caso não se tome cuidado ao gerar um modelo. Mitchell (1997) fornece uma definição formal sobre o problema:

> Dada um modelo H, é dito que H causa overfitting no conjunto de dados de treinamento se existe uma segunda hipótese H', em que a taxa de erros de H < H', em relação aos dados de treinamento, mas a taxa de erros de H' < H em relação ao conjunto total de dados. (Mitchell ,1997, p.67).

Em outras palavras, é importante entender que um modelo que apresenta alta precisão em um conjunto de dados de treinamento não garante necessariamente um bom desempenho em novas instâncias. Um exemplo disso é um classificador que alcança 100% de precisão nos dados de treinamento, mas apenas 50% de precisão nas instâncias de teste, o que indica um problema de ajuste excessivo (*overfitting*) aos dados de treinamento. Isso significa que o modelo está se adaptando muito bem aos dados específicos do treinamento, mas não consegue generalizar adequadamente para novos dados. É fundamental encontrar um equilíbrio entre o ajuste aos dados de treinamento e a capacidade de generalização para garantir um desempenho consistente em diferentes conjuntos de dados.

Figura 4 – Exemplo Geral de *overfitting* em um modelo de Árvore de Decisão



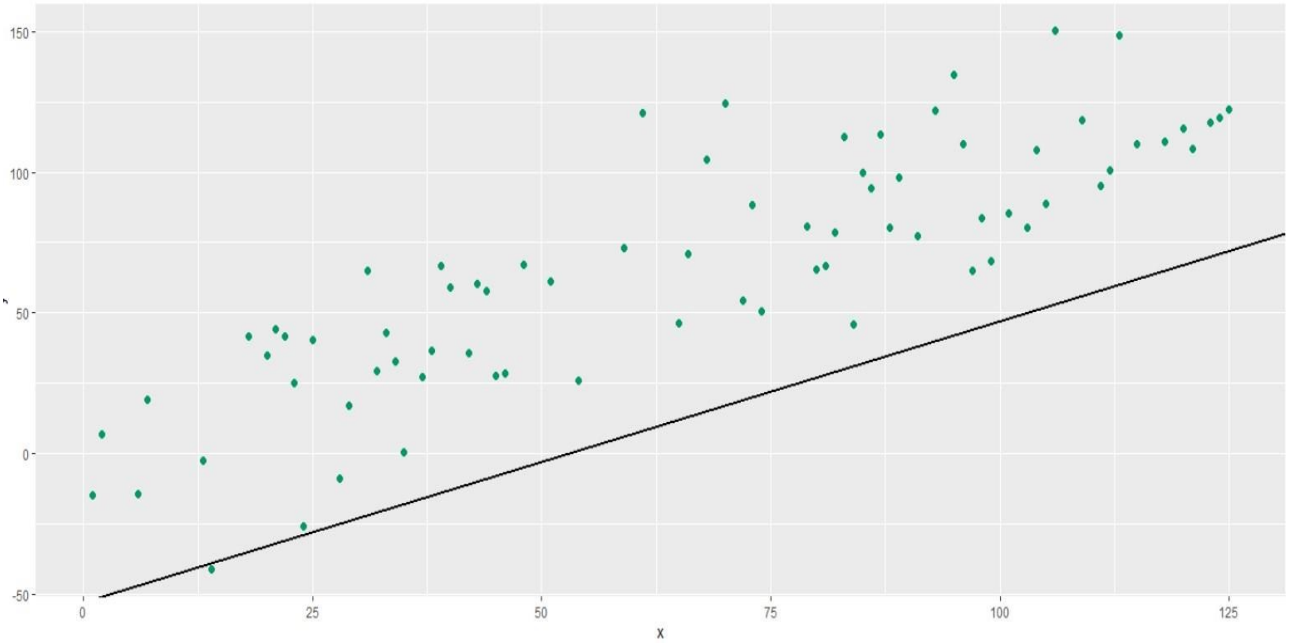
Fonte: MITCHELL (1997)

Algumas, outras causas também podem levar a esse problema, tais como a falta de representatividade do conjunto de treinamento em relação à base de dados completa ou o uso de um conjunto de treinamento muito pequeno. Portanto, é crucial selecionar os dados de treinamento com cautela, a fim de evitar esse tipo de problema.

2.9 Underfitting

De acordo com Monard e Baranauskas (2003), se poucos exemplos representativos forem dados ao sistema de aprendizado ou o usuário pré-defina o tamanho do classificador como muito pequeno (por exemplo, um número insuficiente de neurônios e conexões são definidos em uma rede neural ou um alto fator de poda é definido para uma árvore de decisão) ou uma combinação de ambos. Portanto, é também possível induzir hipóteses que possuam uma melhora de desempenho muito pequena no conjunto de treinamento, assim como em um conjunto de testes. Neste caso, diz-se que a hipótese ajusta-se muito pouco ao conjunto de treinamento ou que houve um *underfitting*.

Figura 5 – Exemplo Geral de *underfitting*



Fonte: DIDÁTICA TECH (2020)

O *underfitting* ocorre normalmente quando o modelo gerado por determinado algoritmo não consegue trabalhar bem com as instâncias do conjunto de dados utilizados.

Portanto, além de se escolher com cuidado os dados de treinamento, deve-se verificar se determinado algoritmo é capaz de solucionar o problema proposto.

2.10 Estudos na área de Segurança Pública

De acordo com (UEDA, 2021), onde o autor realizou um estudo sobre modelagem por aprendizado de máquina para prevenção de roubos de celulares em São Paulo, a implementação do modelo possibilita beneficiar a prática do policiamento preventivo, permitindo a alocação eficiente de diversos recursos policiais, uma vez que esses são escassos. Com o policiamento em áreas onde há mais criticidade de ocorrências é possível dificultar a ação dos criminosos. Outro benefício seria a caráter informativo para a população como um alerta para áreas de risco de roubos de celulares, a previsão dos pontos críticos de crime abre a possibilidade de evitar a esses locais e auxiliar na tomada de decisão de utilizar o celular, mesmo que a área tenha um alto índice de crime.

3. METODOLOGIA

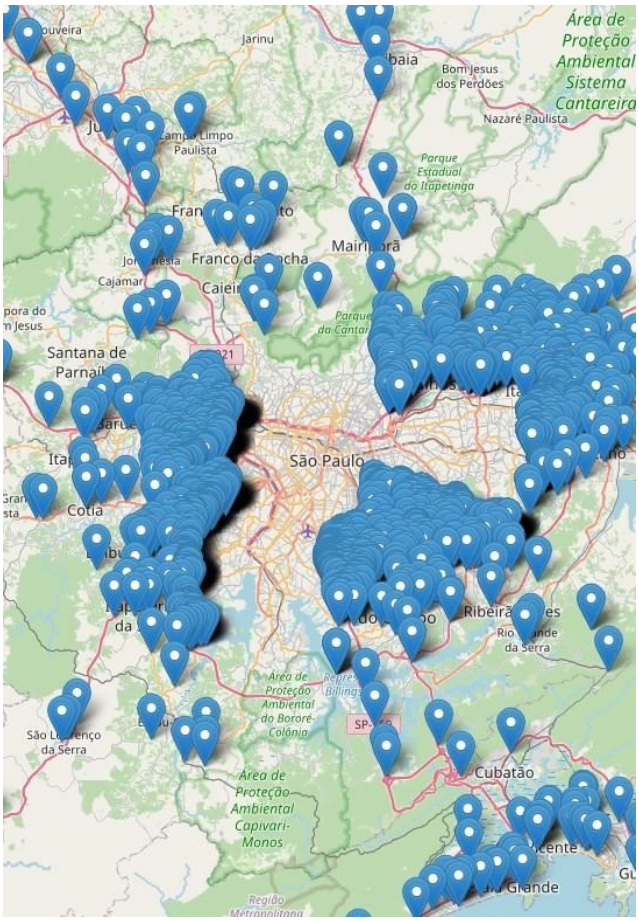
Na condução deste estudo, utilizou-se a linguagem de programação Python, dada a sua capacidade para a manipulação de dados geoespaciais e a implementação de algoritmos de aprendizado de máquina (LOPES et al. ,2021). Os dados de interesse foram coletados diretamente da base de dados disponibilizada pela Secretaria de Segurança Pública de São Paulo, que contém registros de 2017 a 2022 e do portal GEOSAMPA da Prefeitura de São Paulo. Esses dados foram cuidadosamente filtrados para garantir que apenas os incidentes de roubo e furto de *smartphones* ocorridos dentro dos limites da cidade de São Paulo fossem considerados.

3.1 Pré-processamento dos dados

O primeiro passo na análise dos dados foi o pré-processamento. Nessa fase, os dados foram limpos e preparados para a modelagem. O conjunto de dados original foi obtido do portal de dados abertos da Prefeitura de São Paulo e contém informações detalhadas sobre os crimes na cidade, incluindo a data, localização e mais dados sobre a ocorrência.

No processo de limpeza de dados, identificaram-se e removeram-se registros incompletos ou inconsistentes. Também foram identificados e removidos registros duplicados. Na base de dados original, a duplicidade significava que mais de um telefone havia sido roubado ou furtado em um único incidente. A remoção desses registros ajudou a evitar um aumento artificial no número de casos em determinadas regiões.

Mapa 1 - Ponto de ocorrências fora da cidade de São Paulo

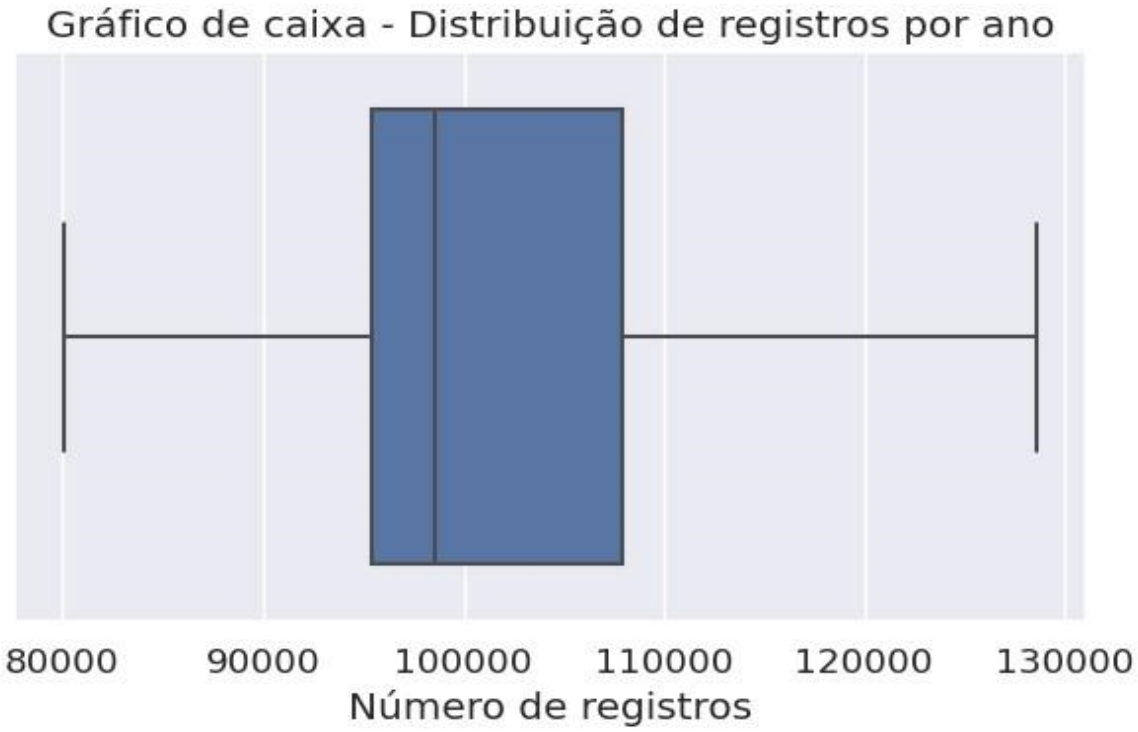


Fonte: Elaborado pelos autores (2023)

Entre esses registros, havia uma variedade de 72 valores para o ano em que a ocorrência efetivamente aconteceu. Para evitar inconsistências nos modelos de aprendizado de máquina, optou-se por filtrar os dados incluindo apenas as ocorrências com a data do acontecimento dentro do período de 2017 a 2022.

O gráfico a seguir apresenta o *boxplot* representando o número de registros por ano depois da aplicação desse filtro, respectivamente. O Gráfico 1, é representado de maneira regular pois são apenas os dados que ocorreram entre 2017 e 2022 que estão de acordo com o objetivo do estudo. Sendo que variam entre no mínimo 80.000 registros a até 130.000 registros no ano.

Gráfico 1 - Distribuição de ocorrências de roubo e furto de celular que ocorreram na cidade de São Paulo entre 2017 e 2022.



Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO,2023)

Os dados foram posteriormente agrupados por mês e ano. O gráfico 2 demonstra a quantidade de roubos e furtos por mês. Em seguida, os dados foram divididos em conjuntos de treinamento e teste, com 80% dos dados usados para treinamento e 20% para teste.

Gráfico 2 - Distribuição da totalidade de ocorrências por mês de roubo e furto de celular na cidade de São Paulo entre 2017 e 2022.

Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO,2023)

Resumindo, a base de dados continha 1.215.412 registros de roubos e furtos de *smartphones* ocorridos na cidade de São Paulo. No entanto, diversas exclusões foram necessárias para atender aos critérios do estudo.

Primeiro, 194.168 registros duplicados foram identificados e removidos. Em seguida, 2.028 registros, apesar de listados como pertencentes à cidade de São Paulo, foram identificados como originários de outros locais e, portanto, também excluídos. Um total de 242.527 registros correspondentes a eventos que ocorreram antes do período de estudo, de 2017 a 2022, também foram excluídos para garantir a relevância e precisão dos dados. Além disso, parte da base de dados inicial não apresentava informações completas nas colunas LONGITUDE e LATITUDE. O total de registros que não possuíam estas colunas foi de 165.414.

Ao final deste processo de limpeza e tratamento dos dados, permaneceram 611.275 registros, que foram utilizados para treinar e testar os modelos de *machine learning*. Embora esse número represente aproximadamente 50% do número de registros original, essa redução foi crucial para garantir a qualidade dos dados e, por consequência, a confiabilidade do modelo de previsão.

Tabela 1 – Resumos do total de dados

Dados	Número de dados
Todos registros	1.215.412
Registros sem latitude e longitude	165.414
Registros duplicados	194.168
Registros anteriores a 2017	242.527
Registros fora da cidade de São Paulo	2.028
Registros que serão utilizados	611.275

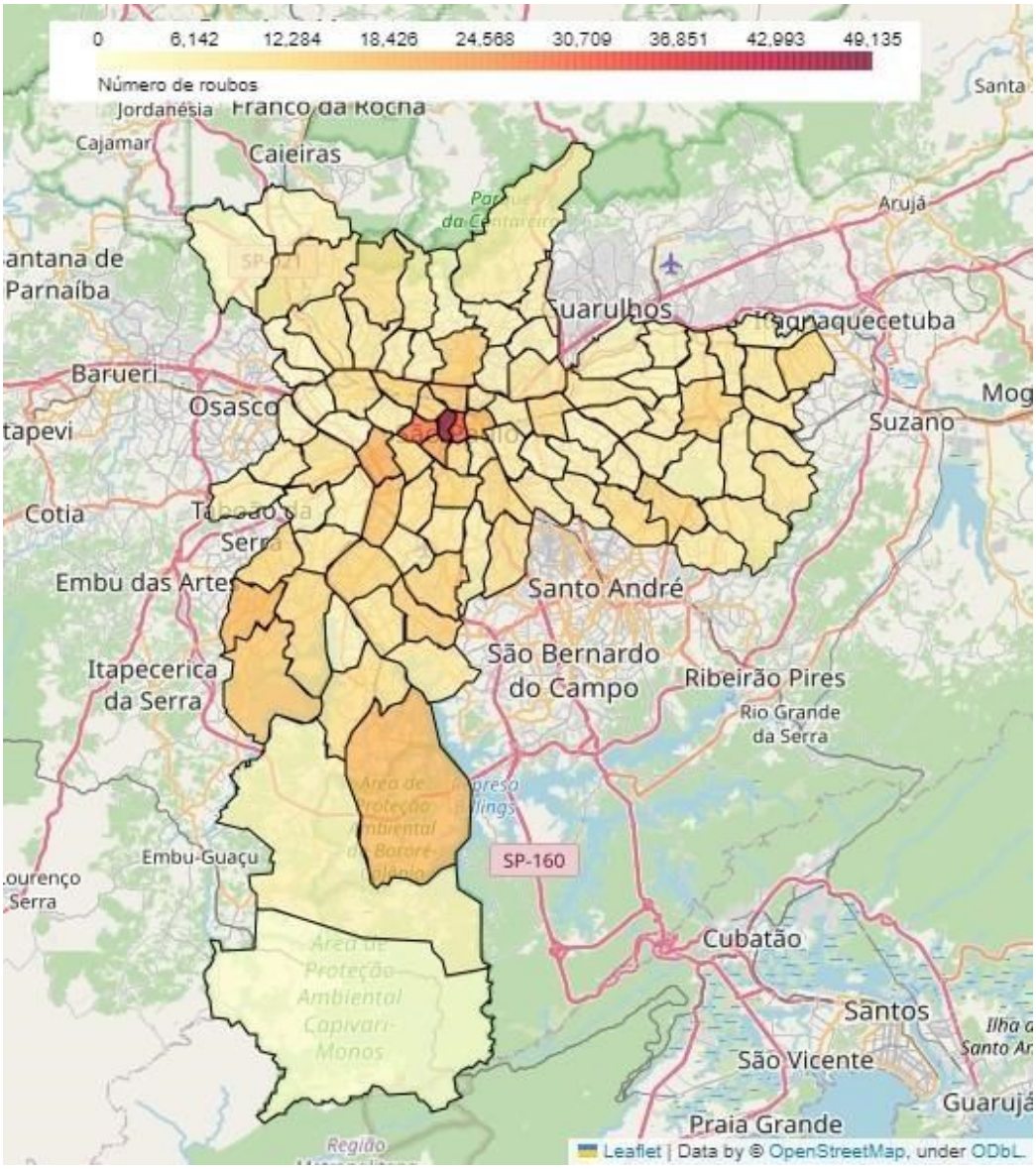
Fonte: Elaborado pelos autores (2023)

3.2 Modelagem dos dados

Os modelos de aprendizado de máquina empregados para a modelagem de dados foram a Árvore de Decisão e a Floresta Aleatória. Os mesmos foram treinados para prever o número de roubos e furtos de celulares.

O recurso ao sistema de informações geográficas GEOSAMPA foi duplamente benéfico para a pesquisa. Em primeiro lugar, foi utilizado para verificar a localização de cada incidente registrado, assegurando assim a fidelidade geográfica dos dados. Em segundo lugar, permitiu a segmentação dos dados por distrito, proporcionando uma análise mais detalhada e a elaboração de um mapa de calor georreferenciado que destaca as áreas de maior incidência de roubos e furtos de celulares é possível verificar no Mapa 2.

Mapa 2 - Distribuição das ocorrências por mês de roubo e furto de celular na cidade de São Paulo entre 2017 e 2022.



Fonte: Secretaria de Segurança Pública do Estado de São Paulo, portal da Transparência, (SÃO PAULO,2023)

Em relação às variáveis dos dados, as informações geográficas de cada distrito, incluindo área, perímetro, coordenadas x e y do centro dos distritos, foram extraídas dos dados do GEOSAMPA com o objetivo de proporcionar maior flexibilidade aos modelos de aprendizado de máquina. Resultado nas seguintes características de entrada mês, área do distrito e as coordenadas do centro do distrito, conforme a Tabela 2.

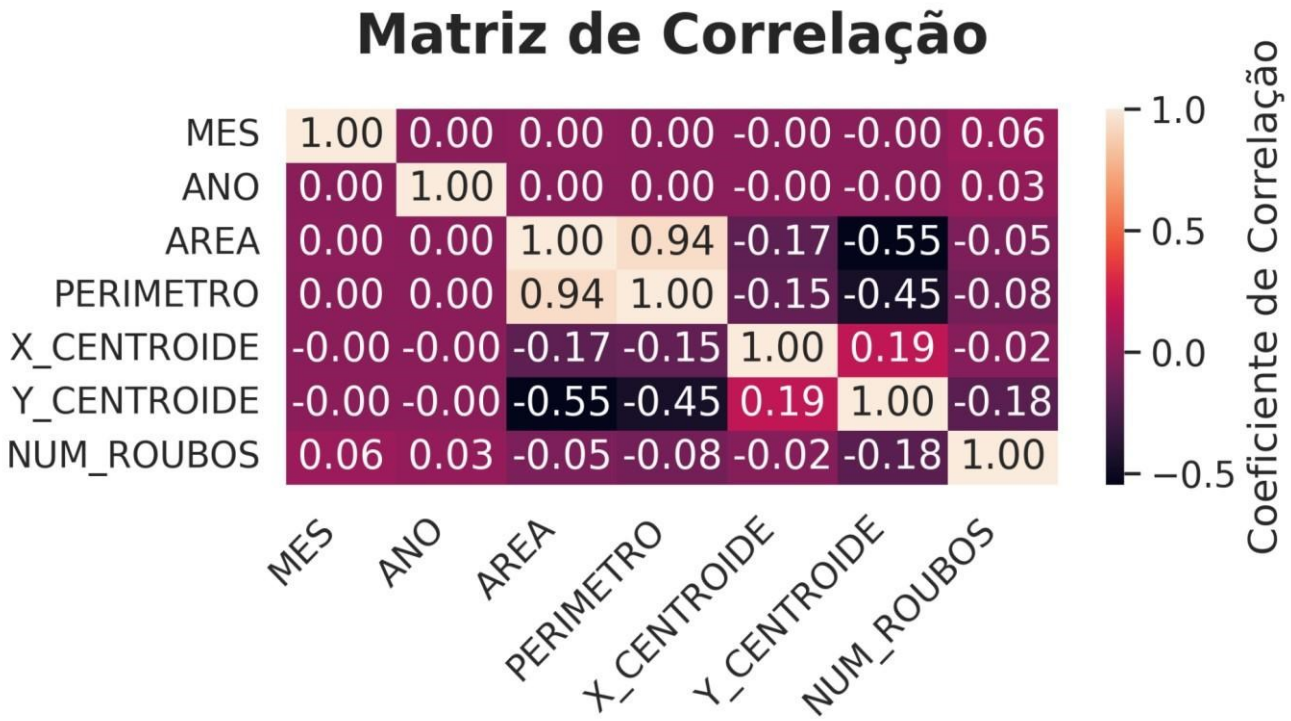
Tabela 2 – Características de entrada

Variável	Descrição
Mês	O mês das ocorrências
Área	A área do distrito
Coordenada X	O ponto X no mapa que identifica o centro do distrito
Coordenada Y	O ponto Y no mapa que identifica o centro do distrito

Fonte: Elaborado pelos autores (2023)

A informação do perímetro não foi incluída no modelo devido à sua alta correlação com a área na Figura 8 é possível verificar a correlação entre os dados.

Figura 8 - Matriz de correlação das variáveis para treinar e testar o modelo.



Fonte: Elaborado pelos autores (2023)

A modelagem de aprendizado de máquina foi inicialmente realizada através do algoritmo de Árvore de Decisão e na sequência foi avaliado o modelo *Random Forest*.

3.3 Avaliação do modelo

Os modelos foram avaliados com base em duas métricas principais: o coeficiente de determinação (R^2) e o erro absoluto médio (MAE). O R^2 mede a proporção da variância na variável de resposta que é previsível a partir das características, enquanto o MAE mede a média das diferenças absolutas entre as previsões e os valores reais.

Em suma, a metodologia empregada permitiu a construção de um sistema de previsão de roubos de *smartphones*, possibilitando a análise dos registros e a implementação de algoritmos de aprendizado de máquina adequados.

A incorporação de variáveis geográficas ao modelo ofereceu uma nova dimensão à análise, que poderia ser explorada ainda mais em estudos futuros.

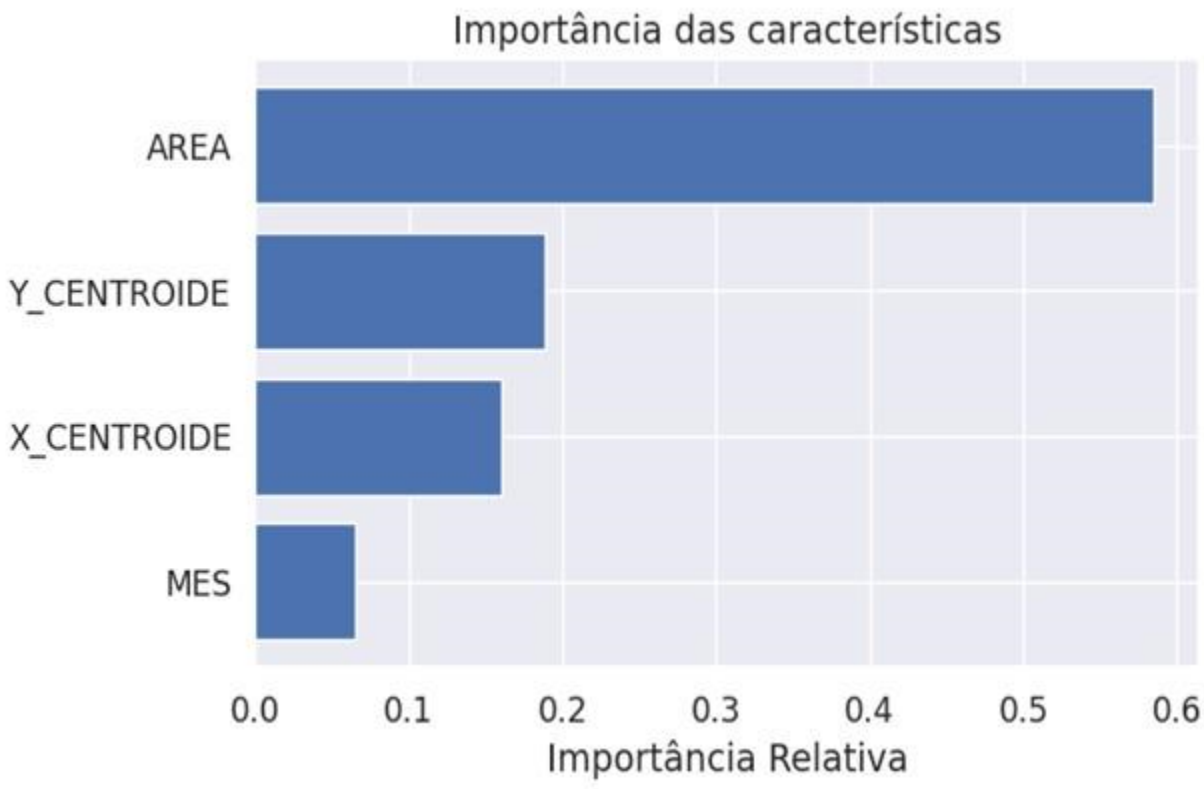
4. RESULTADOS

Os modelos de aprendizado de máquina treinados apresentaram resultados promissores em termos de previsão da quantidade de roubos e furtos de celulares na cidade de São Paulo em conjunto aos dados do GEOSAMPA.

Foram treinados dois modelos de aprendizado de máquina, um modelo de Árvore de Decisão e um modelo de Floresta Aleatória. Ambos os modelos foram utilizados para avaliar a importância das variáveis mês, área, coordenada X e coordenada Y na previsão.

A Gráfico 3 mostra a importância das características atribuídas pelo modelo de Árvore de Decisão. Como se pode observar, todas as quatro variáveis têm alguma contribuição para o modelo, com a “ÁREA” mostrando-se a variável de maior influência.

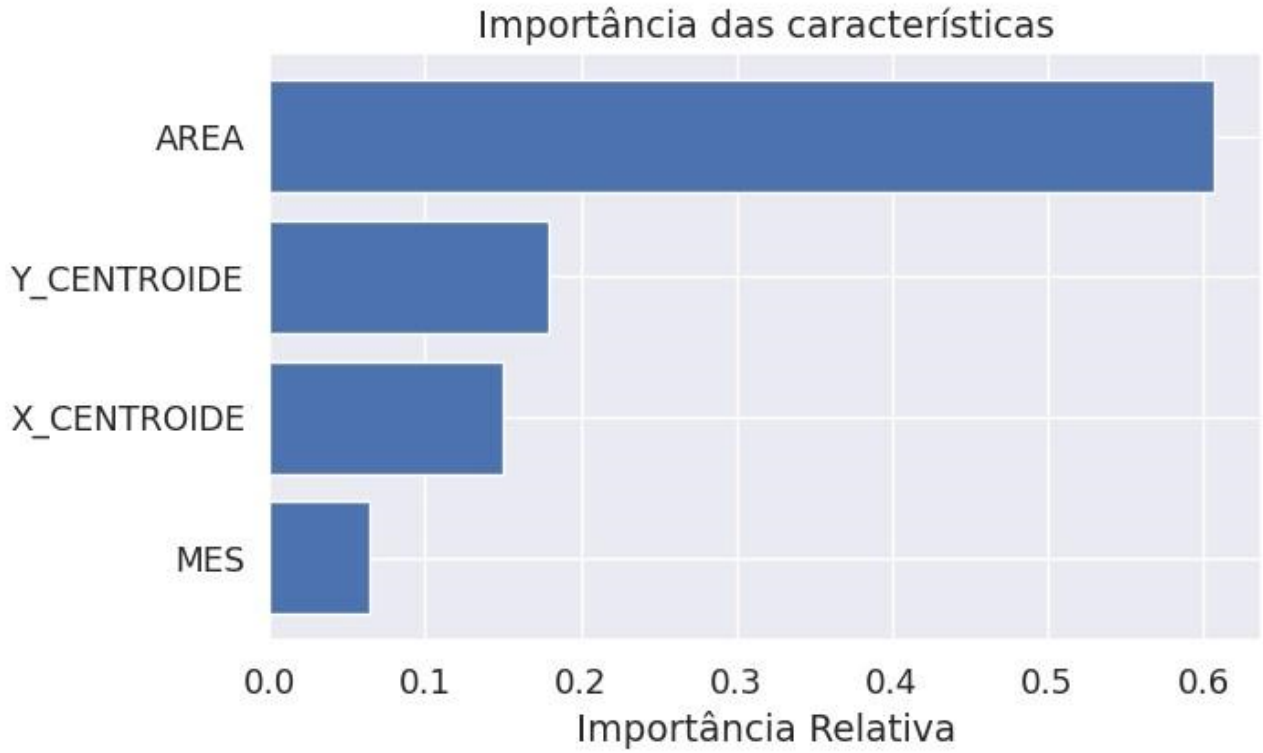
Gráfico 3 – Importância das características de entrada no modelo Árvore de Decisão



Fonte: Elaborado pelos autores (2023)

O Gráfico 4, por outro lado, apresenta a importância das características conforme avaliadas pelo modelo de Floresta Aleatória. A variável área continua sendo a mais significativa.

Gráfico 4 – Importância das características de entrada no modelo Floresta Aleatória



Fonte: Elaborado pelos autores (2023)

Nota-se que a área foi a característica com maior importância em ambos os modelos, indicando uma forte relação entre a área do distrito e a incidência de roubos e furtos de *smartphones*.

O modelo de Árvore de Decisão, com os dados de teste, produziu um coeficiente de determinação (R^2) de 0.921, indicando que aproximadamente 92% da variância no número de roubos pode ser explicada pelo modelo. O Erro Absoluto Médio (MAE) para o mesmo modelo foi de 59.40, sugerindo que, em média, as previsões do modelo diferem do número real de roubos por cerca de 59-60 ocorrências. Com os dados de treinamento, os resultados foram R^2 de 1.0 e um MAE de 0.0.

Tabela 3 – Resultado do modelo Floresta Aleatória

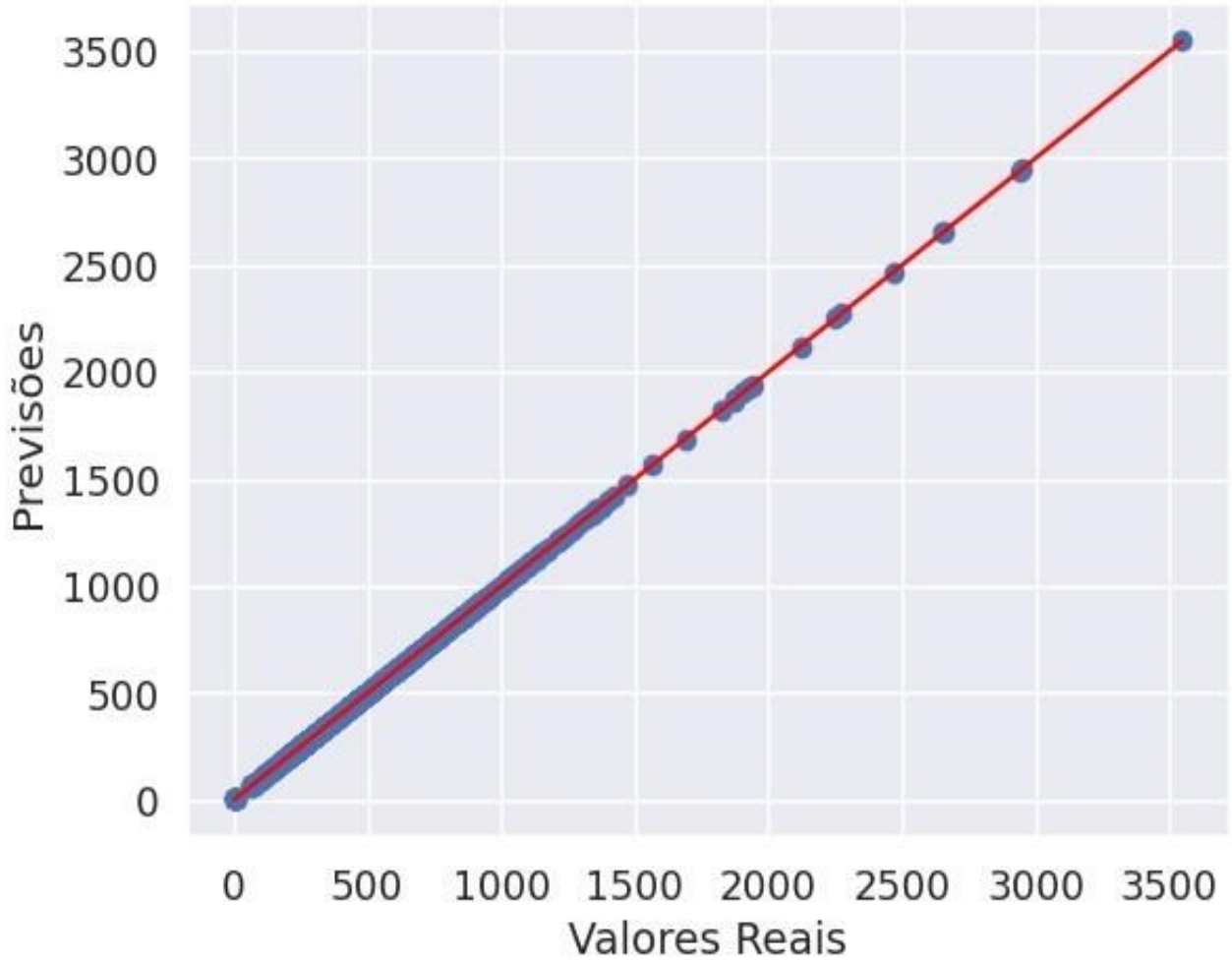
Dados	Coeficiente de Determinação (R^2)	Erro Absoluto Médio (MAE)
Treinamento	1.0	0.0
Teste	0.921	59.40

Fonte: Elaborado pelos autores (2023)

O Gráfico 5, compara os valores previstos pela Árvore de Decisão aos valores reais no conjunto de treino. Cada ponto no gráfico representa uma observação do conjunto de treino, com a eixo Y indicando o valor real e a eixo X mostrando a previsão do modelo. A linha diagonal vermelha representa a correspondência perfeita entre valores reais e previstos.

É notável que todos os pontos estão perfeitamente alinhados com a linha diagonal, isso indica que o modelo foi capaz de prever corretamente todos os valores de treinamento e possui um bom resultado com os dados de teste.

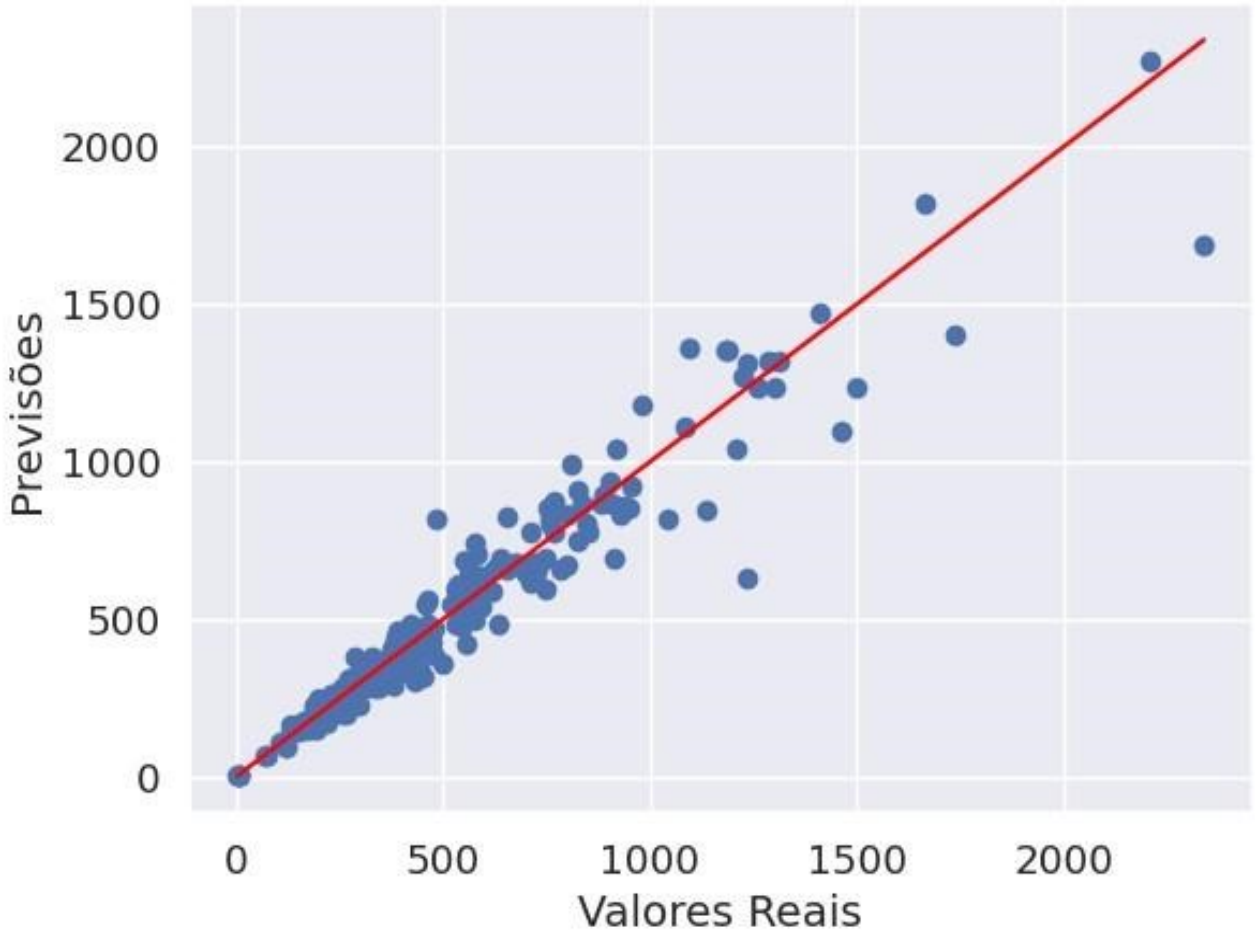
Gráfico 5 – Previsões com os dados de Treino do modelo Árvore de Decisão



Fonte: Elaborado pelos autores (2023)

O Gráfico 6 é similar ao anterior, mas aqui consideramos o conjunto de testes. Assim, os valores previstos pela Árvore de Decisão são comparados aos valores reais no conjunto de testes. A distribuição dos pontos neste gráfico nos fornece uma visão sobre o quão bem o modelo é capaz de generalizar seu aprendizado para novos dados. Mas devido a 100% de precisão com o conjunto de treino, é um indicativo que existe a possibilidade de melhora na precisão com o modelo Floresta Aleatória

Gráfico 6 – Previsões com os dados de Teste do modelo Árvore de Decisão



Fonte: Elaborado pelos autores (2023)

Por outro lado, o modelo Floresta Aleatória demonstrou um desempenho superior com os dados de teste. O R² com os dados de teste foi de 0.951, sugerindo que o modelo é capaz de explicar aproximadamente 95% da variância no número de roubos e furtos. O MAE para este modelo foi de 45.22, o que significa que as previsões do modelo diferem do número real de roubos e furtos por cerca de 45 ocorrências em média. Com os dados de treinamento, o R² foi de 0.989 e o MAE foi de 19.46, indicando uma melhora significativa em relação ao modelo de "Árvore de Decisão".

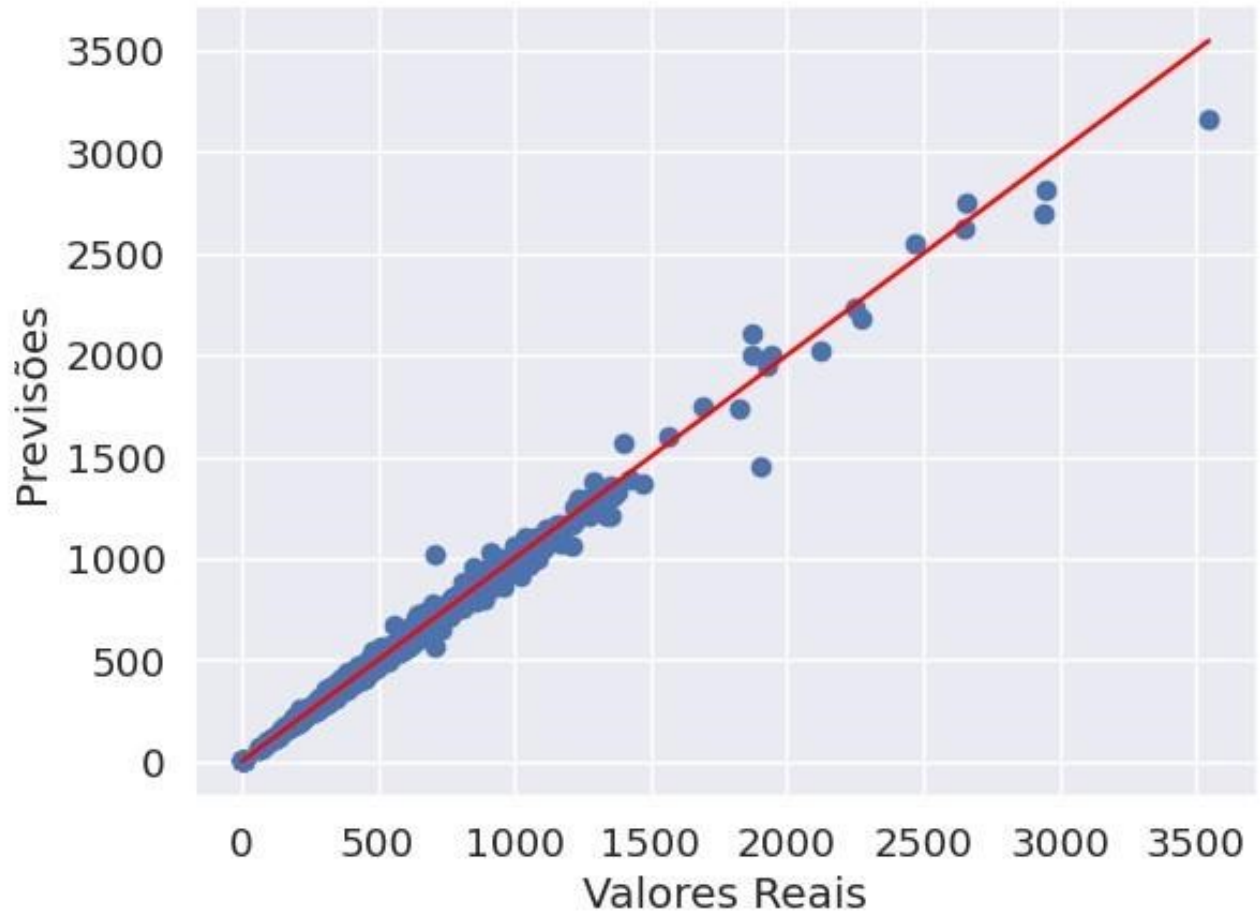
Tabela 4 – Resultado do modelo Floresta Aleatória

Dados	Coeficiente de Determinação (R²)	Erro Absoluto Médio (MAE)
Treinamento	0.989	19.46
Teste	0.951	45.22

Fonte: Elaborado pelos autores (2023)

O Gráfico 7, os valores previstos pelo modelo de *Random Forest* são comparados aos valores reais no conjunto de treino. Similar ao Gráfico 5, cada ponto representa uma observação do conjunto de treino, com a eixo Y indicando o valor real e a eixo X mostrando a previsão do modelo.

Gráfico 7 – Previsões com os dados de Treino do modelo Floresta Aleatória



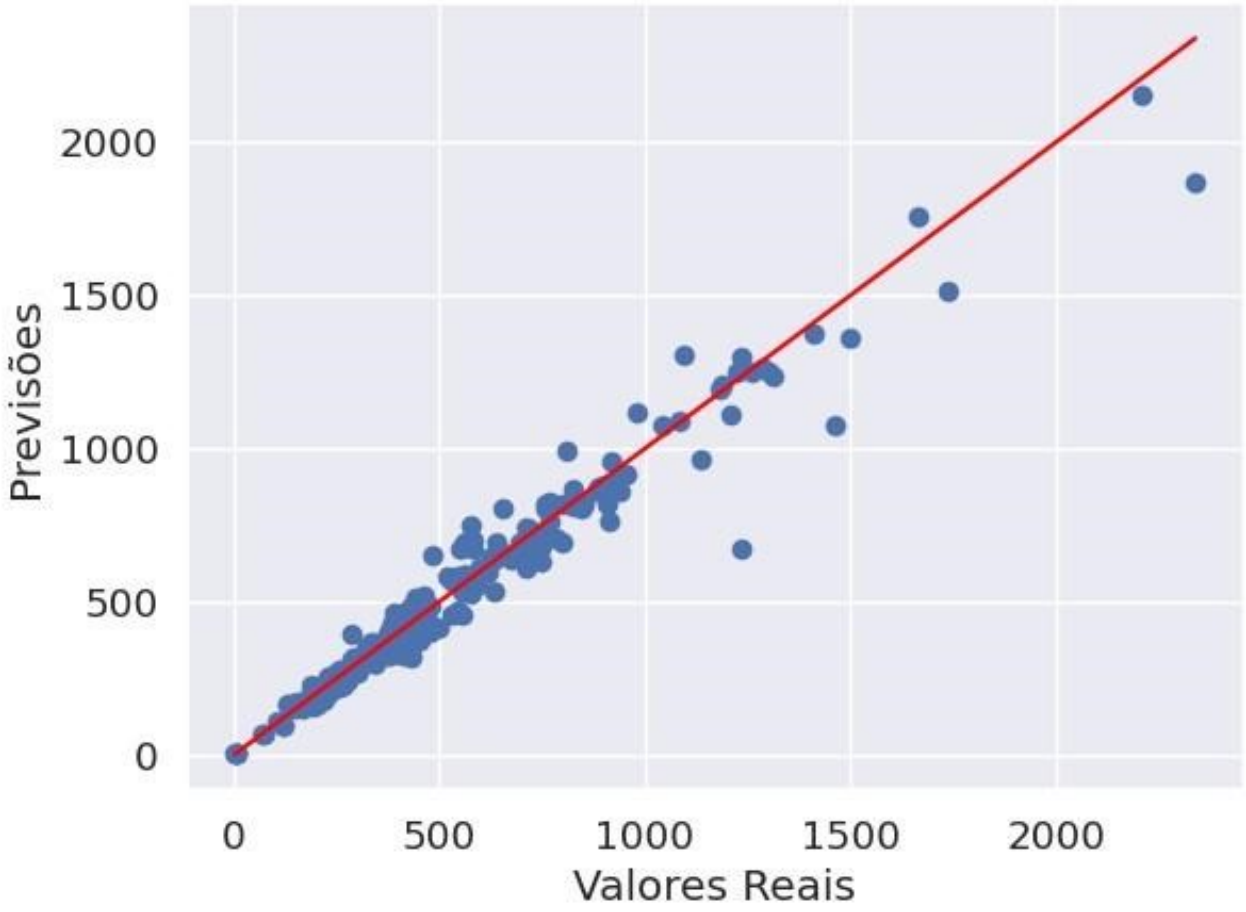
Fonte: Elaborado pelos autores (2023)

O Gráfico 8 é uma continuação do Gráfico 7, mas agora estamos considerando o conjunto de testes. A distribuição dos pontos neste gráfico nos fornece uma ideia de quão bem o modelo de Random Forest é capaz de generalizar seu aprendizado para dados não vistos anteriormente.

A semelhança entre os gráficos de treino e teste sugere que a *Random Forest* conseguiu generalizar eficazmente o aprendizado do conjunto de treino para o conjunto de testes. Este é um resultado positivo, já que um dos principais desafios no aprendizado de máquina é garantir que os modelos sejam capazes de realizar previsões precisas em dados novos e não vistos, não apenas replicar os resultados nos dados em que foram treinados.

A boa performance nos dados de teste sugere que o modelo *Random Forest* provavelmente será capaz de realizar previsões confiáveis em novos dados, desde que esses dados sejam semelhantes aos dados de treino e teste usados aqui. A *Random Forest*, com sua abordagem de conjunto e a capacidade de capturar complexidades não lineares, provou ser um modelo robusto para essa tarefa de regressão.

Gráfico 8 – Previsões com os dados de Teste do modelo Floresta Aleatória



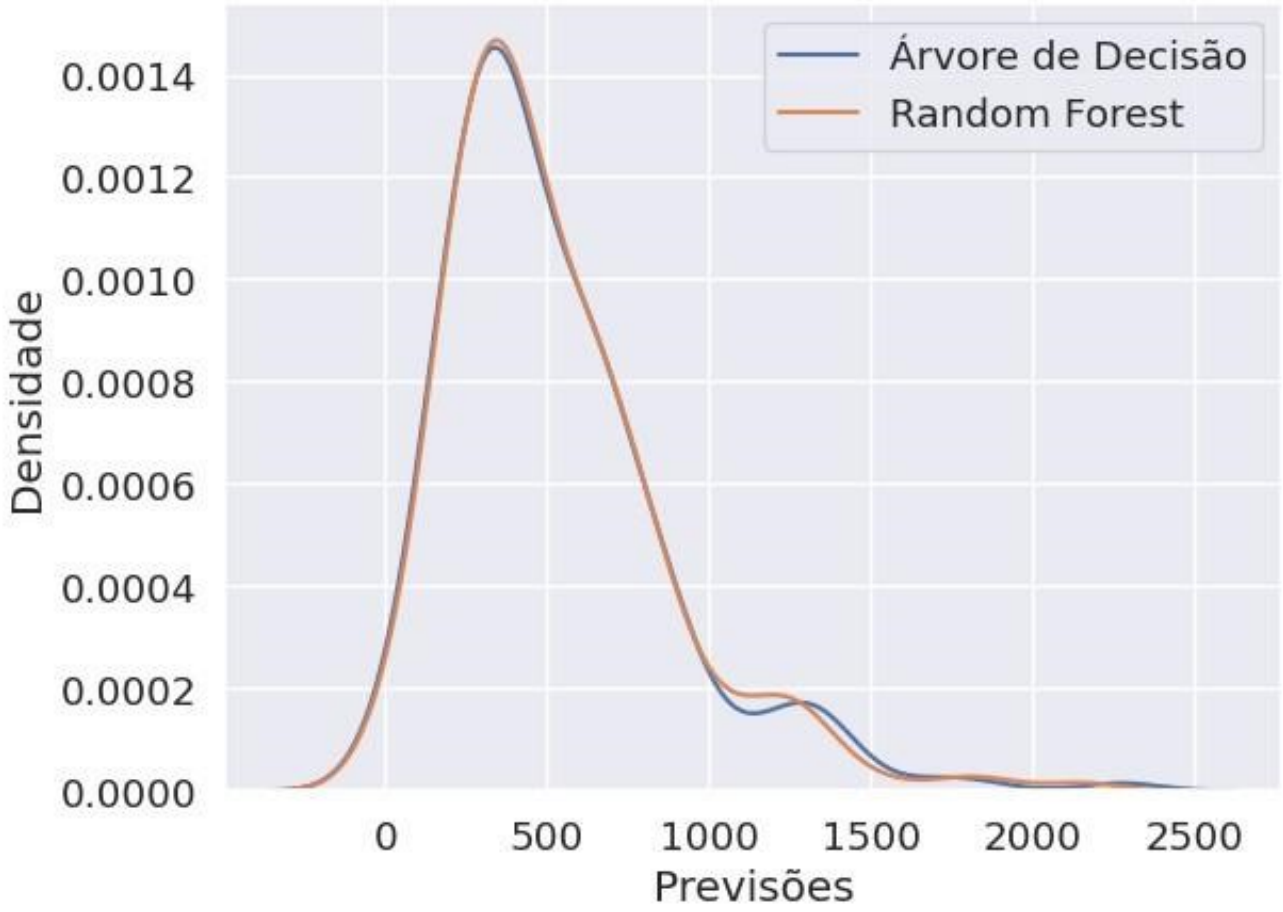
Fonte: Elaborado pelos autores (2023)

Ao comparar os resultados dos modelos com a média de roubos e furtos por distrito e mês (529.71), pode-se ver que ambos os modelos apresentaram previsões com um erro absoluto em torno dos 10%. No entanto, o modelo Floresta Aleatória apresentou uma melhor compreensão dos dados e uma precisão levemente superior, com previsões mais próximas à média real.

Para avaliar e comparar as previsões feitas pelos modelos de Árvore de Decisão e Floresta Aleatória, utilizamos gráficos de densidade de *kernel*. Esses gráficos representam uma estimativa suavizada da distribuição das previsões de cada modelo.

No gráfico 9, o eixo x representa os valores previstos pelos modelos e o eixo y representa a densidade estimada dessas previsões com os dados de testes. As duas linhas no gráfico representam os dois modelos estudados.

Gráfico 9 – Comparação das previsões dos modelos



Fonte: Elaborado pelos autores (2023)

Os resultados deste estudo destacam o potencial dos modelos de aprendizado de máquina para prever a quantidade de roubos e furtos de celulares. O modelo Floresta Aleatória, em particular, demonstrou um pequeno ganho com relação ao modelo de Árvore de Decisão.

5. CONSIDERAÇÕES FINAIS

Este estudo apresentou uma metodologia para a previsão de roubos e furtos de celulares na cidade de São Paulo, empregando técnicas de aprendizado de máquina e análise de dados geográficos. Os resultados obtidos destacam o valor e a eficácia dessas abordagens na prevenção de crimes.

Ambos os modelos de aprendizado de máquina, Árvore de Decisão e Floresta Aleatória, demonstraram um bom desempenho na previsão da quantidade de roubos. No entanto, o modelo *Random Forest* mostrou-se ligeiramente mais preciso comparado com a Árvore de Decisão. Este resultado ressalta o potencial do uso de técnicas mais avançadas de aprendizado de máquina na previsão de crimes.

Além disso, é importante notar que a prevenção e previsão de crimes é um desafio complexo. Enquanto este estudo considerou características geográficas e temporais, existem outros fatores socioeconômicos que podem influenciar a incidência de roubos e furtos de celulares (BERNASCO; BLOCK, 2011). Futuras pesquisas poderiam se beneficiar da inclusão desses fatores para aprimorar ainda mais a precisão e aplicabilidade dos modelos de previsão.

Em resumo, este estudo apresentou um resultado promissor de como o aprendizado de máquina e a análise de dados podem ser empregados para melhorar a segurança pública. Ao permitir previsões mais precisas de roubos e furtos de celulares, estas abordagens podem ajudar as autoridades a desenvolver estratégias de prevenção de crimes mais eficazes, contribuindo para uma cidade mais segura.

6. REFERÊNCIAS BIBLIOGRÁFICAS

Anatel - **Resolução nº 632, de 7 de março de 2014** Disponível em: <https://informacoes.anatel.gov.br/legislacao/resolucoes/2014/750-resolucao-632# art:4 anexo I>. Acesso em: 19 mai. 2023.

BAEZA-YATES, R. RIBEIRO-NETO, B. **Recuperação de Informação: Conceitos e tecnologia das máquinas de busca.** Tradução técnica: Leandro Krug Wives, Viviane Pereira Moreira. 2. ed. Porto Alegre: Bookman, 2013.

Bernasco, W., & Block, R. (2011). **Robberies in Chicago: A block-level analysis of the influence of crime generators, crime attractors, and offender anchor points.** Journal of Research in Crime and Delinquency, 48(1), 33-57.

BORGES, H. B. **Classificador Hierárquico Multirrótulo Usando uma Rede Neural Competitiva**. Disponível em: https://www.ppgia.pucpr.br/pt/arquivos/doutorado/teses/2012/helyane_borges.pdf. Acesso em: 14 mai. 2023.

BREIMAN, L., "**Random Forests Leo Breiman and Adele Cutler**". Disponível em: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#workings. Acesso em: 19 Mai. 2023.

CHEESEMAN, P. & J. STUTZ. **Bayesian Classification (Auto Class): Theory and Results.** In: FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. (Eds.), Advances in Knowledge Discovery and Data Mining. American Association for Artificial Intelligence. Menlo Park, CA. 1996, p. 153-180

Didática Tech. **Overfitting e Underfitting**. Disponível em: <https://didatica.tech/underfitting-e-overfitting/>. Acesso em: 09 mai 2023.

GeoSampa. **Mapa Digital da Cidade de São Paulo** – Disponível em: <https://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx#>. Acesso em 05/04/2023.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. **Data mining: conceitos, técnicas, algoritmos, orientações e aplicações**. Rio de Janeiro: Elsevier, 2015.

GONÇALVES, M. Classificação de Textos. In: BAEZA-YATES, R. RIBEIRO-NETO, B. **Recuperação de Informação: Conceitos e tecnologia das máquinas de busca**. Tradução técnica: Leandro Krug Wives, Viviane Pereira Moreira. 2. ed. Porto Alegre: Bookman, 2013. p. 277-338.

JACOB, Elin K. **Classification and categorization: a difference that makes a difference.** 2004. Disponível em: <https://www.ideals.illinois.edu/handle/2142/1686>. Acesso em: 23 mai. 2023.

Jornal da Band. **SP: Polícia descobre comércio clandestino de celulares.** Disponível em: <https://videos.band.uol.com.br/16770853/sp-policia-descobre-comercioclandestinode-celulares.html>. Acesso em: 13 mai. 2023.

Jornal Nacional - **JN flagra ação de ladrões em viaduto de São Paulo - 24/10/2016.** Disponível em: <https://globoplay.globo.com/v/5400197/.> Acesso em 13/05/2023.

LOPES, Gesiel Rios; DELBEM, Alexandre CB; DE SOUSA, Joélcio Braga. **Introdução à análise de dados geoespaciais com python**. Sociedade Brasileira de Computação, 2021.

MARKOV, Z.; LAROSE, D. T. **Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage. New Britain**: Wiley, 2007.

MATOS, D. **Conceitos Fundamentais de Machine Learning.** Disponível em: <http://www.cienciaedados.com/conceitos-fundamentais-de-machine-learning/>. Acesso em 23 mai. 2023

Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (1994). **Machine Learning, Neural, and Statistical Classification**. Prentice Hall.

MITCHELL, T. **Machine Learning. New York, NY: McGraw-Hill, 1997**. ISBN 0-07042807-7.

MONARD, M. C.; BARANAUSKAS, J. A. **Sistemas Inteligentes: Fundamentos e Aplicações, capítulo Conceitos sobre Aprendizado de Máquina**, pp. 89–114. Manole, 2003.

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). **Introduction to time series analysis and forecasting**. John Wiley & Sons.

MÜLLER, A. C. e GUIDO, S. **Introduction to Machine Learning with Python**. O'Reilly Media, 2017.

RUSSELL, Stuart; NORVIG, Peter. **The history of artificial intelligence. Artificial Intelligence: A Modern Approach**. 3rd. Upper Saddle River, NJ, USA: Prentice Hall Press, p. 16-28, 2009.

Secretaria da Segurança Pública de São Paulo. **Transparência** – Disponível em: <http://www.ssp.sp.gov.br/transparenciassp/Consulta2022.aspx>. Acesso em 26/03/2023.

[SP1 2017] SPTV. SP1 - **SP1 flagra venda de celulares roubados em feiras - 30/08/2017.** Disponível em: <https://globoplay.globo.com/v/6114067/>. Acesso em 13/05/2023.

[SP2 2020] SPTV- **Quadrilha que roubava celulares em São Paulo é presa - 04/11/2020.** Disponível em: <https://globoplay.globo.com/v/8996397/>. Acesso em 13/05/2023.

TELOKEN, Alex et al. **Estudo Comparativo entre os algoritmos de Mineração de Dados Random Forest e J48 na tomada de decisão. Simpósio de Pesquisa e Desenvolvimento em Computação**, v. 2, n. 1, 2016.

Ueda, J. K. " **Previsão de Pontos Críticos de Crime: Uma modelagem por aprendizado de máquina para prevenção de roubos de celulares em São Paulo**." Dissertação (Bacharelado) – Universidade de São Paulo [UOL 2022] - **Roubos e furtos de celulares levam a 'corrida' para bloquear chaves Pix.** Disponível em: <https://economia.uol.com.br/noticias/estadaoconteudo/2022/04/09/roubos-furtos-celulares-bloqueio-chaves-pix.ndhtm>. Acesso em 21/05/2023.