UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS DE SOBRAL
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA
ELÉTRICA E DE COMPUTAÇÃO (PPGEEC)

ANDRIO RODRIGO CORRÊA DA SILVA

Predição de localização de crimes em região urbana usando algoritmos de análise de regressão

SOBRAL
2020

ANDRIO RODRIGO CORRÊA DA SILVA

Predição de localização de crimes em região urbana usando algoritmos de análise de regressão

Dissertação apresentada ao Programa de Pós-graduação em Engenharia Elétrica e de Computação (PPGEEC) do Campus de Sobral da Universidade Federal do Ceará, como parte dos requisitos necessários para a obtenção do título de Mestre em Engenharia Elétrica e da Computação. Área de concentração: Sistemas de Informação.

Orientador: Prof. Dr. Iális Cavalcante da Paula Junior

SOBRAL

2020

AGRADECIMENTOS

Agradeço à minha família por ter me apoiado durante essa jornada de graduação e pós-graduação. Agradeço principalmente à minha mãe por estar sempre ao meu lado me aconselhando e me fazendo continuar.

Agradeço a meu orientador e mentor, Prof. Dr. Iális Cavalcante, por ter me dado a oportunidade de ser seu orientando, por ter sido paciente em vários momentos e por ter sempre me ajudado nas dificuldades. Além de ser um brilhante profissional é também um grande ser humano.

Agradeço também aos colegas de mestrado pelo conhecimento compartilhado durante esse período.

RESUMO

Há índices relevantes de violência no Brasil que tem aumentado nos últimos anos. Isso exige que o combate à violência seja inteligente e eficiente, a fim de reduzir o gasto público e tempo dos agentes públicos. Existem várias soluções, este trabalho também apresenta uma delas, que usam sistemas inteligentes para prever onde e quando um crime ocorrerá, isso permite organizar rotas policiais para áreas com maior risco de perigo. Nos experimentos realizados, foram utilizados duas bases de dados, uma da cidade da Filadélfia, Estados Unidos, disponível de forma pública, e outra da cidade de Fortaleza, disponibilizada pela Secretaria de Segurança Pública do Estado do Ceará. Para esses experimentos, diferentes métodos de regressão foram aplicados para a realização das predições dos locais onde os crimes poderiam ocorrer. Com a base de dados da cidade da Filadélfia, Estados Unidos, foi realizado somente um experimento utilizando esses regressores. Para a base de dados da cidade de Fortaleza, foi realizado quatro conjuntos de experimentos utilizando os mesmos métodos de regressão utilizados anteriormente para a cidade da Filadélfia, Estados Unidos. Para cada uma dessas técnicas, foi gerado um resultado referente aos resíduos, sendo a diferença entre os valores preditos e os valores reais. A utilização dos métodos regressores também proporcionou a geração de gráficos de dispersão de pontos, onde cada predição realizada é plotada e comparada aos pontos originais. Houve também a necessidade de exibir os pontos preditos nos mapas das respectivas cidades, dessa forma é possível verificar as áreas em que há maiores incidências criminais. Para cada um dos métodos regressores foi calculado o valor de erro utilizando as métricas MSE (*Mean Squared Error*) e RMSE (*Root Mean Squared Error*), os menores valores para MSE e RMSE permitem inferir que o modelo apresentou ótimas predições. Os resultados obtidos pelos métodos de regressão se mostraram eficientes na tarefa de predição de localização de crimes. É possível concluir que a utilização desses métodos para problemas de aspectos criminais, tornam a tarefa preditiva muito mais tangível.

Palavras-chave: Aprendizagem de máquina, Análise de Regressão, Predição de crimes.

ABSTRACT

There are relevant rates of violence in Brazil that have increased in recent years. Intelligence and efficiency are required to combat this issue, in order to reduce public money spending and time from public officials. There are several solutions, this work also presents one of them, which use intelligent systems to predict where and when a crime will occur, this allows organizing police routes to areas with a higher risk of danger. In the experiments carried out, two databases were used, one from the city of Philadelphia, publicly available, and another from the city of Fortaleza, provided by the Department of Public Safety. For these experiments, different regression methods were applied to make predictions of the places where crimes could occur. For the dataset of the city of Philadelphia, only one experiment was performed using these regressors. For the dataset of the city of Fortaleza, four sets of experiments were carried out using the same regression methods used previously for the city of Philadelphia. For each of these techniques, a result was generated regarding the residues, which is the difference between the predicted values and the actual values. The use of regression methods also provided the generation of point dispersion plots, where each prediction made is plotted and compared to the original points. There was also a need to display the predicted points on the maps of the respective cities, so, this way, it is possible to check the areas where there are major criminal incidents. For each of the regressor methods, the error value was calculated using the metrics MSE (Mean Squared Error) and RMSE (Root Mean Squared Error), the lowest values for MSE and RMSE allow to infer that the model presented excellent predictions. The results obtained by the regression methods proved to be efficient in the task of predicting the location of crimes. It is possible to conclude that the use of these methods for problems of criminal aspects, make the predictive task much more tangible.

Keywords: Machine learning, regression analysis, crime prediction.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Nos últimos anos, a taxa de violência no Brasil apresentou crescimento em índices relevantes, principalmente no que diz respeito aos números relacionados a homicídios. No relatório do IPEA (Cerqueira *et al.*, 2019) é relatado uma diminuição de homicídios nas regiões Sudeste e Centro-Oeste, no entanto, houve um aumento nas regiões Norte e Nordeste.

Dados oficiais do Ministério da Saúde (Cerqueira *et al.*, 2019) apontam que em 2017 houve 65.602 homicídios registrados no território brasileiro, um número alto quando comparado ao período de 2007, onde a taxa de homicídios foi próxima a 50.000. O relatório do IPEA também aponta o Ceará como o estado em que houve o maior crescimento de homicídios em 2017, atingindo um número de 5.433 mortes causadas por armas de fogo, drogas ilícitas e conflitos interpessoais.

O combate à violência precisa ser intensificado, sendo necessário alocar recursos para um combate eficaz e inteligente. Na cidade de Los Angeles, Estados Unidos, é usado um sistema chamado PredPol (Capellán and Otero, 2017). Este sistema propõe fazer uma predição de um crime, mais especificamente, o tipo, a localização, a data e a hora do mesmo. Este projeto nasceu de uma parceria entre o Departamento de Polícia de Los Angeles e a UCLA (Universidade da Califórnia). O uso desses sistemas inteligentes permite alocar recursos de patrulhamento mais efetivos, contribuindo mais para a segurança da população.

Realizar a predição de um crime não é tarefa fácil, apenas uma grande quantidade de dados não é suficiente, visto que encontrar padrões nos crimes é um grande obstáculo a ser resolvido (Sathyadevan and Gangadhara, 2014), além do fator sazonalidade que está intrínseco nos crimes cometidos. No trabalho em (Araújo *et al.*, 2018) é proposto uma estrutura que classifica lugares e horários em graus de perigo. No artigo os autores descrevem que o *framework* é proposto a plataformas e sistemas de segurança. Os autores também apontam que o *framework* gera *hotspots* para cada janela de tempo, que no caso é semanal o que implica em novas áreas de crimes que são detectadas ao longo do tempo, evitando assim patrulhas sempre nos mesmos lugares. Esse artigo utiliza dados de um estado do nordeste do Brasil, o Rio Grande do Norte.

Algumas pesquisas, como (Ingilevich and Ivanov, 2018), propõem uma solução de predição de crimes baseadas em fatores relacionados ao ambiente. Os autores afirmam que a urbanização cria vários problemas sociais, inclusive o crime. No estudo publicado, os autores identificaram indicadores espaciais que influenciam em determinados tipos de crimes. Esses indicadores são: tamanho da população, estações de polícia, escolas, *shoppings*, igrejas, conveniências, número de prédios e bares. Baseados nessa hipótese, os autores conseguiram realizar a predição do número de crimes em determinadas áreas da cidade de São Petersburgo, Rússia.

Alguns fatores externos estão diretamente ligados a incidências de crimes. No artigo (Xu *et al.*, 2017), os autores mostram que a iluminação das ruas é uma variável que contribui para a redução ou aumento de crime em uma determinada área. Ou seja, em uma rua escura é mais provável acontecer roubos e invasões, diferentemente do que ocorreria em uma rua com maior nível de iluminação. O estudo foi realizado utilizando dados da cidade de Detroit, localizada nos Estados Unidos. Há outros estudos que tentam identificar a relação entre temperatura e o aumento de crimes, no entanto ainda não há um consenso de que o aumento ou diminuição da temperatura global está diretamente ligado as altas taxas de crimes.

As redes sociais se tornaram nos últimos anos uma grande fonte de dados. Foi pensando nessa quantidade de dados que os autores do estudo (Abbass *et al.*, 2020) criaram um *framework* que permite realizar a predição de crimes através de *tweets* da rede social *Twitter*. No trabalho publicado, os autores coletam dados relacionados a diferentes categorias de palavras chaves, como perseguição, assédio, *bullying* etc. Utilizando esse *framework* é possível monitorar e prevenir comportamentos criminais tanto de indivíduos como também de grupos.

Alguns estudos, como (Dash, Safro, and Sakrepatna, 2018), consideram fatores sociais, como educação e condições econômicas, relevantes para realizar a predição de crimes, os autores afirmam que essa perspectiva contribui para melhorar a qualidade da predição. Os autores também utilizam informações sobre os diferentes tipos de comunidades dentro de uma cidade, tentando dessa forma definir padrões de crimes em determinadas áreas. Os experimentos do trabalho foram realizados utilizando a base de dados da cidade de Chigago, Estados Unidos.

No Brasil já há algumas iniciativas sobre o uso de dados para realizar a predição e o monitoramento de crimes. O governo do estado de São Paulo utiliza o sistema Detecta (SSP, 2020) que permite auxiliar os policiais no trabalho investigativo, realizando o cruzamento de informações sobre determinado crime ou suspeito. No Rio de Janeiro foi criado o sistema CrimeRadar pelo Instituto Igarapé (Igarapé, 2020). Esse sistema gera uma mapa de calor em que é possível visualizar as áreas com maiores incidências criminais.

Há uma similaridade entre os estudos apresentados acima, e o termo *machine learning* ou aprendizado de máquina. O aprendizado de máquina permite que computadores tomem decisões sem necessariamente haver uma intervenção humana (Kim *et al.*, 2018). Esse campo da computação é bastante amplo, há algoritmos de aprendizagem supervisionada e não supervisionada. Nos problemas de aprendizagem supervisionada, os dados são rotulados, ou seja, para os valores de entrada, já temos os valores de saída conhecidos. Dessa forma, precisaremos apenas encontrar uma função que os relacione. Já para os problemas de aprendizagem não supervisionada, não temos os dados rotulados, ou seja, será necessário analisar a semelhança entre os dados para realizar um eventual

agrupamento, mais conhecido como *clustering*.

A classe de problemas de aprendizagem supervisionada pode ser dividida em dois tópicos: classificação e regressão. Nos problemas de classificação, é necessário identificar a qual classe aquele conjunto de dados pertence. Um exemplo disso, é o problema da classificação de flores, o conhecido *Iris dataset* (Pinto, Kelur, and Shetty, 2018), em que há as dimensões da sépala e da pétala de uma determinada flor, e o objetivo é determinar, baseado nessas dimensões, se aquela flor é uma setosa, versicolor ou virgínica. Já nos problemas de regressão, a saída sempre será um número. Um exemplo de regressão pode ser o problema de predição de preço de casas (Madhuri, G, and Pujitha, 2019), dado um conjunto de dados que representa uma casa, como número de quartos, número de banheiros, área da casa etc, a saída esperada deve ser o preço daquele imóvel.

Neste trabalho, propõe-se uma arquitetura para realizar a predição do local onde um crime ocorrerá, mais especificamente, onde acontecerá o crime de roubo a pessoa. O problema será tratado como regressão. Inicialmente foi utilizado, como uma espécie de laboratórios de testes, a base de dados pública de crimes da cidade da Filadélfia, Estados Unidos. Posteriormente foi utilizado a base de dados de crimes da cidade de Fortaleza. Realizou-se dois tipos de experimentos, sendo que cada um possui duas categorias de testes, totalizando quatro cenários de resultados.

1.1 TRABALHOS RELACIONADOS

Em (Capellán and Otero, 2017) é introduzido um sistema chamado PredPol que é utilizado na cidade de Los Angeles, Estados Unidos. Esse sistema tem como principal objetivo realizar uma predição, não somente do tipo de crime, mas também onde, a data e a hora que aquele crime irá ocorrer. Esse sistema tem uma grande contribuição da Universidade da Califórnia. Os autores, apesar de nem um outro estudo independente confirmar, afirma que o algoritmo desenvolvido por eles possui duas vezes mais acurácia quando comparado ao ser humano em se tratando de predizer onde um crime irá ocorrer.

No trabalho apresentado por (Sathyadevan and Gangadhara, 2014), é descrito um sistema capaz de predizer regiões que possuem maior probabilidade de ocorrer um crime. Os autores utilizaram algoritmo de Naive Bayes para realizar o treinamento com dados relacionados a vandalismo, assassinato, roubo, abuso sexual etc. Os testes realizados mostraram que a utilização de Naive Bayes apresentou 90/

No artigo de (Araújo *et al.*, 2018) é introduzido um *framework* que ajuda a melhorar o planejamento de patrulha já que o mesmo pode fornecer precsiamente lugares e horários que são mais propícios de serem perigosos. O textitframework oferece a possibilidade de gerar novas predições a cada novo intervalo de tempo, dessa forma novos lugares são preditos.

O trabalho descrito em (Ingilevich and Ivanov, 2018) apresenta características do ambiente que podem contribuir para o aumento de crimes. Os autores discutem que

tamanho da população, estações de polícia e até escolas podem estar diretamente relacionadas com o aumento ou diminuição do número de crimes em uma determinada área, bairro etc. Os autores obtiveram uma melhor acurácia no trabalho utilizando modelos de *boosting*, tais resultados podem ajudar a polícia a encontrar os melhores lugares para colocar uma estação policial e assim reduzir a taxa de criminalidade.

Alguns estudos, como (Xu *et al.*, 2017), citam que fatores externos também interferem na incidências de crimes. Os autores mostram que a iluminação nas ruas contribui para um aumento ou redução de crime em determinada localidade. Em uma rua escura, por exemplo, há uma maior chance de ocorrer um crime do que em uma rua com uma luminosidade maior. O trabalho foi realizado na cidade de Detroit.

1.2 OBJETIVO

1.2.1 OBJETIVO GERAL

Analisar dados criminais e realizar a predição de onde ocorrerá o crime de roubo a pessoa relacionado a cidade de Fortaleza, utilizando técnicas de regressão.

1.2.2 OBJETIVOS ESPECÍFICOS

- Analisar dados de segurança pública para associação entre ocorrências criminais e informações de tempo e localização advindas das bases de dados exploradas.
- Avaliar diferentes algoritmos de regressão para aplicação em bases de dados de crimes.
- Realizar um comparativo de metodologias para predição de localização de crimes.
- Construir uma sequência eficiente de visualização de dados para identificação de localização de crimes.

1.3 ORGANIZAÇÃO DO TRABALHO

Este trabalho está organizado da seguinte forma: na Capítulo 2, os materiais e métodos, como conjuntos de dados, tanto da cidade da Filadélfia, Estados Unidos, quanto de Fortaleza, e técnicas de regressão, são descritos. No Capítulo 3, são mostrados os procedimentos que foram seguidos para a metodologia proposta a realização das predições e consequentemente para a obtenção dos resultados. Apresenta-se no Capítulo 4 o que foi gerado a partir dos resultados alcançados, os resultados obtidos são apresentados em tabelas e também de forma gráfica para um melhor entendimento para o público geral. Por fim, traz-se a conclusão do trabalho, além de perspectivas futuras em relação ao problema explorado.

2 MATERIAIS E MÉTODOS

Neste capítulo, os dados, as ferramentas e procedimentos utilizados para a composição do presente trabalho serão explanadas. Na Seção 2.1, será apresentada a base de dados utilizada para a cidade da Filadélfia, Estados Unidos, e para a cidade de Fortaleza. Na Seção 2.2, será apresentados os métodos regressores usados para o treinamento dos dados.

2.1 BASE DE DADOS

A base de dados da cidade da Filadélfia, Estados Unidos, está disponibilizada de forma pública no site *Kaggle* (Kaggle, 2020), que é especializado em competições de aprendizado de máquina. Essa base de dados é composta por 2.237.605 linhas e 14 colunas, distribuídos entre os anos de 2005 e 2017.

A base de dados utilizada para a cidade de Fortaleza foi fornecida pela Secretaria de Segurança Pública do Estado do Ceará. Por motivos de segurança e confidencialidade, os dados não podem ser disponibilizados publicamente neste momento.

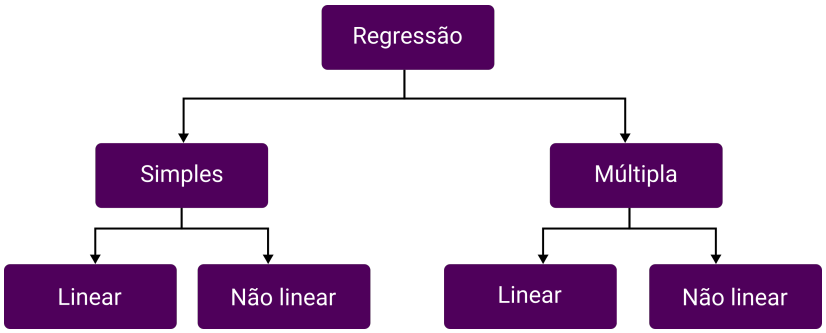
O conjunto de dados disponível possui 122.005 instâncias e 11 atributos relacionados a um total de 25 diferentes tipos de crimes ocorridos na cidade de Fortaleza, que foram coletados entre 2015 e 2019. Novos atributos foram criados e outros removidos do conjunto de dados. O processo de criação de novos atributos é detalhado na seção de metodologia.

2.2 REGRESSÃO

O problema, de predizer onde ocorrerá o crime de roubo a pessoa, explorado neste trabalho é tratado como regressão, uma técnica estatística que tenta definir a relação entre variáveis através de um modelo matemático. A regressão é comumente utilizada em problemas em que é necessário definir a causa e efeito entre as variáveis de entrada e as variáveis de saída. A diferença básica entre problemas de classificação e de regressão está no fato de que na primeira o resultado é uma classe, uma categoria, já nos problemas de regressão, o resultado final é um valor numérico, como latitude ou longitude, por exemplo (Jaiswal and Samikannu, 2017).

Dentro da regressão há a divisão entre regressão simples e regressão múltipla, e cada uma delas pode ser divida em linear e não linear. Todas elas pertencem a uma sub-categoria do aprendizado de máquina, conhecido como aprendizagem supervisionada, onde os dados de saídas já são previamente rotulados, ou seja, já são conhecidos. A Figura 1 exibe os tipos de regressão existentes.

Figura 1: Regressão



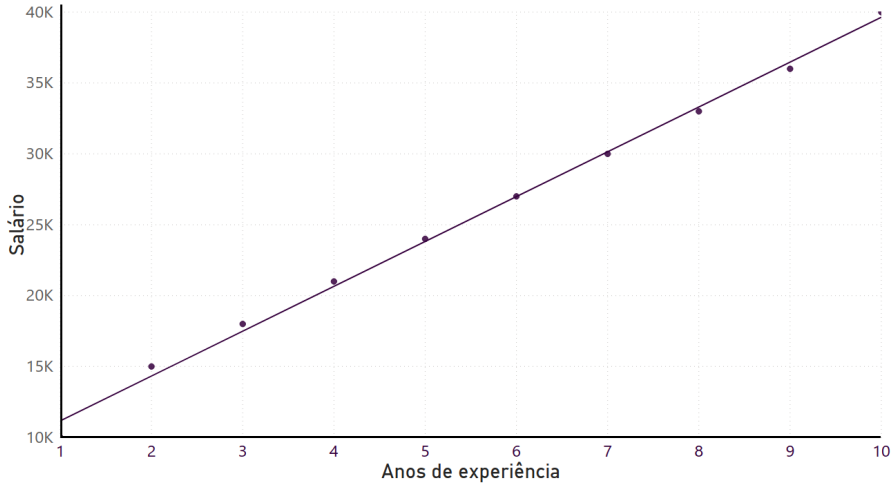
Fonte: Autoria própria.

Nos problemas de regressão simples haverá somente uma variável de entrada relacionada com uma variável de saída. Já nos problemas de regressão múltipla poderá haver diversas variáveis de entrada relacionadas com uma ou mais variáveis de saída.

2.2.1 REGRESSÃO SIMPLES - LINEAR

Nos problemas de regressão simples e linear há apenas uma variável de entrada, também conhecida como variável independente, relacionada com apenas uma variável de saída, também conhecida como variável dependente. A Figura 2 apresenta um exemplo simples de regressão linear.

Figura 2: Regressão simples - linear



Fonte: Autoria própria.

A Figura 2 exibe um exemplo de regressão linear simples, com a relação entre os anos de experiência e o salário de determinada pessoa. É possível notar que a medida que os anos de experiência crescem, o salário também acompanha esse crescimento, portanto, há uma causa e efeito, nesse caso linear.

$$Y = a + bX \tag{1}$$

A Equação 1 exibe, matematicamente, a regressão simples e linear, onde Y representa a variável dependente, o valor que será predito. Enquanto que X, representa a variável independente, a variável preditora. O valor de a indica o valor do ponto Y quando $X = 0$ e b representa a inclinação da reta.

2.2.2 REGRESSÃO SIMPLES - NÃO LINEAR

Nesse tipo de problema a dificuldade persiste no fato de que não há como calcular a relação entre X e Y, visto que elas não são linearmente relacionadas (Kumar, 2015). A regressão simples e não linear também é conhecida como regressão polinomial. Há vários tipos de regressão simples e não linear, as mais conhecidas são: cubica, quadrática, exponencial e logarítmica, sendo essa última definida formalmente na Equação 2.

$$Y = log(X) \tag{2}$$

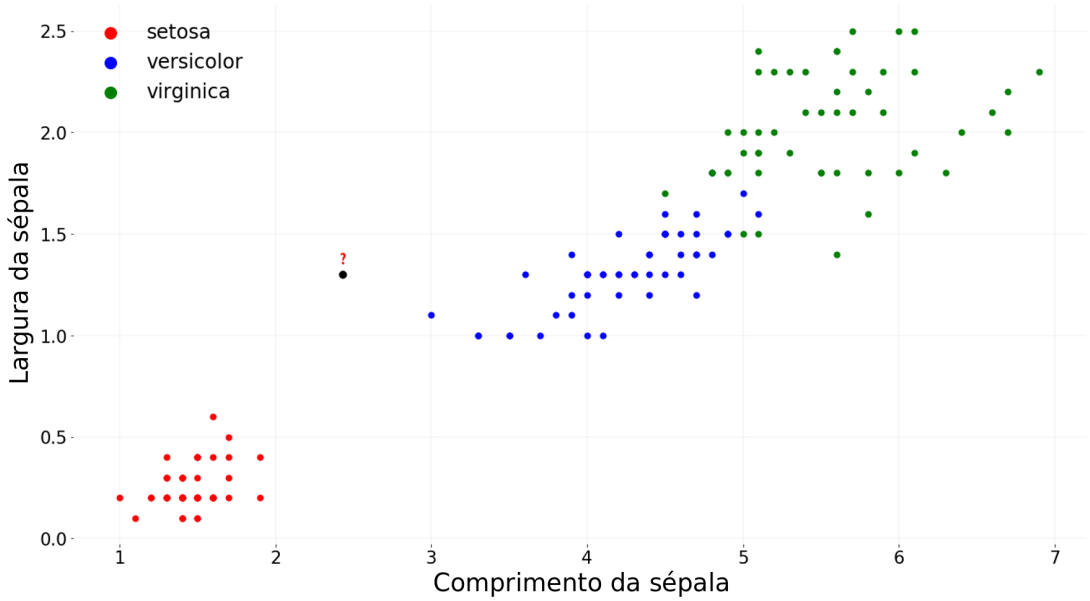
2.2.3 REGRESSÃO MÚLTIPLA - LINEAR E NÃO LINEAR

Na regressão múltipla e linear o valor de saída, Y, está relacionado com múltiplos valores de entrada, X, sendo X um vetor de dados. Já na regressão múltipla e não linear há um variável dependente, chamada de Y, que não está relacionada linearmente com as variáveis independentes, chamadas de X. O problema proposto nesse trabalho está categorizado como regressão múltipla e não linear, com duas variáveis de saída e múltiplas variáveis de entrada.

2.2.4 REGRESSOR K-NEAREST KNEIGHBOR

O algoritmo de *K-Nearest Kneighbor* (KNN) é comumente utilizado em problemas de classificação, entretanto, ele também pode ser utilizado em problemas de regressão. A principal característica do KNN está em conseguir realizar a predição de novos valores baseado na similaridade entre as variáveis. No artigo (Ortiz-Bejar *et al.*, 2018) o autor cita que quando KNN é utilizado em problemas de regressão, a saída y é calculada como o valor do vizinho mais similar, sendo esse valor pertencente ao conjunto original dos dados. A Figura 3 exibe um exemplo simples de como funciona o KNN, utilizando o exemplo da planta Iris. Essa base de dados contém informações de largura e comprimento da sépala e pétala de três flores, a Iris Versicolor, Iris Setosa e Iris Virginica. A largura e comprimento da sépala e da pétala identifica o tipo de flor.

Figura 3: Exemplo do K-Nearest Kneighbor Regressor



Fonte: Adaptado de (Vidhya, 2020) .

Na Figura 3 é plotado os dados de comprimento, no eixo x, e de largura, no eixo y da sépala. E se fosse necessário, por exemplo, descobrir o valor de largura da pétala? No gráfico é possível notar um ponto preto com um sinal de interrogação vermelho, esse é um ponto que ainda não possui rótulo. Para descobrir a qual classe esse ponto pertence é necessário identificar os valores relacionados a pétala (largura e comprimento). Para realizar a predição desses valores alguns passos precisam ser tomados, são eles:

- Calcular a distância entre o ponto a se rotular e cada um dos outros pontos.
- Os pontos que estão mais próximos, dependendo do valor de k estabelecido, são selecionados.
- Caso o valor de k seja maior que 1, será realizado uma soma entre todos os valores que será posteriormente dividido por k. Se k for igual a 1, então o valor da saída, o valor da predição, será *igual* ao valor do vizinho mais próximo.

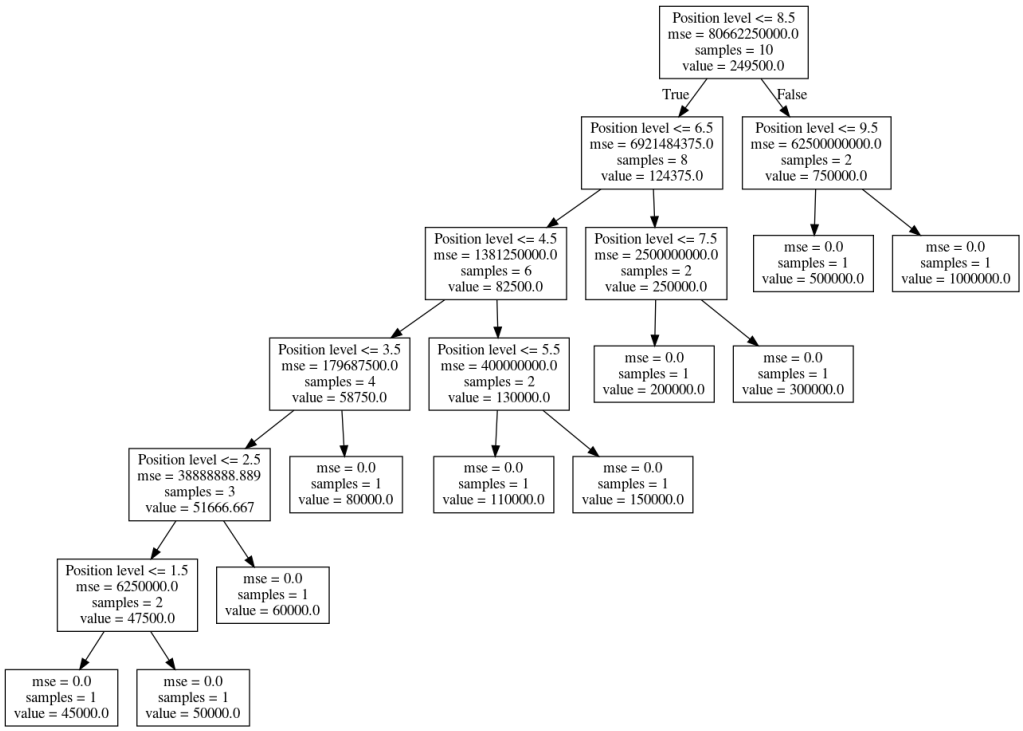
No exemplo da Figura 3, foi utilizado o k sendo igual a 1, portando o valor do comprimento da pétala que o ponto preto assumiu foi o mesmo valor pertencente ao vizinho mais próximo dele, o ponto azul. O mesmo processo pode ser feito para tentar descobrir o valor da sépala. O cálculo da distância entre os pontos pode ser calculado utilizando a medida Euclidiana, mais comumente utilizada, ou pode ser utilizado também a medida de Manhattan.

2.2.5 REGRESSOR DECISION TREE

O método de *Decision Tree* (Árvore de Decisão), funciona baseado em uma série de divisões que são feitas sobre os dados, cada divisão vai definindo os próximos nós da árvore. É um dos algoritmos mais usados para classificação e regressão, possui

poucos parâmetros de configuração, o que facilita e agiliza o seu uso. Outra característica importante dessa técnica é a capacidade de aprender com o ruído dos dados, o que também evita o ajuste excessivo dos parâmetros (Rathan, Sai, and Manikanta, 2019). A Figura 4 apresenta um exemplo simples de *Decision Tree*.

Figura 4: Decision Tree exemplo



Fonte: Entendendo Decision Trees (Dezhic, 2020)

A Figura 4 representa a predição do salário de um funcionário baseado na posição que ele ocupa na empresa, utilizando *Decision Tree*. No topo da árvore é verificado se o nível dele é menor ou igual a 8.5, se a reposta for verdadeira, então deve-se ir para o próximo nível esquerdo da árvore, caso a resposta for falsa, deve-se ir para o nível direito da árvore, isso é feito até chegar ao último nível.

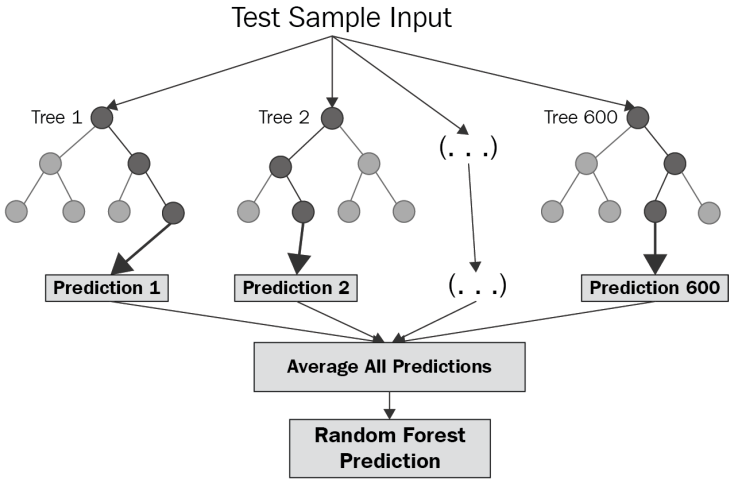
2.2.6 REGRESSOR RANDOM FOREST

No artigo (Breiman, 2001), Leo Breiman introduz o método Random Forest. Na pesquisa, o autor define essa técnica como sendo uma combinação de árvores preditoras. No artigo (Cutler, Cutler, and Stevens, 2011) é citado que o desenvolvimento dessa técnica se deu para competir com os algoritmos tipo *boosting*. Ainda em (Cutler, Cutler, and Stevens, 2011) os autores citam algumas vantagens do Random Forest, como por exemplo, a rapidez para treino e predição, a dependência de poucos parâmetros e a possibilidade de ser utilizado diretamente em problemas de alta dimensionalidade. Random Forest pode ser usado em problemas de classificação, regressão e aprendizado não supervisionado (Bravo and Moreno, 2019).

Em aprendizado de máquina, um dos pontos mais importantes a se evitar é o *overfitting*, ou seja, evitar que o modelo se ajuste de forma ótima aos dados de treino mas que ao ser apresentado a novos dados não consiga realizar uma generalização tão boa quanto no treinamento, passando a realizar predições incertas. A técnica de Random Forest é bastante resistente ao *overfitting*, isso é possível, em maior parte, pelo número de árvores que é gerado na floresta. É importante também destacar que Random Forest é computacionalmente vantajoso até certo ponto, visto que, com um aumento significativo no número de árvores, o processo de predição irá se tornar lento e ineficiente. Para aplicações que requerem predições em tempo real e que utilizam uma vasta gama de dados, essa técnica pode não ser ideal.

No artigo (Cutler, Cutler, and Stevens, 2011) os autores definem que para um vetor aleatório de dimensão p, $X = (X_1, ...X_p)^T$, e um valor aleatório Y, representando um atributo de saída, o objetivo é determinar uma função $f(x)$ que possar relacionar X e Y, proporcionando uma predição. Os autores também citam que a função $f(x)$ é determinada pela minimização de uma função de perda. A função de perda ajuda a avaliar os dados preditos pelo modelo. Formalmente, a função de perda é escrita da seguinte forma: $L(Y, f(x))$. Há vários tipos de funções de perda que podem ser utilizadas, como *Log-loss*, *Cross Entropy Loss*, *Huber*, *MAE* (*Mean absolute error*) ou erro absoluto médio e *MSE* (*Mean squared error*) ou erro quadrático médio, sendo essa última utilizada nesse trabalho. A imagem abaixo exemplifica de maneira mais clara como funciona o método de *Random Forest*.

Figura 5: Random Forest exemplo



Fonte: Implementação de Random Forest (Chakure, 2020)

A Figura 5 exibe um exemplo de *Random Forest* com 600 árvores. O algoritmo é iniciado com os valores de entrada X. Para cada árvore, k variáveis aleatórias são selecionadas. Para cada um dos nós filhos das árvores, é realizado uma predição baseado em condições das k variáveis selecionadas inicialmente. Cada árvore irá produzir uma

predição, ao final a média das predições será calculada e esse será o valor da predição final, isso em um problema de regressão. Já na classificação, o valor que for predito mais vezes será o valor final da predição, funcionando como um sistema de votos.

2.2.7 REGRESSOR EXTRA TREES

O método *Extra Trees*, também conhecido como *Extremely Randomized Tree*, é bastante similar ao método *Random Forest*, entretanto, há algumas pequenas diferenças entre essas técnicas. A principal está na forma em que os nós são criados. Em *Random Forest* os nós são criados a partir do melhor nó, já em *Extra Trees* os nós são criados a partir da escolha de um nó aleatório.

No artigo (Geurts, Ernst, and Wehenkel, 2005), os autores citam que as principais vantagens do método *Extra Trees* está principalmente na acurácia e também na eficiência computacional. Enquanto *Random Forest* testa todas as divisões possíveis sobre uma fração das variáveis, *Extra Trees* testa as divisões aleatórias sobre uma fração das variáveis, o que torna o custo computacional menor e consequentemente também o tempo de treinamento. Esse algoritmo também possui resistência a *overfitting* como sua principal característica (John, John, and Guo, 2018).

2.2.8 REGRESSOR BAGGING

Bagging ou *Boostrap aggregating* é uma técnica que permite reduzir a variância nos dados (Buhlmann, 2012). Essa técnica é popularmente conhecida pela simples implementação. É usado para melhorar a estabilidade e a precisão dos algoritmos de aprendizado de máquina e, assim, como as anteriores, evita o *overfitting*. É um algoritmo amplamente usado para problemas com pequenos conjuntos de dados (Dutt and Krishna, 2019). O valor da predição final é definido tanto por votação, em problemas de classificação, quanto por média, em problemas de regressão.

2.2.9 CONSIDERAÇÕES FINAIS

Esse capítulo tratou sobre a base de dados das cidades da Filadélfia, Estados Unidos, e de Fortaleza. Foi apresentado também os métodos de regressão que foram utilizadas para a predição da localização das ocorrências criminais.

No capítulo 3, Metodologia, será apresentado o processo para a obtenção dos resultados utilizando as bases de dados e as técnicas de regressão apresentadas neste capítulo.

3 METODOLOGIA

Neste capítulo será apresentado o processo metodológico utilizado neste trabalho para atingir os objetivos propostos. Será apresentado como foi realizado a análise exploratória dos dados, desde a obtenção até a geração dos resultados. Será apresentado também os experimentos que foram realizados com a base de dados da cidade de Fortaleza. Por fim, será explanado sobre a engenharia e seleção de atributos e também sobre o treinamento dos modelos.

3.1 ANÁLISE EXPLORATÓRIA DOS DADOS

A depuração e a análise dos dados devem ser o ponto de partida em qualquer problema de aprendizado de máquina, pois dessa forma, há um melhor entendimento sobre o que os dados representam e como eles podem ser utilizados na fase de predição. Analisar os dados faz parte do ciclo de vida de um projeto de ciência de dados. A Figura 6 exibe os passos, conhecido como *framework* OSEMN (Dineva and Atanasova, 2018), que são necessários para um projeto de ciência de dados.

Figura 6: *Framework* OSEMN

Fonte: Autoria própria.

Na Figura 6, o primeiro passo é conhecido como *obtain* ou obter. Os dados podem ser coletados praticamente de qualquer lugar, como redes sociais, exames médicos, sensores etc. O conjunto de dados utilizado nesse trabalho foi coletado ao longo dos anos na cidade da Filadélfia, Estados Unidos, e na cidade de Fortaleza através de registros policiais. A maioria das bases coletadas apresentam falhas, como dados faltantes, por exemplo. Para realizar o tratamento desses dados é utilizado o segundo passo, *scrub* ou limpeza, nessa etapa os dados desnecessários são removidos ou substituídos. Na terceira etapa, *explore* ou explorar, a propriedade dos dados é verificada. Em uma base de dados há diferentes tipos de dados, como numérico, categóricos, datas e etc, para cada um desses dados é necessário realizar um tratamento diferente, seja para extração de novos dados ou para conversão. O quarto passo, *model* ou modelo, é onde os algoritmos de aprendizado de máquina serão utilizados para realizar classificação ou regressão sobre os dados, essa etapa é completamente dependente da etapa anterior, ou seja, uma ótima análise exploratória dos dados influi diretamente nas predições do modelo. Após o modelo ter realizado as predições, é necessário interpretar os dados, que é a última etapa, *interpret*. Aqui é necessário dar significado ao que o modelo apresentou como saída, o que aquela predição representa e como ela pode ser aplicada. Esse tipo de inferência pode ser apresentada de

forma gráfica, permitindo um melhor entendimento por parte do público.

3.1.1 ANÁLISE EXPLORATÓRIA DA BASE DE DADOS DA CIDADE DA FILADÉLFIA

A base de dados da cidade da Filadélfia, Estados Unidos, é constituída inicialmente por 2.237.605 linhas e 14 colunas. A primeira verificação feita foi em relação ao número de dados faltantes em cada uma das colunas. A Figura 7 apresenta as colunas da base de dados e sua respectiva quantidade de dados faltantes.

Figura 7: Dados faltantes - base de dados Filadélfia, Estados Unidos

Dc_Dist	0
Psa	0
Dispatch_Date_Time	0
Dispatch_Date	0
Dispatch_Time	0
Hour	0
Dc_Key	0
Location_Block	0
UCR_General	663
Text_General_Code	663
Police_Districts	19930
Month	0
Lon	17349
Lat	17349

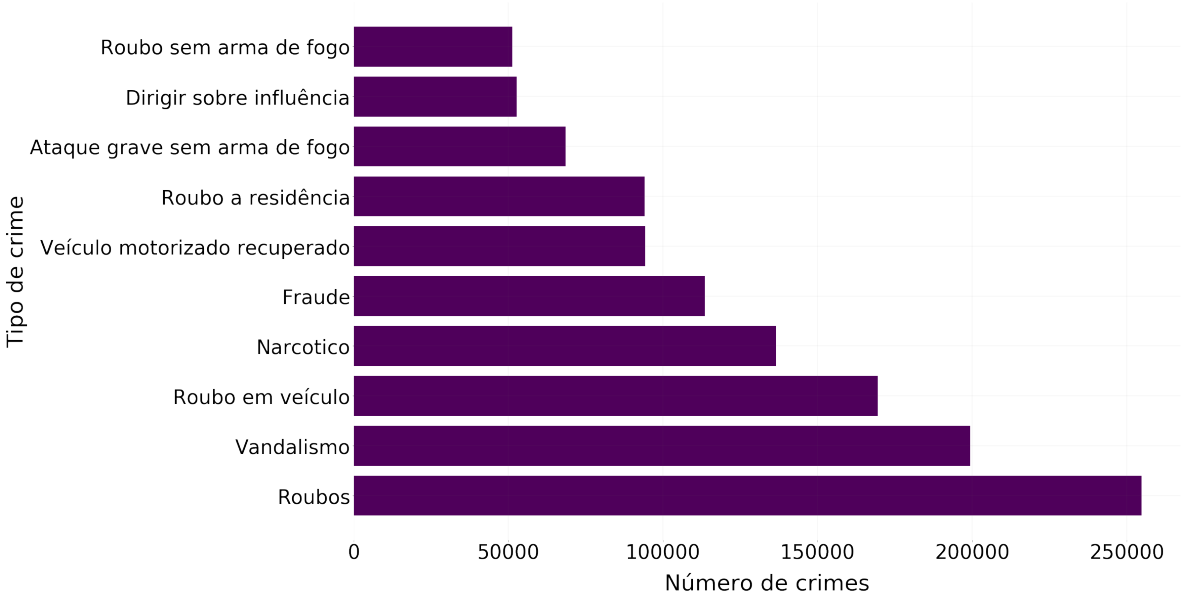
Fonte: Autoria própria.

É possível notar na Figura 7 que as principais colunas como *UCR_General*, que armazena o código do tipo de crime cometido, *Lon* e *Lat*, que armazenam os dados de longitude e latitude, respectivamente, possuem dados faltantes, sendo necessário tratar ou remover esses dados. Foi identificado no total 18,012 linhas que possuem dados faltantes, esse valor foi encontrado da seguinte forma: a soma dos dados faltantes de *Lon* ou *Lat* com os dados faltantes de *UCR_General*, totalizando 18.012. A coluna *Police_Districts* não foi removida por isso ela foi contabilizada. Assim a soma conjunta daria 18,012. Dividindo o resultado encontrado da soma pelo total de linhas, 2,237,605, um percentual de 0.8% é encontrado, trata-se de um valor muito pequeno, dessa forma é mais viável apenas remover as linhas com valores faltantes do que tratá-los. Após a remoção, a nova base ficou com 2,219,593 linhas.

Com a nova base de dados foi realizado a extração de novas variáveis. A partir da coluna *Dispatch_Date*, que representa a data em que ocorreu o crime, foi criado as seguintes variáveis: *Year*, *Day* e *DayofWeek*, que representam o ano, dia e o dia da semana, respectivamente, em que houve a ocorrência do crime. Essas novas variáveis representam tempo, aspecto importante na predição de crimes. A base de dados passou a ter 16 colunas.

A base de dados da Filadélfia, Estados Unidos, possui informações sobre diversos crimes, como: dirigir sobre influência de alguma substância, roubo de veículos motorizados etc. É necessário descobrir quais são os crimes que ocorrem com maior frequência para assim realizar a predição sobre um crime específico. A Figura 8 exibe os dez crimes que mais ocorreram na cidade da Filadélfia, Estados Unidos.

Figura 8: Crimes com maior ocorrência



Fonte: Autoria própria.

Pela Figura 8 é possível perceber que o tipo de crime *Thefts* ou roubos possui a maior frequência, com exatos 254.714 ocorrências. Esse será o tipo de crime que o modelo irá realizar a predição do local onde poderá acontecer futuras ocorrências. Com um tipo de crime selecionado, a nova base de dados passará a ter 254,714 linhas e 16 colunas.

3.1.2 ANÁLISE EXPLORATÓRIA DA BASE DE DADOS DA CIDADE DE FORTALEZA

O processo de análise dos dados para a base de dados de Fortaleza seguiu o mesmo padrão utilizado na cidade da Filadélfia, Estados Unidos. Inicialmente foi verificado se havia dados faltantes no conjunto de dados relacionados a cidade de Fortaleza, os dados estavam completos, não foi necessário remover ou tratá-los. A base de dados de Fortaleza também possui uma coluna de data e a partir dela foi extraída quatro novas variáveis, são elas: ano, mês, dia e dia da semana.

Na análise exploratória da base de dados da cidade de Fortaleza, um ponto a ser notado foi em relação aos valores de latitude e longitude, para a cidade de Fortaleza. Esses valores estavam em -3.71839 com variações para latitude e -38.5434 e variações para a longitude, entretanto, havia alguns valores que não estavam de acordo, como por exemplo, -4.185814913 e -38.136553005, latitude e longitude, respectivamente. Não

seria viável pesquisar a localização desses valores de forma individual, assim, foi utilizado inicialmente a *API* do Google (Google, 2020), onde era passado como parâmetro o nome da rua e o retorno era constituído pelos valores de latitude e longitude, entretanto, o custo financeiro para depois de 2,500 requisições era um pouco grande. A alternativa encontrada foi a utilização da biblioteca *GeoLocator* (GeoPy, 2020), que também recebe a rua como parâmetro e retorna algumas informações, como latitude e longitude, por exemplo. O *GeoLocator* foi utilizado juntamente com um *proxy* para permitir realizar um maior número de requisições em um curto período de tempo. A Figura 9 exibe o resultado das requisições e mesmo os dados sendo rotulados como Fortaleza, as informações de latitude e longitude estavam incorretas.

Figura 9: *Resultado requisição GeoLocator*

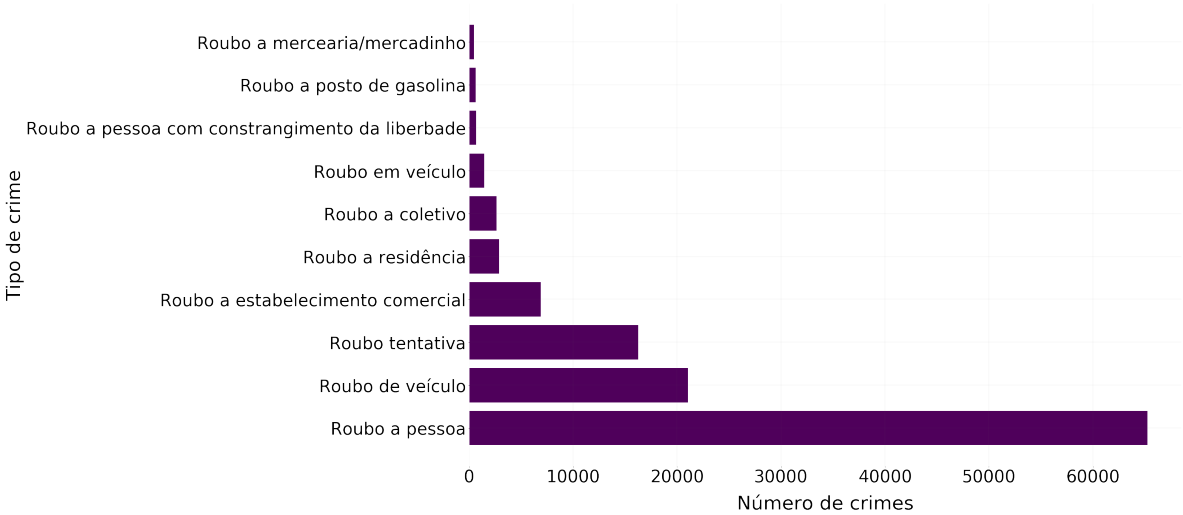
localizacao	endereco
-3.7303581789, -38.5329432917	Rua 24 de Maio, 936-1002 - Centro, Fortaleza - CE, 60020-000, Brazil
-4.185814913, -38.136553005	Beberibe - State of Ceará, 62840-000, Brazil
-3.6789516394, -40.3721261833	Terrenos Novos, Sobral - CE, 62031-050, Brazil
-3.7032230879, -40.340070009499996	Cidade Gerardo Cristino de Menezes, Sobral - CE, Brazil
-7.2328276025, -39.3398809637	R. Cecília Silva de Souza, 140 - São José, Juazeiro do Norte - CE, 63024-480, Brazil

Fonte: Autoria própria.

Na Figura 9 é possível notar que valores das cidades de Sobral, Juazeiro do Norte e Beberibe estavam contidos na base de dados, havia valores de outras cidades também. No total foram removidas 3.781 linhas com informações de outras cidades. Uma análise inicial dos *outliers*, valores que fogem do padrão dos outros dados, possibilitaria identificar essas discrepâncias. A nova base possui 118.224 linhas.

A base de dados da cidade de Fortaleza também possui informações sobre diversos crimes como: roubo a mercearia, roubo a coletivo, roubo de veículo, etc. Foi feito a verificação de qual crime ocorreu mais durante o período de 2015 a 2019. A Figura 10 apresenta os dez crimes com maior ocorrência.

Figura 10: Ocorrência de crimes na cidade de Fortaleza



Fonte: Autoria própria.

Pela Figura 10 é possível notar que o crime de roubo a pessoa é o que possui maior frequência, 65.240, para ser mais específico. Esse será o tipo de crime explorado para a base de dados de Fortaleza.

A fase de exploração dos dados foi importante pois permitiu validar algumas hipóteses, como em que bairro a taxa de violência é maior, que tipo de crime ocorre mais em uma determinada rua etc. Com base nessas validações, os dados foram modelados para que as técnicas de aprendizado de máquina pudessem cumprir o objetivo principal do trabalho, que é predizer os locais onde os crimes irão ocorrer.

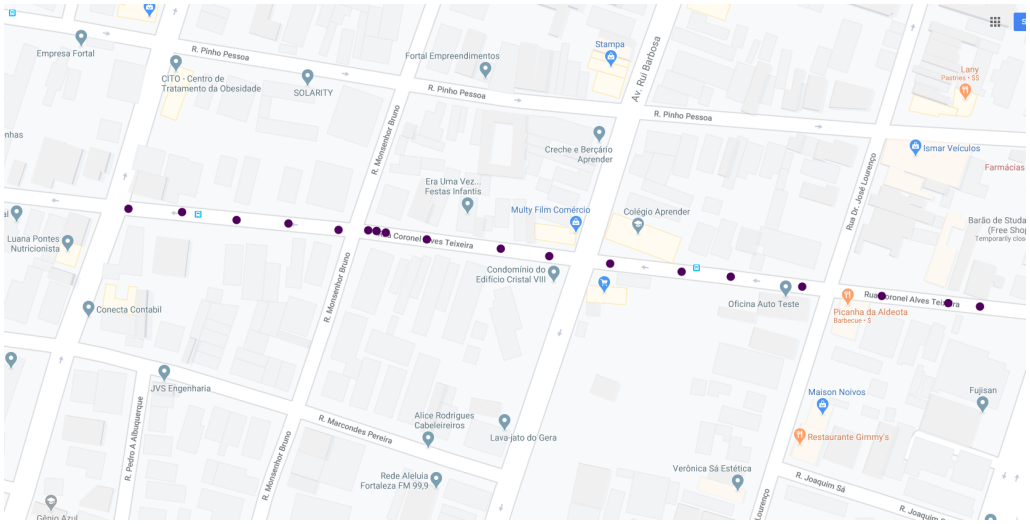
3.2 EXPERIMENTOS

Esse trabalho é dividido em dois experimentos, sendo que cada experimento possui mais duas categorias de testes. Esses experimentos poderiam, futuramente, ajudar a criar uma rota para realização de patrulhas em uma determinada área. As rotas poderiam ser baseadas em periculosidade de uma determinada rua ou até mesmo bloco. Os experimentos possuem também como objetivo identificar se o modelo de aprendizado de máquina consegue aprender os padrões com os dados distribuídos ou com os dados concentrados. Cada um dos experimentos serão explanados a seguir.

3.2.1 EXPERIMENTO 1

No experimento um é considerado que os crimes de roubo a pessoa ocorrem em múltiplos pontos de uma rua. Isso significa que para uma mesma rua possa existir diversos pontos onde ocorreu determinado crime, não tendo assim um padrão que possa ser aprendido pelo modelo. A Figura 11 exibe um exemplo sobre a distribuição dos pontos de crime em uma rua.

Figura 11: Múltiplos pontos de crime em uma rua

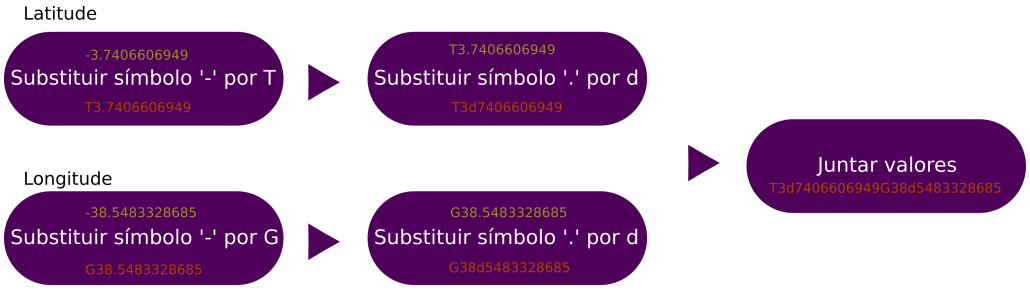


Fonte: Autoria própria.

Na imagem acima é possível notar que na Rua Coronel Alves Teixeira, em Fortaleza, ocorreram diversos crimes em diversos pontos diferentes, onde, para cada um dos crimes registrados, há uma latitude e longitude diferentes. Nesse experimento será avaliado se essa diferença de pontos irá interferir no aprendizado do modelo.

O experimento um é dividido em duas categorias. Na primeira categoria é considerado que latitude e longitude são valores distintos, portanto, na predição, a saída esperada do modelo também serão dois valores distintos. Já para a segunda categoria, os valores de latitude e longitude tornam-se um só. Para que isso seja possível, é criado um valor *hash* a partir dos valores originais. Essa valor *hash* deve ser convertido posteriormente em um valor numérico, para ser aceito pelos algoritmos de aprendizado de máquina, podendo ser usado o método de *LabelEncoder*, por exemplo. A Figura 12 exibe um exemplo simples de conversão dos valores de latitude e longitude para *hash*. Essa estratégia, de *hash*, foi adotada nesse trabalho como forma de experimentação.

Figura 12: Conversão de latitude e longitude para *hash*



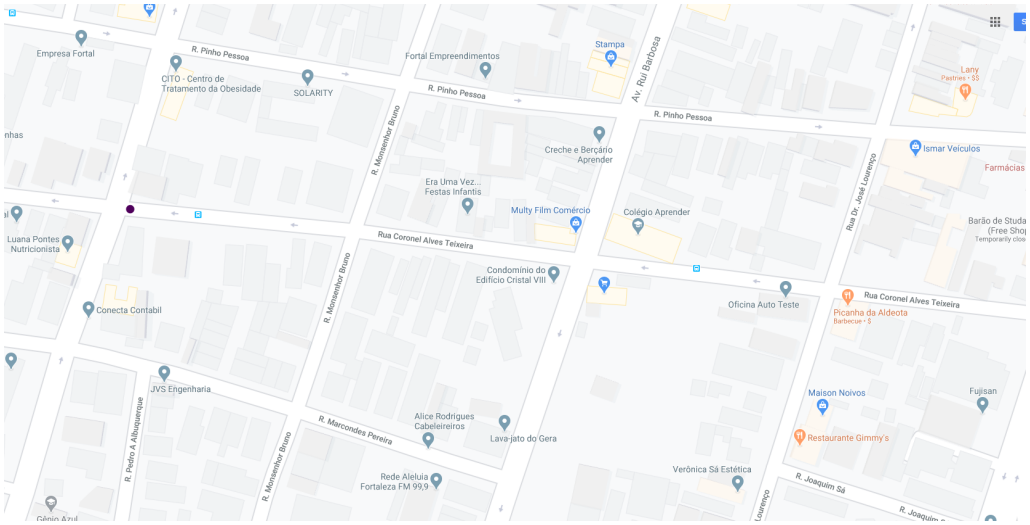
Fonte: Autoria própria.

Para realizar o processo contrário, ou seja, converter de *hash* para valores de latitude e longitude será necessário apenas realizar o processo de forma inversa. Essa estratégia de utilizar apenas o valor *hash* permite reduzir, computacionalmente, a complexidade do problema pois agora haverá apenas uma variável de saída para realizar a predição.

3.2.2 EXPERIMENTO 2

No experimento dois, o crime de roubo a pessoa é quantificado como se ocorresse em apenas um ponto da rua, ou seja, agora ao invés os valores de latitude e longitude estarem distribuídos, agora esses valores estão concentrados em um único local. A Figura 11 exibiu dezoito registros de roubo a pessoa que ocorreu na Rua Coronel Alves Teixeira. A Figura 13 exibe o mesmo número de crimes, entretanto, que ocorreu em apenas um ponto.

Figura 13: Único ponto de crime em uma rua



Fonte: Autoria própria.

A Figura 13 exibe a estratégia utilizada no segundo experimento, a quantidade de registros permanece a mesma, dezoito. Com a concentração da quantidade de crimes para apenas um valor de latitude e um valor de longitude, pode ajudar o modelo de aprendizado a encontrar um padrão, dessa forma auxiliando a realizar uma predição de roubo a pessoa que possa ocorrer naquela determinada rua, permitindo também concentrar um patrulhamento mais intensivo para aquela localidade.

Assim como o experimento um, o experimento dois também é dividido em duas categorias. Na primeira categoria também haverá duas saídas e na segunda categoria haverá apenas uma saída, o valor convertido em *hash*. O valor de *hash* também pode ser encontrado utilizando o mesmo fluxo apresentado na Figura 12.

3.3 ENGENHARIA DE ATRIBUTOS

A engenharia de atributos permite criar variáveis a partir de outras já existentes, esse processo é geralmente realizado no pré-processamento dos dados. Na base de dados da Filadélfia, Estados Unidos, foi criada três novas variáveis, *Year*, *Day* e *DayofWeek* a partir de um atributo já existente, o *Dispatch_Date*. A criação de novas variáveis permite que o modelo realize predições mais condizentes com a realidade, o que não implica dizer que a não criação de novas variáveis resultará em um modelo ruim. O processo de engenharia de atributos requer uma série de alterações e testes no modelo para verificar se o novo atributo criado melhora ou não o modelo preditivo.

Na base de dados da cidade de Fortaleza, novos atributos também foram criados. A partir da data, foram criados os atributos ano, mês, dia e dia da semana. A partir do atributo hora, foi criada a variável noite ou dia, que representa se um crime foi cometido durante o dia ou durante a noite.

Data e hora são dados cíclicos, isso implica uma variação ao longo do tempo,

ou seja, os dados se repetem com uma determinada frequência, definida por um ciclo de horas do dia, ou dias da semana ou do mês, dessa forma, o modelo deve saber que os dados utilizados pertencem a classe cíclica. Para realizar o processamento desses dados, é necessário usar a trigonometria, mais especificamente, deve-se converter esses dados na representação de seno e cosseno. A partir do mês, dia, dia da semana, hora e minuto, novas variáveis foram geradas para seno e cosseno.

A partir dos dados de latitude e longitude, foi gerada uma nova variável chamada rua, que representa o nome da rua onde ocorreu um determinado crime. Após gerar essa variável, foi notado que os dados de crimes de outras cidades como Sobral e Itapipoca, por exemplo, ainda permaneciam na base de dados, o que poderia acarretar em mal aprendizado por parte dos algoritmos. Esses dados foram removidos da base.

Após gerar o novo atributo rua, foram criadas duas novas variáveis, chamadas *new-lat* e *new-log*. Como citado inicialmente, para a base de dados de Fortaleza, foi realizado quatro experimentos divididos em duas categorias. No segundo experimento de cada categoria, será utilizado as novas variáveis *new-lat* e *new-log*, pois elas indicam a latitude e longitude com maior frequência para uma mesma rua. A Tabela 1 exibe as variáveis criadas para a base de dados de Fortaleza.

Tabela 1: Variáveis criadas - *Fortaleza*

Variável original	**Variáveis criadas**
data	ano, mês, dia e dia da semana
mês	seno do mês e cosseno do mês
dia	seno do dia e cosseno do dia
dia da semana	seno do dia da semana e cosseno do dia da semana
rua	new-lat e new-log

Fonte: Autoria própria.

3.4 SELEÇÃO DE ATRIBUTOS

Para o conjunto de dados da cidade da Filadélfia, Estados Unidos, não foi utilizado nenhuma técnica de seleção de atributos. Como se tratava de uma base para experimentos, foi utilizado somente os atributos relacionados a temporalidade. A Tabela 2 exibe as variáveis que foram utilizadas.

Tabela 2: Variáveis utilizadas - Filadélfia, Estados Unidos

Variáveis de entrada	**Variáveis de saída**
Hour, Month, Year Day, DayofWeek	Lat e Long

Fonte: Autoria própria.

Já para a base de Fortaleza, após todo o processo de engenharia de atributos ter sido realizado, doze novos atributos foram adicionados a base de dados, totalizando

dezenove variáveis: dezessete independentes e duas dependentes. Nem todos esses atributos são considerados importantes no momento do treinamento, dessa forma, foi utilizado o *Multi Task Lasso CV* para selecionar os atributos com maior grau de importância. Essa técnica linear permite selecionar os melhores atributos através do teste de validação cruzada, testando diversas combinações de atributos, permitindo selecionar as melhores variáveis do conjunto de dados. Para cada um dos experimentos realizados, foi necessário a utilização do *Multi Task Lasso CV* para selecionar as variáveis. Nas subseções seguintes são exibidas os atributos selecionados para cada um dos experimentos.

3.4.1 EXPERIMENTO 1 - CATEGORIA 1

Na primeira categoria do experimento um, foi utilizado os valores de latitude e longitude para todos os pontos onde ocorreram crimes em determinada rua.

Com a utilização do *Multi Task Lasso CV*, quinze variáveis foram selecionadas. A Tabela 3 exibe as variáveis de entrada e também de saída que serão utilizadas para o treinamento desse experimento.

Tabela 3: Variáveis utilizadas - Experimento 1 - Categoria 1

Variáveis de entrada	**Variável de saída**
ano, mes, dia da semana hora, noite ou dia, seno da hora cosseno da hora, seno do minuto cosseno do minuto, seno do mês cosseno do mês, seno do dia da semana cosseno do dia da semana	latitude e longitude

Fonte: Autoria própria.

3.4.2 EXPERIMENTO 1 - CATEGORIA 2

Na segunda categoria do experimento um, os pontos de latitude e longitude foram convertidos em apenas um valor. Esse valor foi utilizado junto ao *Multi Task Lasso CV* para identificar o grau de importância das variáveis.

A Tabela 4 exibe as dez variáveis que foram selecionadas para o treinamento.

Tabela 4: Variáveis utilizadas - Experimento 1 - Categoria 2

Variáveis de entrada	**Variáveis de saída**
ano, sen_hora, mes horan, cos_minuto, dia, minuto dia_da_semana, cos_dia_da_semana, cos_hora	lat_log_encoder

Fonte: Autoria própria.

3.4.3 EXPERIMENTO 2 - CATEGORIA 1

Para categoria um do experimento dois foram considerados os pontos de latitude e longitude apenas para um local específico em determinada rua.

Para uma melhor compreensão das variáveis selecionadas, a Tabela 5 exibe os atributos que serão utilizados nesse experimento.

Tabela 5: Variáveis utilizadas - Experimento 2 - Categoria 1

Variáveis de entrada	**Variáveis de saída**
noite_ou_dia, cos_hora, cos_dia_da_semana dia_da_semana, sen_minuto, sen_dia_da_semana dia, mes, horan, cos_minuto cos_mes, ano, sen_mes, sen_hora	latitude e longitude

Fonte: Autoria própria.

3.4.4 EXPERIMENTO 2 - CATEGORIA 2

Assim como na categoria dois do experimento um, a saída na categoria dois do experimento dois também será convertida para apenas um valor.

Como nos experimentos anteriores, a Tabela 6 sintetiza as variáveis que foram utilizadas para o experimento explanado. Para a categoria dois do experimento dois foram selecionados nove atributos.

Tabela 6: Variáveis utilizadas - Experimento 2 - Categoria 2

Variáveis de entrada	**Variáveis de saída**
sen_hora, ano, mes, horan dia, minuto, dia_da_semana cos_dia_da_semana, cos_hora	lat_log_encoder

Fonte: Autoria própria.

3.5 TREINAMENTO

Com os dados devidamente limpos e organizados, foi realizado a etapa de treinamento dos modelos. Para cada uma das técnicas apresentadas na seção de materiais e métodos, uma série de parâmetros foram testados para obter uma melhor saída (predições). Neste trabalho, as três fases do aprendizado de máquina foram respeitadas, são elas:

O treinamento. Nessa fase os dados foram expostos a cada uma das técnicas para que dessa forma os padrões pudessem ser aprendidos pelo algoritmo. É nessa etapa também em que os dados de treinamento são usados para se ajustar ao modelo. Foi dividido 80% dos dados para treinamento.

O teste. Nessa etapa os dados de testes são utilizados para realizar a comparação de performance com os dados de treinamento, se o modelo conseguir obter um ótimo índice de predição nos dados de treino mas não nos dados de teste, então há um *overfitting* e o modelo precisa ser ajustado. Foi reservado 20% dos dados para a fase de teste.

Validação. Nessa etapa, dados que não foram apresentados ao modelo são utilizados para validar se o algoritmo está realizando as predições de forma esperada.

Após as etapas de engenharia de atributos e de seleção de atributos, 6.950 instâncias do conjunto de dados inicial foram removidas para serem usadas como dados de validação em uma situação real. É importante enfatizar que esses dados não foram apresentados em momento algum aos algoritmos de aprendizado, eles foram usados apenas para realizar as predições finais.

As métricas MSE (*Mean Squared Error*) e RMSE (*Root Mean Squared Error*), respectivamente apresentadas nas Equações (3) e (4), foram utilizadas para avaliar os modelos.

$$MSE = \left(\frac{1}{n}\right) \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3}$$

Na Equação (3), y_i representa a saída esperada e $\hat{y}_i$ representa a predição realizada pelo modelo. MSE calcula o a média quadrada dos erros para as predições. Comumente um valor MSE alto indica que o modelo não é bom e precisa de mais ajustes dos parâmetros.

$$RMSE = \sqrt{\left(\frac{1}{n}\right) \sum_{i=1}^{n} (y_i - \hat{y}i)^2} \tag{4}$$

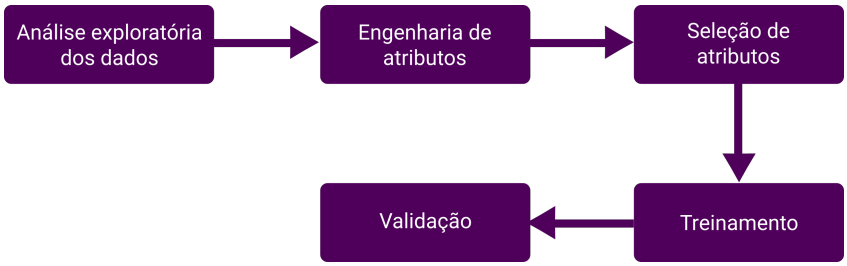
Na Equação (4) y_i e $\hat{y}_i$ representa o mesmo descrito na Equação (3). A diferença é que agora a raiz de MSE é calculada. Isso permite que a escala dos erros seja igual à escala da saída esperada.

O processo de treinamento para cada um dos modelos foi realizado de forma diferente. Métodos, como Floresta aleatória (*Random Forest Regressor*), foram variados usando mais de um parâmetro para encontrar o melhor ajuste, outros métodos, como *Bagging Regressor*, foram variados usando apenas um parâmetro para encontrar um resultado ideal. Os resultados do treinamento e os parâmetros usados são apresentados no capítulo seguinte.

3.6 FLUXO METODOLÓGICO

A metodologia aplicada nesse trabalho precisou ser dividida em várias partes e cada uma das etapas possui uma dependência com a etapa anterior. O fluxo metodológico permite visualizar de forma mais clara os módulos que foram aplicados na pesquisa, a Figura 14 exibe esse fluxo.

Figura 14: Fluxo metodológico



Fonte: Autoria própria.

A Figura 14 exibe os principais passos que foram utilizados para obter os resultados apresentados no capítulo seguinte. Na análise exploratória dos dados foi realizado a limpeza e análise dos dados. Nessa fase foram verificados quais crimes ocorreram com mais frequência, além também de ser verificado em quais períodos do dia e da semana a incidência de crimes foi maior. Em engenharia de atributos foi realizado a criação e extração de novas variáveis a partir das já existentes. Na fase de seleção de atributos foi realizado a verificação de quais atributos apresentavam uma importância maior para serem utilizados no treinamento. A fase de treinamento utilizou-se de algumas técnicas de aprendizado de máquina para encontrar padrões nos dados apresentados e consequentemente realizar as devidas predições. Na última fase, validação, foi verificado o comportamento dos dados preditos em relação aos dados originais de validação, e permitiu indicar qual modelo apresentou o melhor resultado.

3.7 CONSIDERAÇÕES FINAIS

Neste capítulo foram detalhados os passos que foram seguidos para a obtenção dos dados de treinamento. Nas seções apresentadas foi mostrado os detalhes das análises exploratórias que foram realizadas para a base de dados da Filadélfia, Estados Unidos, e também de Fortaleza.

O Capítulo a seguir apresenta os resultados que foram obtidos a partir dos experimentos que foram propostos. Os resultados que serão apresentados exibem as predições que foram realizadas para a cidade da Filadélfia, Estados Unidos, e também para a cidade de Fortaleza.

4 RESULTADOS

Neste Capítulo serão apresentados os resultados obtidos para os experimentos utilizando a base de dados da cidade da Filadélfia, Estados Unidos, e para a base de dados da cidade de Fortaleza, utilizando os métodos citados no capítulo 3.

4.1 BASE DE DADOS FILADÉLFIA

Para a base de dados da Filadélfia, Estados Unidos, os resultados foram obtidos utilizando-se somente considerando que os crimes ocorriam em vários pontos de uma rua. A Tabela 7 apresenta os erros e os parâmetros para cada uma das técnicas apresentadas na seção de materiais e métodos.

Tabela 7: Resultado dos experimentos

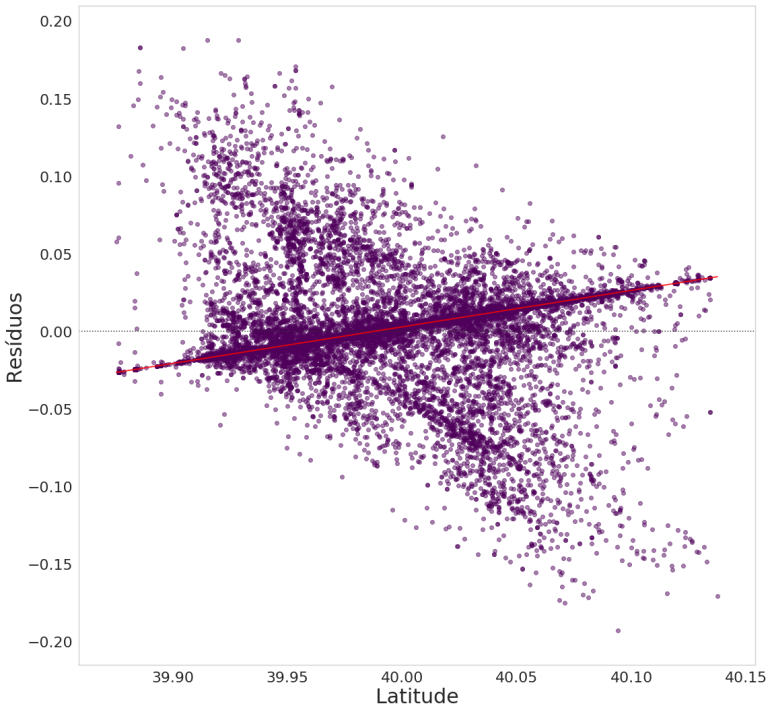
	Parâmetros			**Erro**	
Método	**Número de estimadores**	**Profundidade máxima**	**Número de vizinhos**	**MSE**	**RMSE**
K-Nearest Neighbor	-	-	1	0.00156	0.03951
Random Forest	1	-	-	0.00089	0.02984
Extra Trees	1	24	-	0.00178	0.04228
Decision Tree	**-**	**45**	**-**	**0.00055**	**0.02360**
Bagging	1	-	-	0.00091	0.03024

Fonte: Autoria própria.

Para cada um dos métodos discutidos, foram utilizados diferentes parâmetros para a obtenção de melhores resultados. Alguns métodos, como *K-Nearest Neighbor* e *Bagging* foram treinados com apenas um parâmetro sendo variado. Enquanto que outros métodos, como *Extra Tress* e *Decision Tree* foram treinados utilizando dois parâmetros. Utilizando os dados da cidade da Filadélfia, Estados Unidos, para realizar os experimentos, foi verificado que o método de *Decision Tree* obteve um menor erro, como pode ser verificado nas colunas de **MSE** e **RMSE** na Tabela 7. Entretanto, apenas os valores dos erros não é o suficiente para avaliar se o modelo realizou as predições de maneira correta.

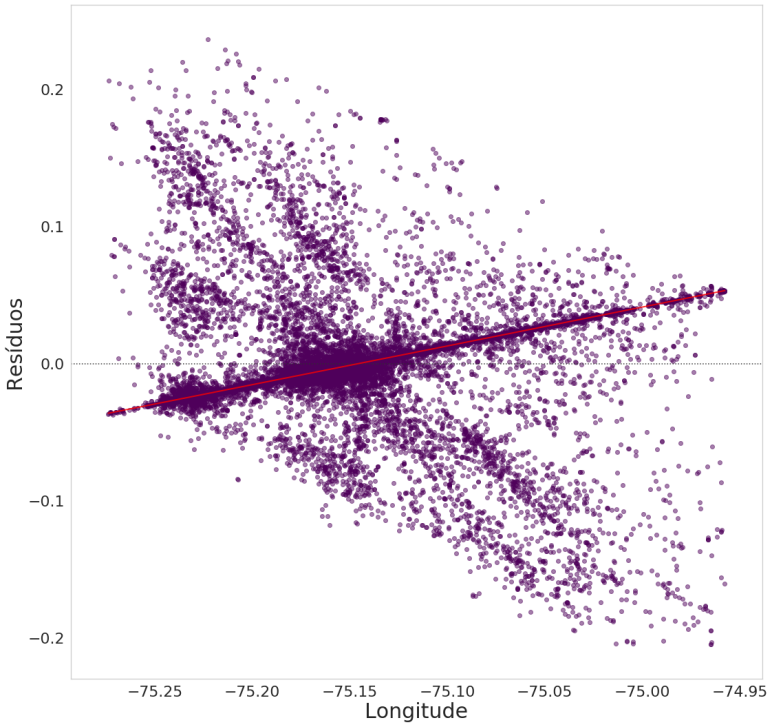
Nesse trabalho, também foi utilizado o cálculo dos resíduos para avaliar os métodos. O resíduo é a diferença entre o valor real e o valor predito, essa medida permite verificar a tendência dos valores encontrados pelo modelo. Figuras 15 a 24 apresentam os resíduos calculados para cada um dos regressores.

Figura 15: Resíduos para Latitude utilizando K-Nearest Neighbor



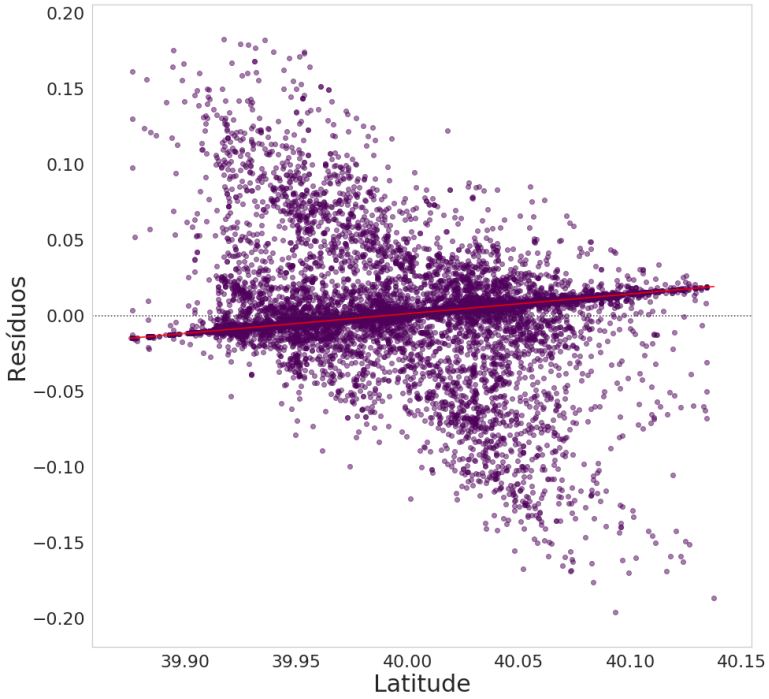
Fonte: Autoria própria.

Figura 16: Resíduos para Longitude utilizando K-Nearest Neighbor



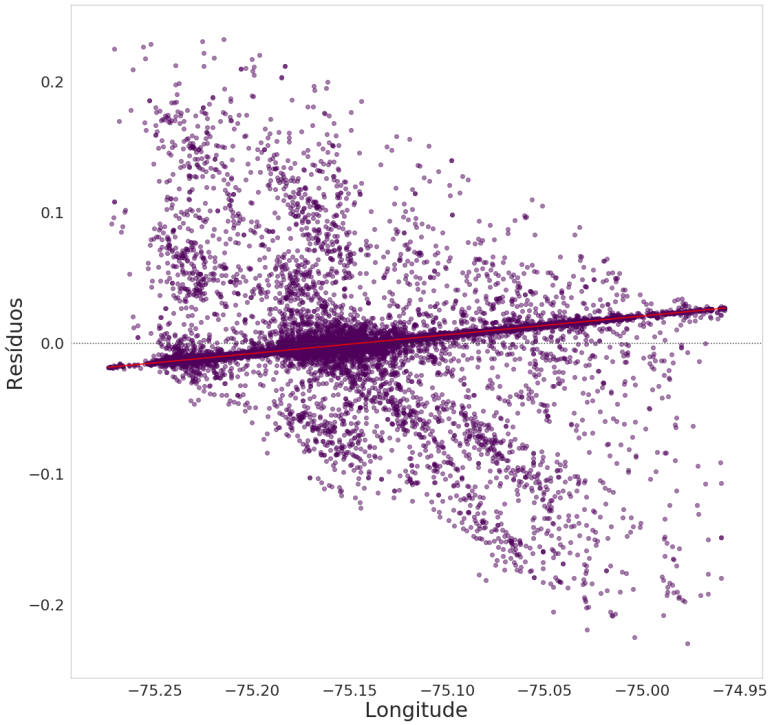
Fonte: Autoria própria.

Figura 17: Resíduos para Latitude utilizando Random Forest



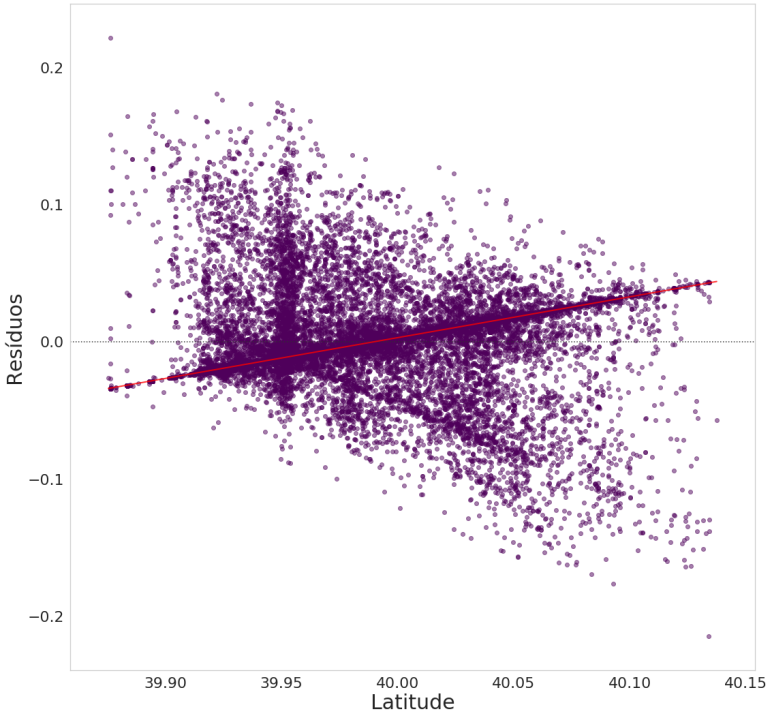
Fonte: Autoria própria.

Figura 18: Resíduos para Longitude utilizando Random Forest



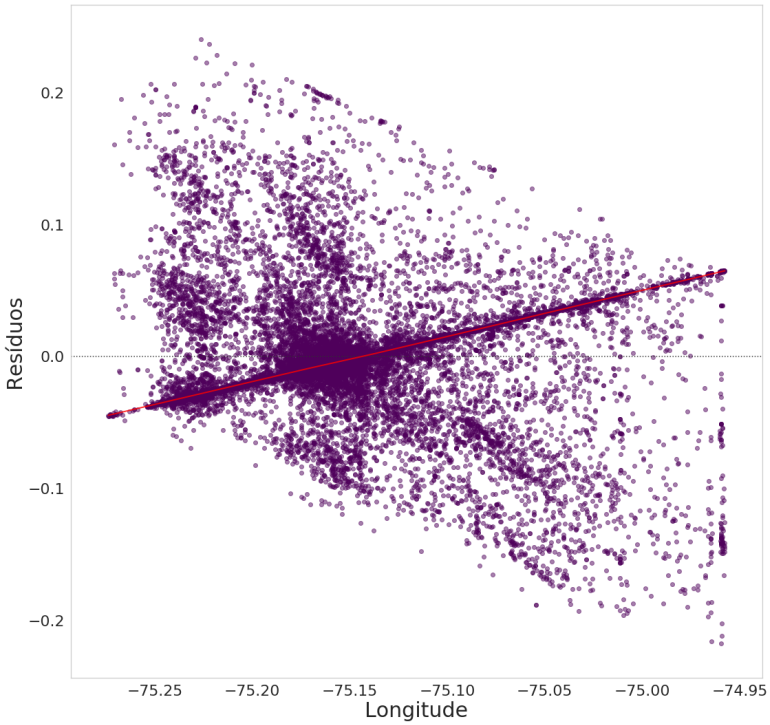
Fonte: Autoria própria.

Figura 19: Resíduos para Latitude utilizando Extra Trees



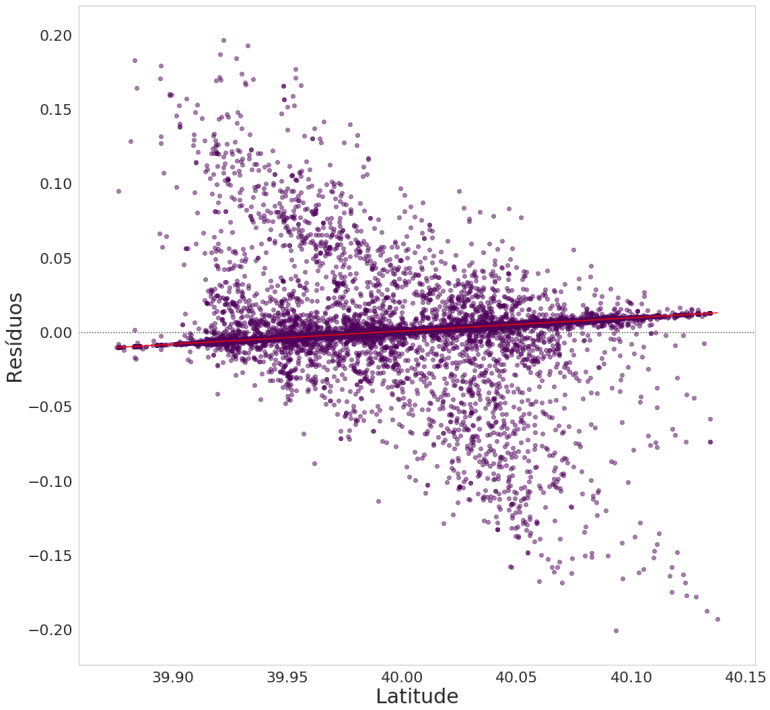
Fonte: Autoria própria.

Figura 20: Resíduos para Longitude utilizando Extra Trees



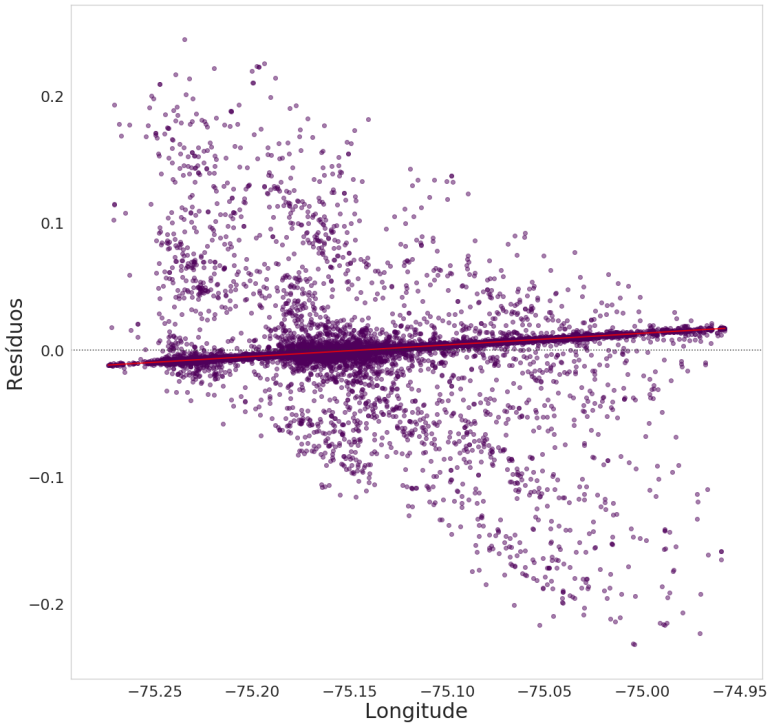
Fonte: Autoria própria.

Figura 21: Resíduos para Latitude utilizando Decision Tree



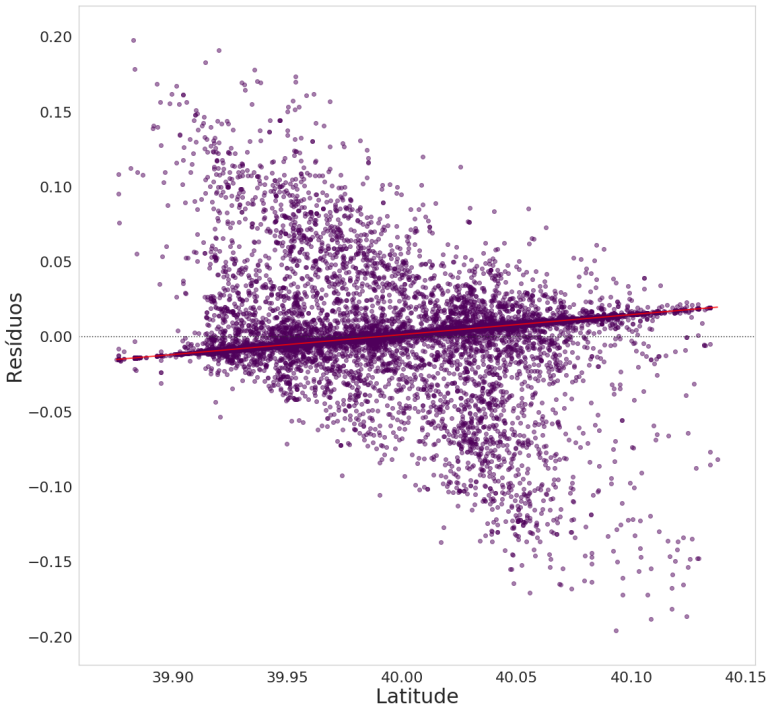
Fonte: Autoria própria.

Figura 22: Resíduos para Longitude utilizando Decision Tree



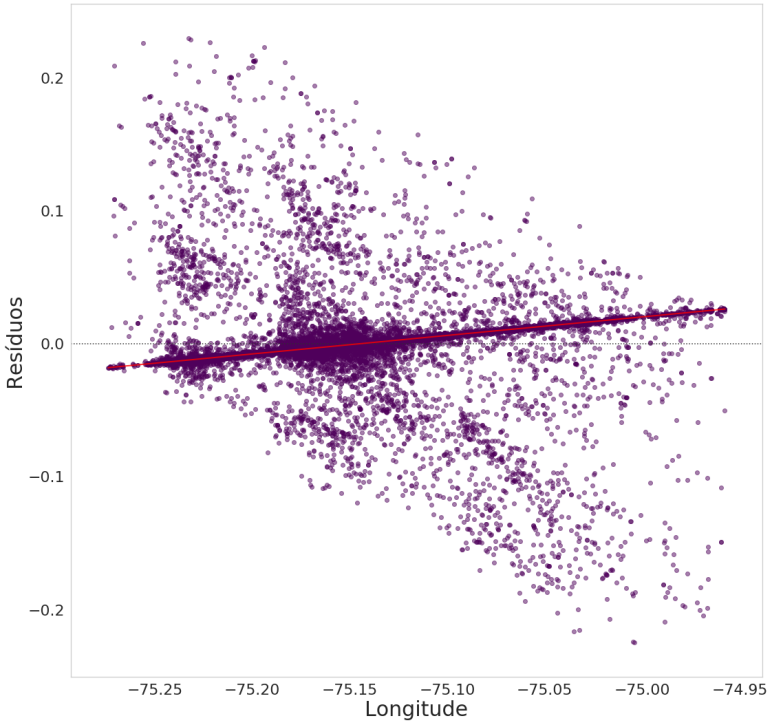
Fonte: Autoria própria.

Figura 23: Resíduos para Latitude utilizando Bagging



Fonte: Autoria própria.

Figura 24: Resíduos para Longitude utilizando Bagging



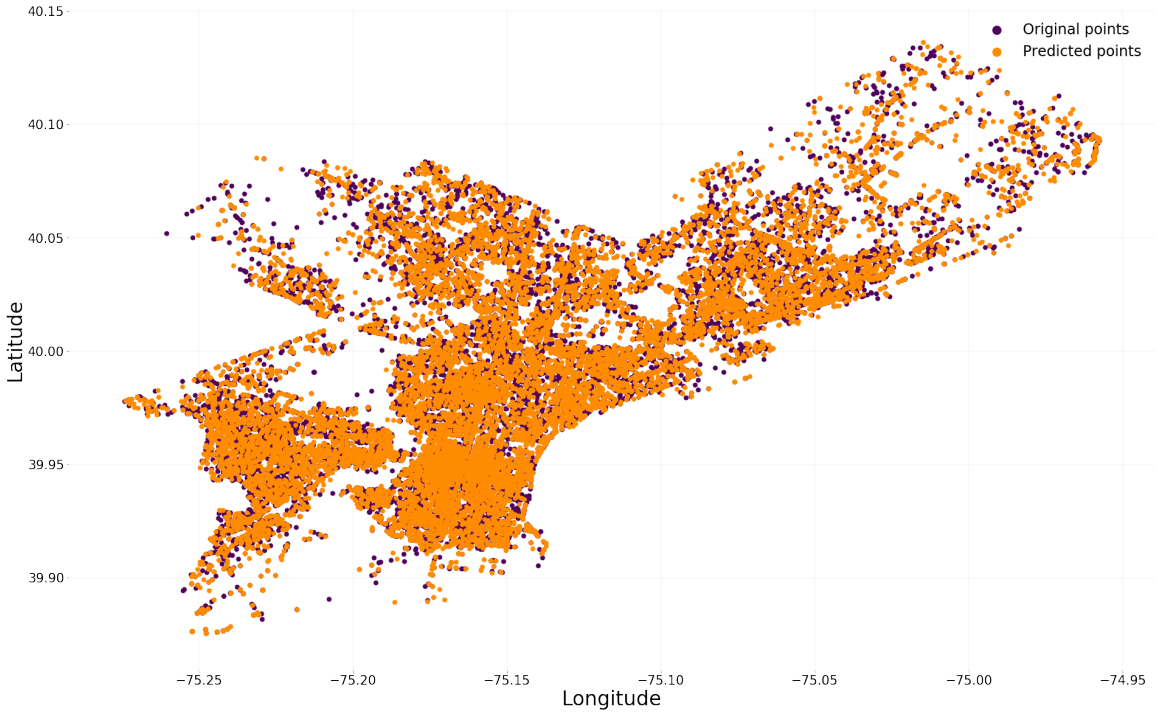
Fonte: Autoria própria.

Nas Figuras acima, que apresentam os valores dos resíduos, que é a diferença entre os valores esperados e os valores preditos, é possível notar uma linha vermelha no sentido horizontal. Essa linha permite avaliar o relacionamento entre os valores de

entrada e os valores de saída, ou seja, o quão os valores preditos diferem dos valores reais. Para um modelo ser considerado ideal, a linha vermelha deve tender a zero. Avaliando cada uma das figuras, podemos observar que para o método de *Decision Tree Regressor* a linha está se comportando da forma esperada, está se aproximando de zero, o que acaba corroborando com o resultado de **MSE** e **RMSE** exibidos na Tabela 7. Comparando também o maior erro da Tabela 7, resultante do método *Extra Trees Regressor*, com os resíduos encontrados nas Figuras 19 e 20, observa-se que a linha vermelha possui uma angulação maior quando comparado aos outros métodos, principalmente em relação aos resíduos dos dados de Longitude, como é exibido na Figura 20.

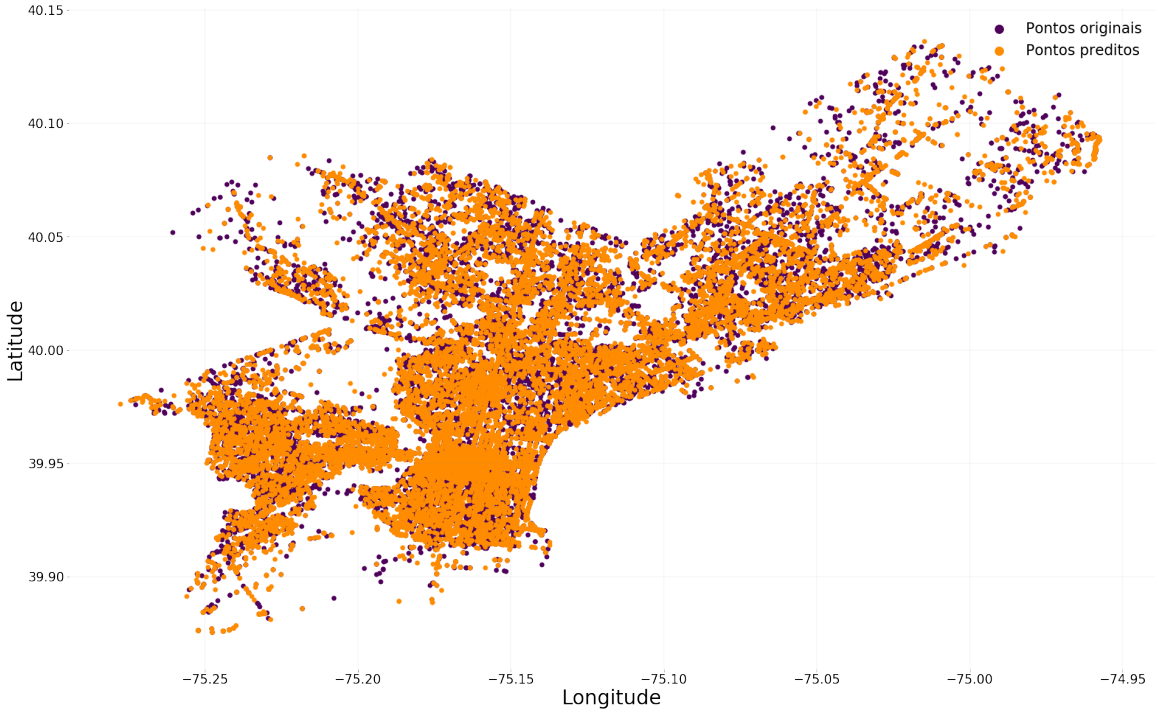
Além dos resíduos, outro resultado encontrado diz respeito aos pontos que foram preditos pelos modelos. A sequência de Figuras 25 a 29 exibe os pontos originais do *dataset* da Filadélfia, Estados Unidos, e sobreposto a eles estão os pontos que foram preditos, esse tipo de gráfico é conhecido como dispersão. O diagrama de dispersão permite comparar dois pontos que estão igualmente, ou quase, dispostos nos eixos X e Y, facilitando dessa forma verificar se os resultados preditos estão análogos aos dados originais.

Figura 25: Pontos preditos para Random Forest Regressor



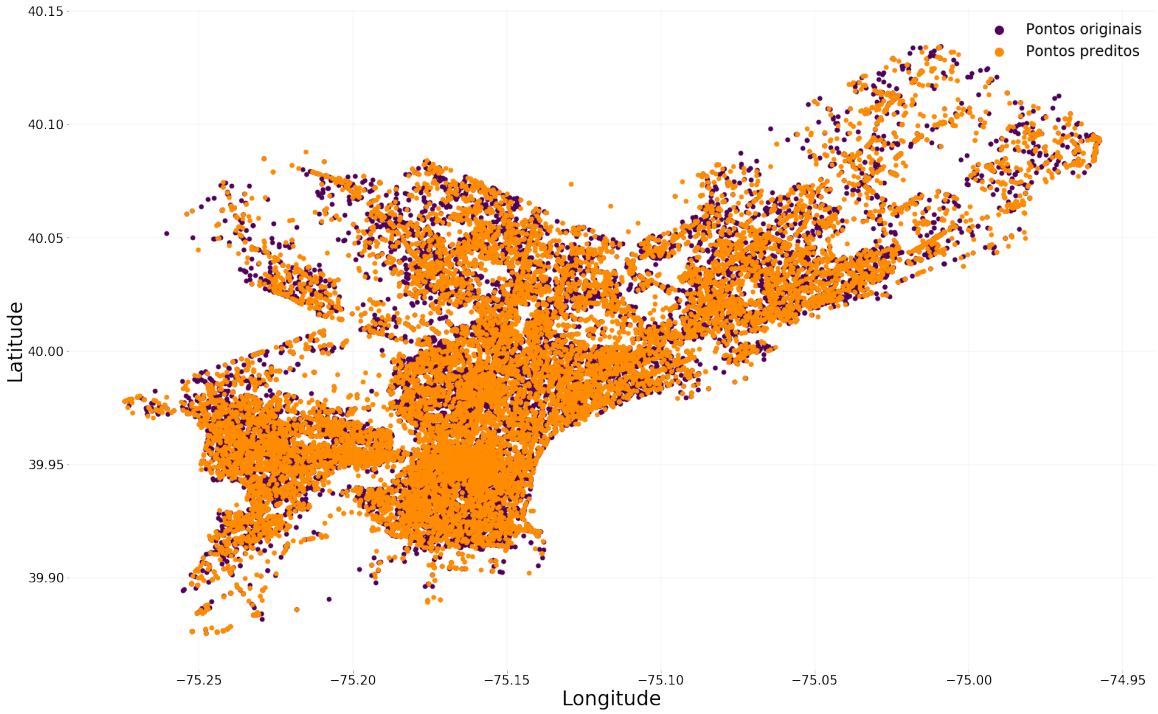
Fonte: Autoria própria.

Figura 26: Pontos preditos para K-Nearest Neighbor Regressor



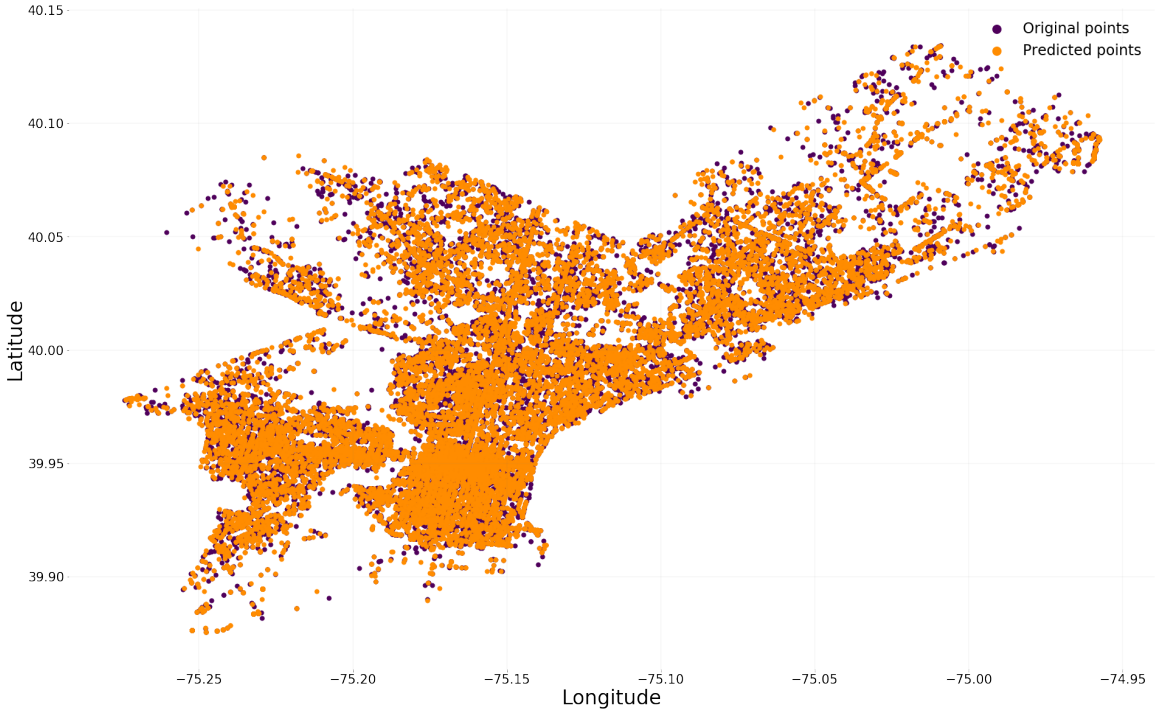
Fonte: Autoria própria.

Figura 27: Pontos preditos para Extra Trees Regressor



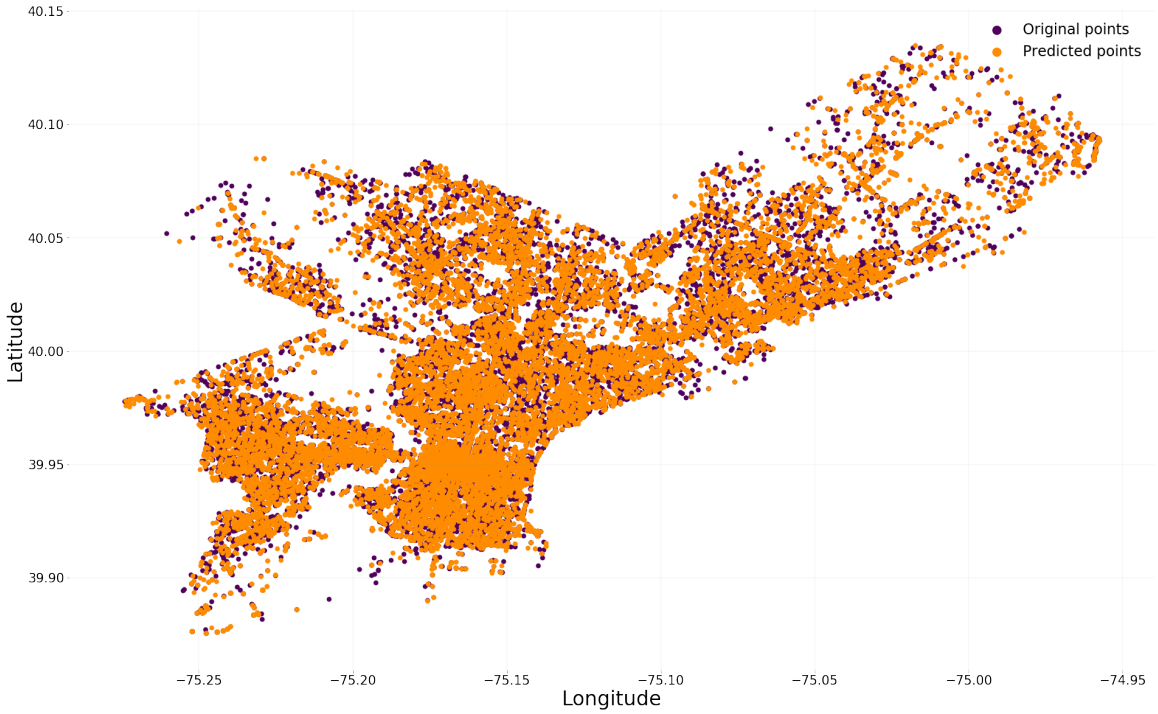
Fonte: Autoria própria.

Figura 28: Pontos preditos para Decision Tree Regressor



Fonte: Autoria própria.

Figura 29: Pontos preditos para Bagging Regressor

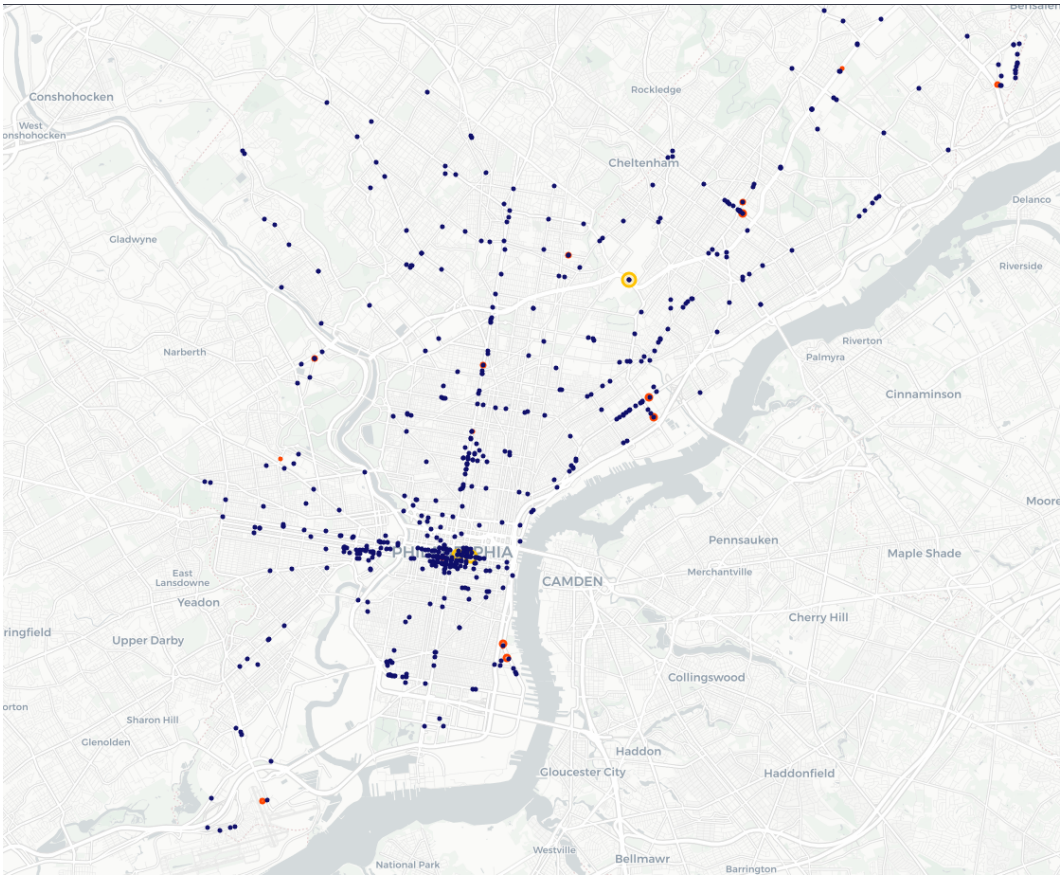


Fonte: Autoria própria.

As Figuras 25 a 29 apresentam os crimes preditos, pontos de cor laranja, e também os pontos originais, pontos de cor azul. Nota-se que as predições realizadas estão de acordo com os dados originais. Alguns métodos, como o K-Nearest Neighbor Regres-

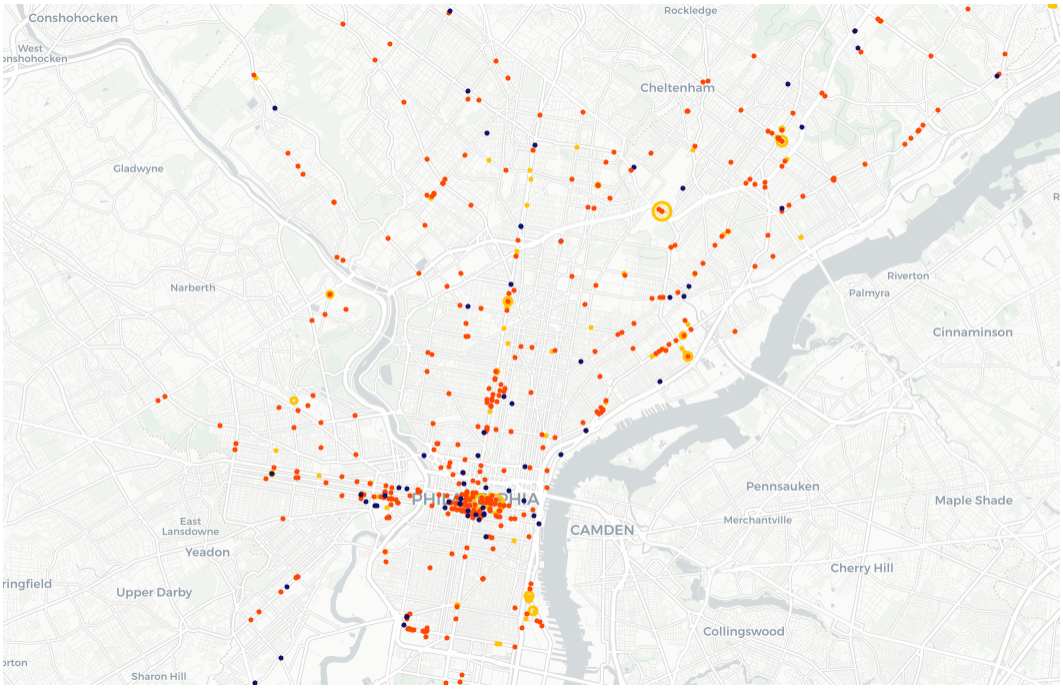
sor, apresentaram dispersões maiores, isso pode ser verificado pela quantidade de pontos originais que são visíveis nos gráficos. Os resultados obtidos com o método *Decision Tree Regressor* apresentaram um maior nível de semelhança ao objetivo proposto inicialmente, a predição de crimes. Esse resultado fica ainda mais visível ao plotar os pontos preditos em um mapa da cidade analisada e compará-los com os dados originais em tal circunstância. As Figuras 31 a 35 exibem esses resultados.

Figura 30: Pontos originais - Mapa da cidade da Philadelphia - Os pontos azuis indicam baixa quantidade de crimes, os pontos vermelhos indicam uma média quantidade de crimes e os pontos amarelos indicam uma alta quantidade de crimes



Fonte: Autoria própria.

Figura 31: Crimes preditos - K-Nearest Neighbor Regressor - Os pontos azuis indicam baixa quantidade de crimes, os pontos vermelhos indicam uma média quantidade de crimes e os pontos amarelos indicam uma alta quantidade de crimes



Fonte: Autoria própria.

Figura 32: Crimes preditos - Random Forest Regressor - Os pontos azuis indicam baixa quantidade de crimes, os pontos vermelhos indicam uma média quantidade de crimes e os pontos amarelos indicam uma alta quantidade de crimes

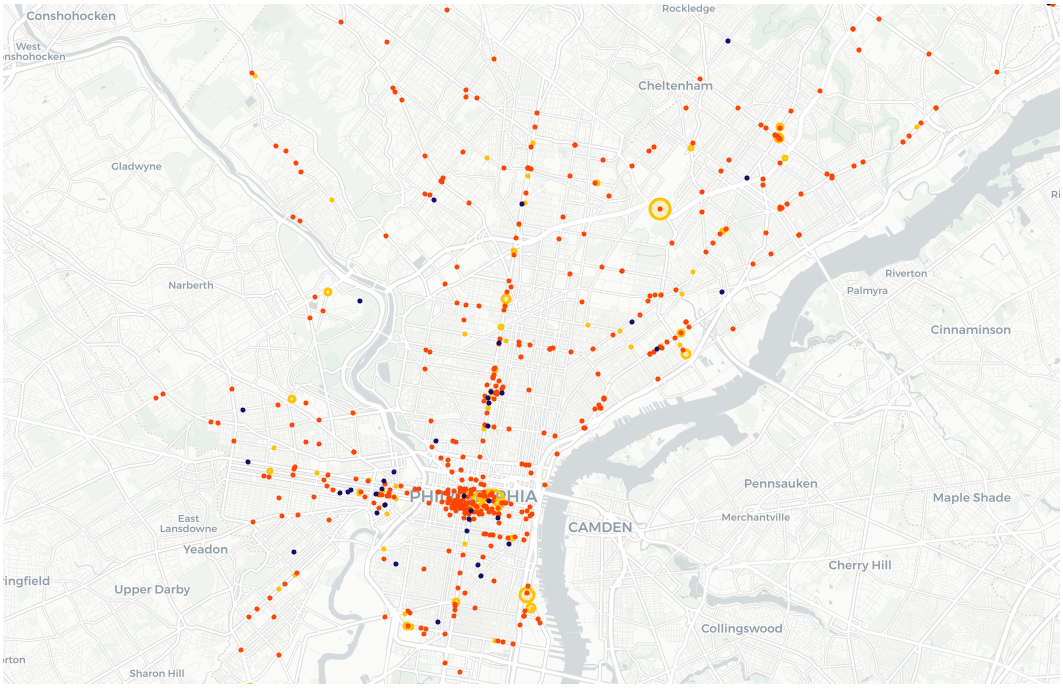
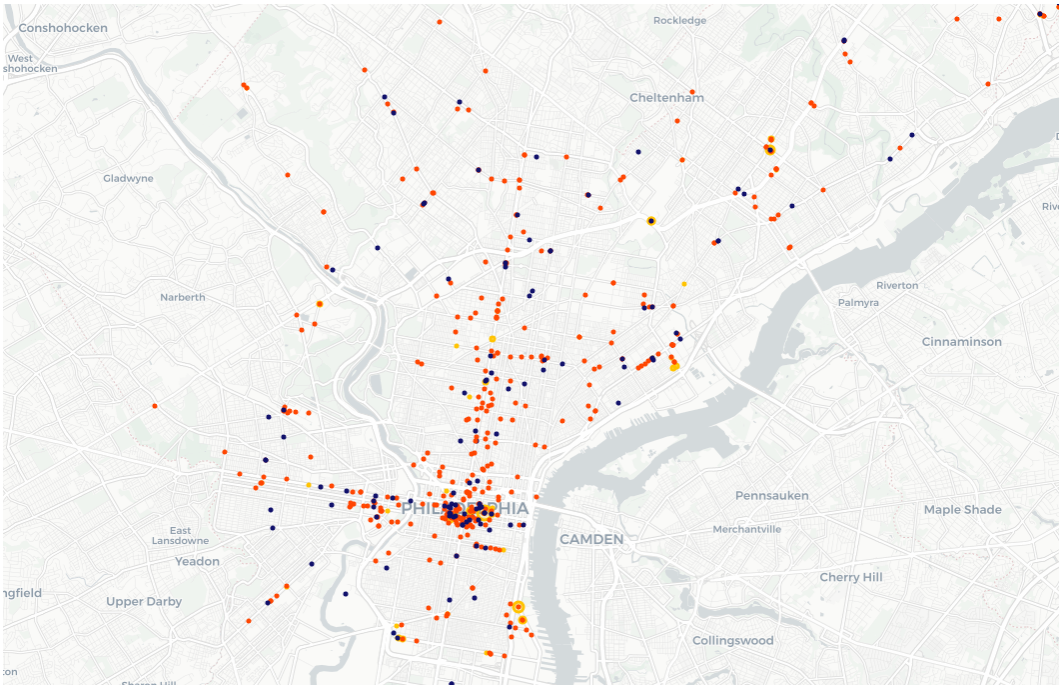
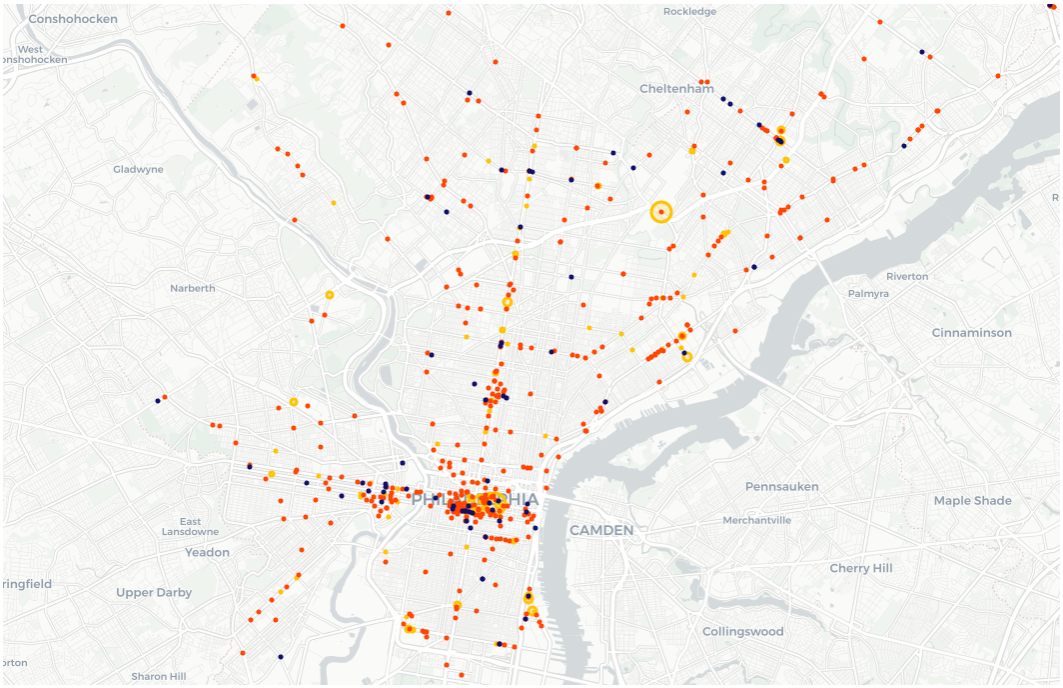


Figura 33: Crimes preditos - Extra Trees Regressor - Os pontos azuis indicam baixa quantidade de crimes, os pontos vermelhos indicam uma média quantidade de crimes e os pontos amarelos indicam uma alta quantidade de crimes



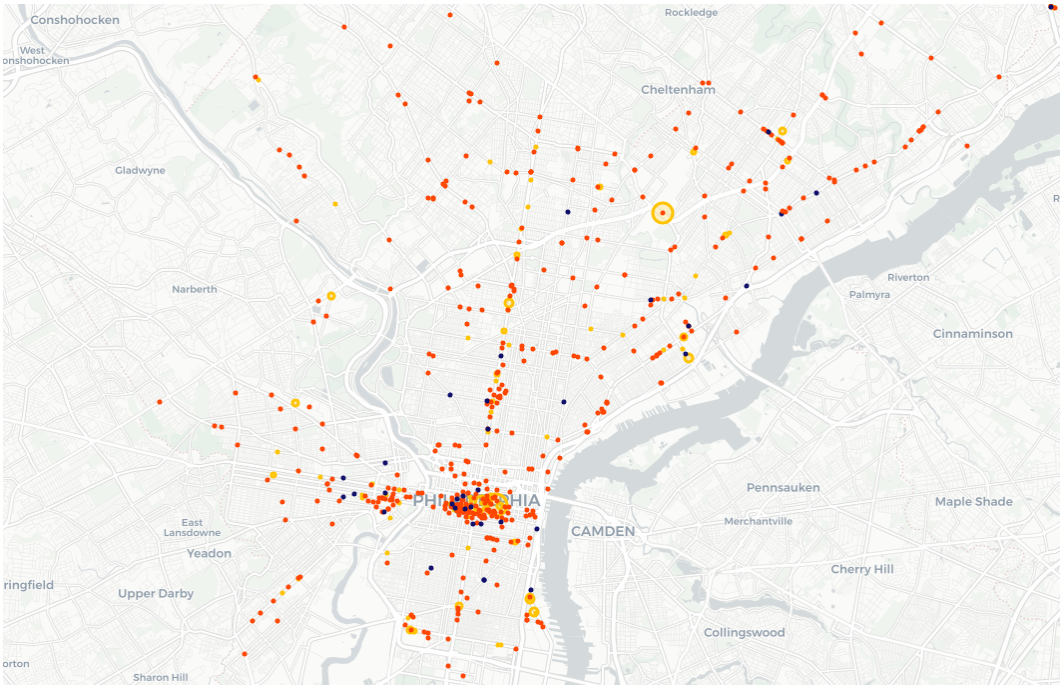
Fonte: Autoria própria.

Figura 34: Crimes preditos - Decision Tree Regressor - Os pontos azuis indicam baixa quantidade de crimes, os pontos vermelhos indicam uma média quantidade de crimes e os pontos amarelos indicam uma alta quantidade de crimes



Fonte: Autoria própria.

Figura 35: Crimes preditos - Bagging Regressor - Os pontos azuis indicam baixa quantidade de crimes, os pontos vermelhos indicam uma média quantidade de crimes e os pontos amarelos indicam uma alta quantidade de crimes



Fonte: Autoria própria.

A Figura 30 apresenta os pontos originais da base de dados da Filadélfia, Estados Unidos, plotados em um mapa. Na plotagem há alguns pontos azuis, vermelhos e amarelos, esses pontos indicam a quantidade de crimes que ocorreu naquela posição. Foram utilizados 600 pontos para a composição dos mapas. Considera-se 20 ocorrências ou mais de crimes em um mesmo ponto como uma alta quantidade. Entre 6 e 19 ocorrências em um mesmo local, considera-se como média e menos que 6, é classificada como baixa. Os pontos azuis indicam baixa quantidade de crimes, os pontos vermelhos indicam uma média quantidade de crimes e os pontos amarelos indicam que houve uma alta quantidade de crimes naquele local.

A Figura 31 apresenta os pontos que foram preditos para o método *K-Nearest Neighbor Regressor*. Em algumas áreas há incidência de grande quantidade de crimes ocorridos, diferentemente dos pontos originais em que há poucos casos. A Figura 32 exibe os resultados utilizando o método *Random Forest Regressor*, as predições realizadas se assemelham aos do método anterior, houve também uma predominância de áreas com alta quantidade de crimes ocorridos. A Figura 33 mostra os dados preditos para o regressor *Extra Trees*, os pontos já se diferem em relação aos métodos anteriores, os crimes em algumas áreas não foram preditos corretamente.

Já na Figura 34, os pontos preditos para o método *Decision Tree* são apresentados, novamente há a incidência de grande quantidade de crimes em algumas áreas. O modelo foi capaz de realizar a predição do ponto amarelo que está presente na Figura

30, tendo uma maior semelhança com os dados originais. Na Figura 35 é apresentado os pontos preditos para o método regressor *Bagging*, foi realizado também a predição do ponto amarelo, assim como no método anterior, presente na Figura 30. Houve também uma incidência alta de crimes de média quantidade ocorridos que foram preditos pelo regressor.

4.2 BASE DE DADOS FORTALEZA

Nessa seção serão apresentados os resultados obtidos através dos experimentos em que foram utilizados a base de dados da cidade de Fortaleza. Para o *dataset* utilizado, foram realizados dois experimentos que foram subdivididos em mais dois experimentos, totalizando quatro experimentos. Cada um deles será explicado nas subseções seguintes.

4.2.1 EXPERIMENTO 1 - CATEGORIA 1

O primeiro resultado apresentado diz respeito a primeira parte, nomeada como categoria 1 do experimento 1. Nesse experimento foi considerado que os crimes na cidade de Fortaleza aconteciam em vários pontos de uma rua, assim como nos dados originais, ou seja, em uma rua poderiam haver registros de crimes que ocorreram no início, no meio ou no fim dela. Outro ponto importante nesse experimento foi considerar que o modelo iria realizar a predição de duas saídas, latitude e longitude.

Para cada um dos métodos regressores, discutido na seção de materiais e métodos, foi gerado um modelo de aprendizado responsável pelas predições. A Tabela 8 exibe os parâmetros que foram utilizados e os erros que foram obtidos para cada uma das técnicas.

Tabela 8: Resultado experimento 1 - categoria 1

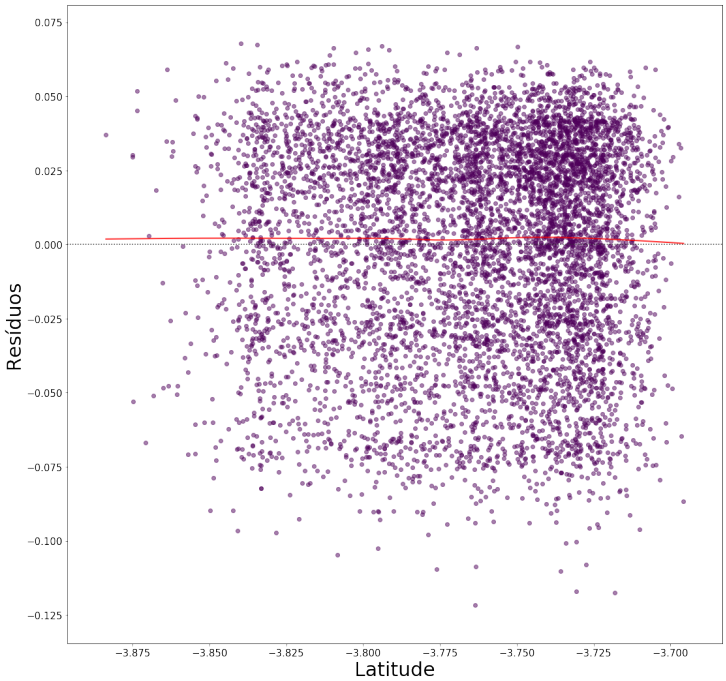
	Parâmetros			**Erro**	
Método	**Número de estimadores**	**Profundidade máxima**	**Número de vizinhos**	**MSE**	**RMSE**
K-Nearest Neighbor Regressor	-	-	1	0.00293	0.05420
Random Forest Regressor	**1**	**-**	**-**	**0.00275**	**0.05251**
Extra Trees Regressor	1	24	-	0.00296	0.05441
Decision Tree Regressor	-	25	-	0.00287	0.05365
Bagging Regressor	1	-	-	0.00292	0.05410

Fonte: Autoria própria.

Na Tabela 8 os menores erros de MSE e RMSE pertencem ao regressor *Random Forest*. Essa técnica utilizou apenas um parâmetro para ser treinada, que foi o *Number of Estimators*. Os outros regressores que utilizaram apenas um parâmetro de treinamento, também obtiveram um erro menor. A quantidade de parâmetros não é necessariamente um indicativo que o modelo apresentará melhores resultados, isso pode ser visto com os erros apresentados pelo *K-Nearest Neighbor Regressor*, o treinamento foi realizado com

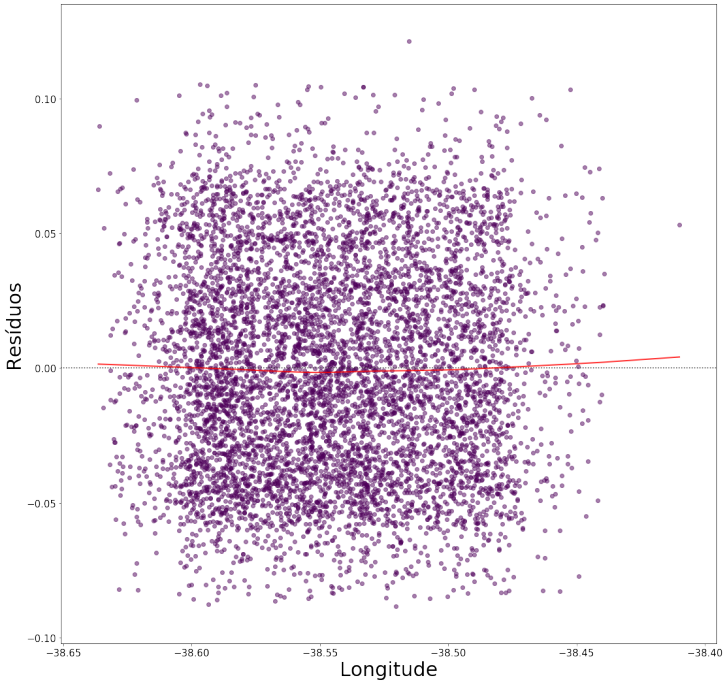
a apenas um parâmetro mas os erros obtidos foi um dos maiores para esse experimento. Assim como nos dados da cidade da Filadélfia, Estados Unidos, para a base de dados de Fortaleza também foram gerados os resultados referentes ao resíduos encontrados. As Figuras 36 a 45 exibem esses resultados.

Figura 36: Resíduos para Latitude de K-Nearest Neighbor Regressor



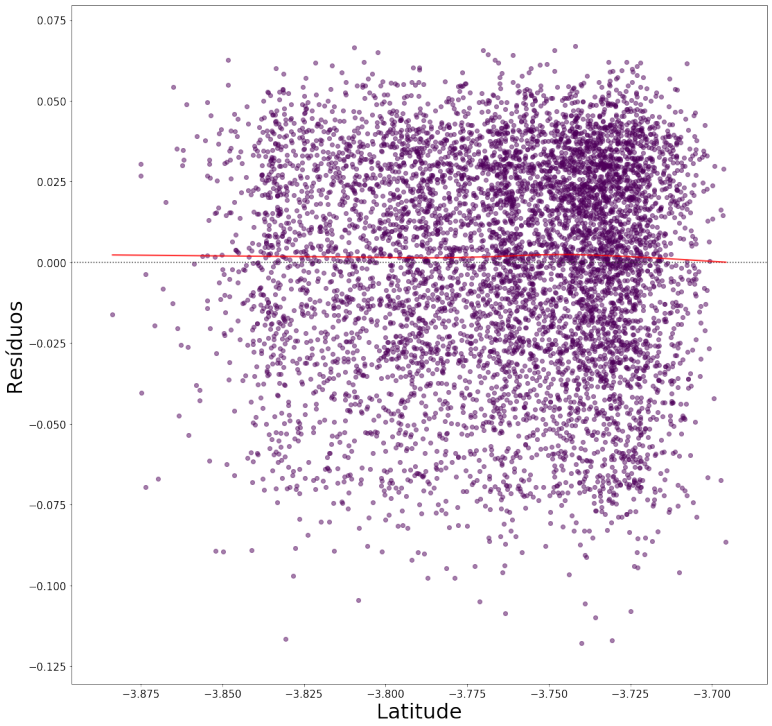
Fonte: Autoria própria.

Figura 37: Resíduos para Longitude de K-Nearest Neighbor Regressor



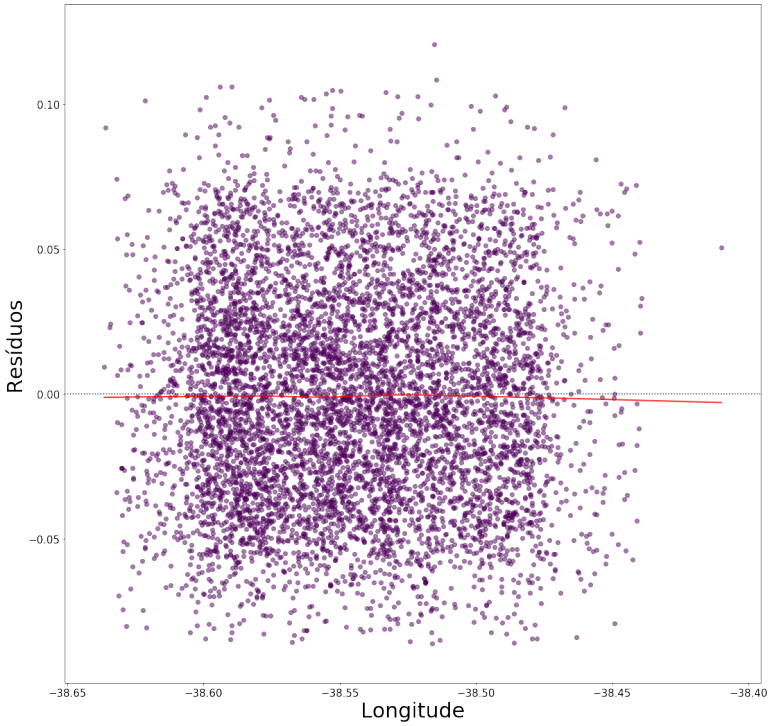
Fonte: Autoria própria.

Figura 38: Resíduos para Latitude de Random Forest Regressor



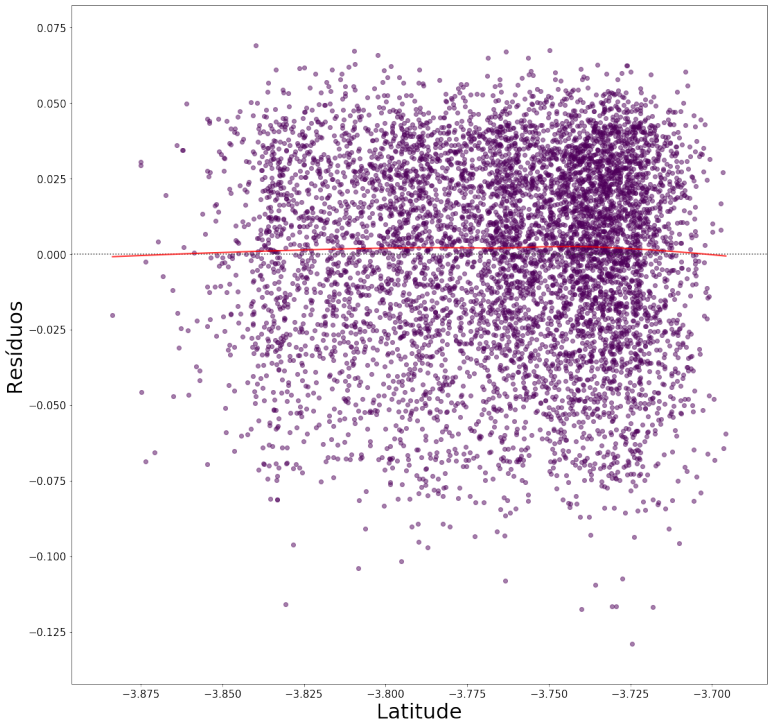
Fonte: Autoria própria.

Figura 39: Resíduos para Longitude de Random Forest Regressor



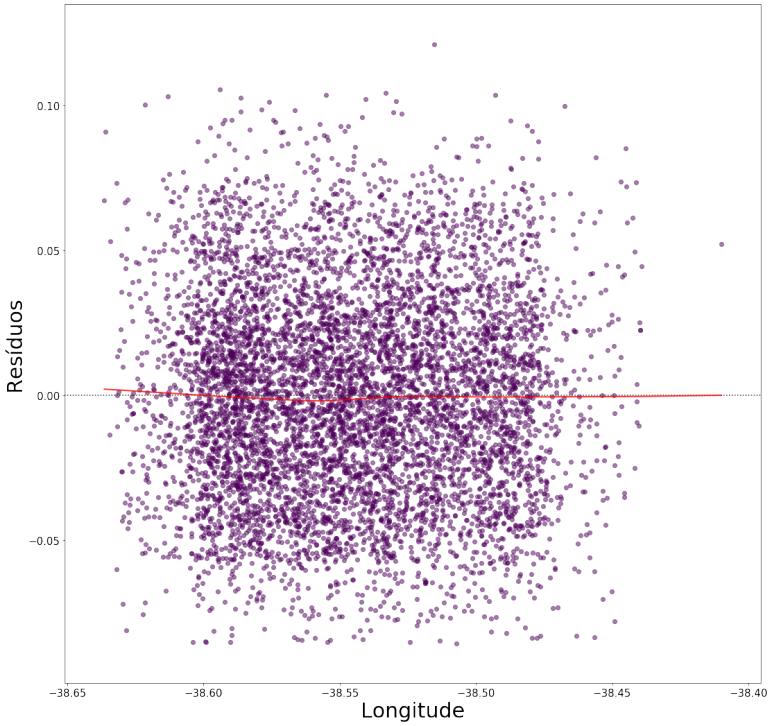
Fonte: Autoria própria.

Figura 40: Resíduos para Latitude de Extra Trees Regressor



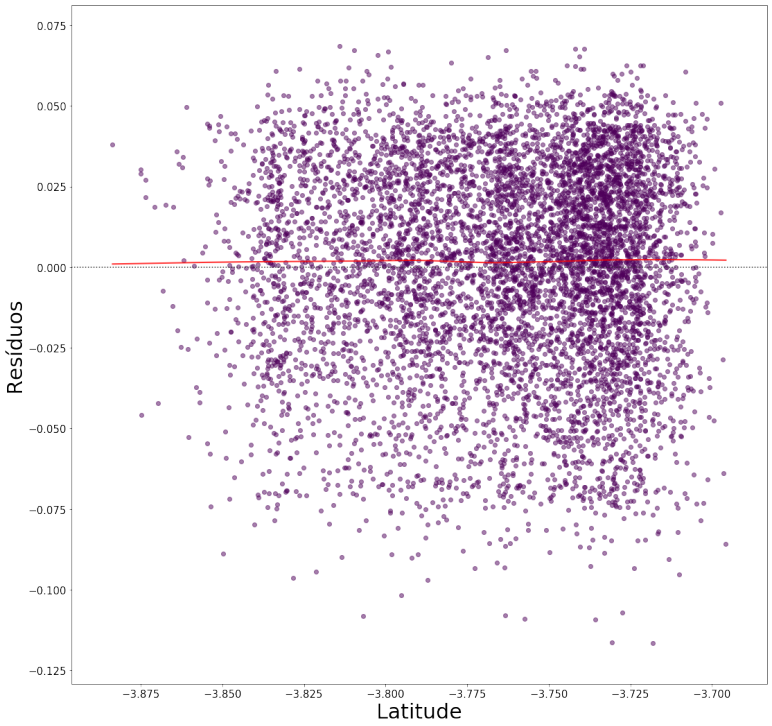
Fonte: Autoria própria.

Figura 41: Resíduos para Longitude de Extra Trees Regressor



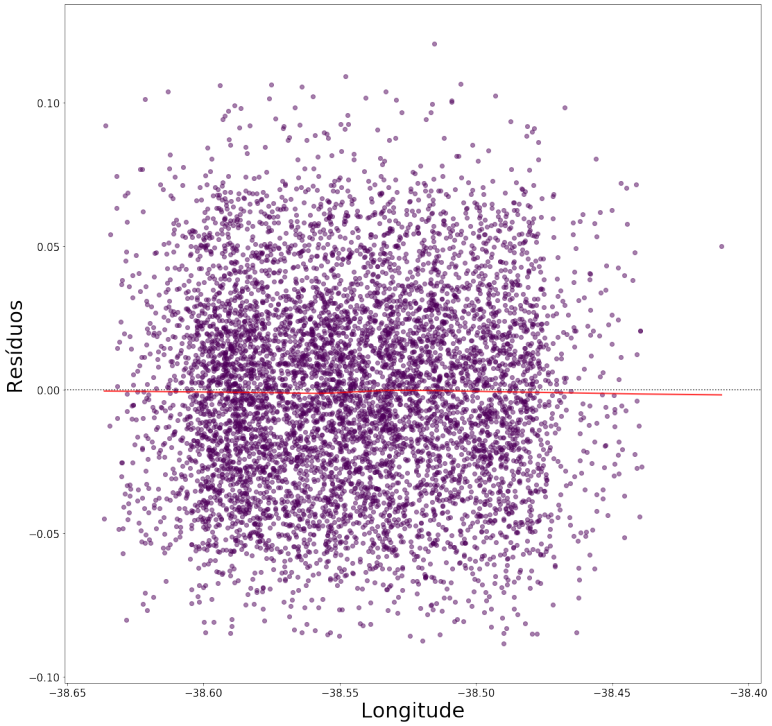
Fonte: Autoria própria.

Figura 42: Resíduos para Latitude de Decision Tree Regressor



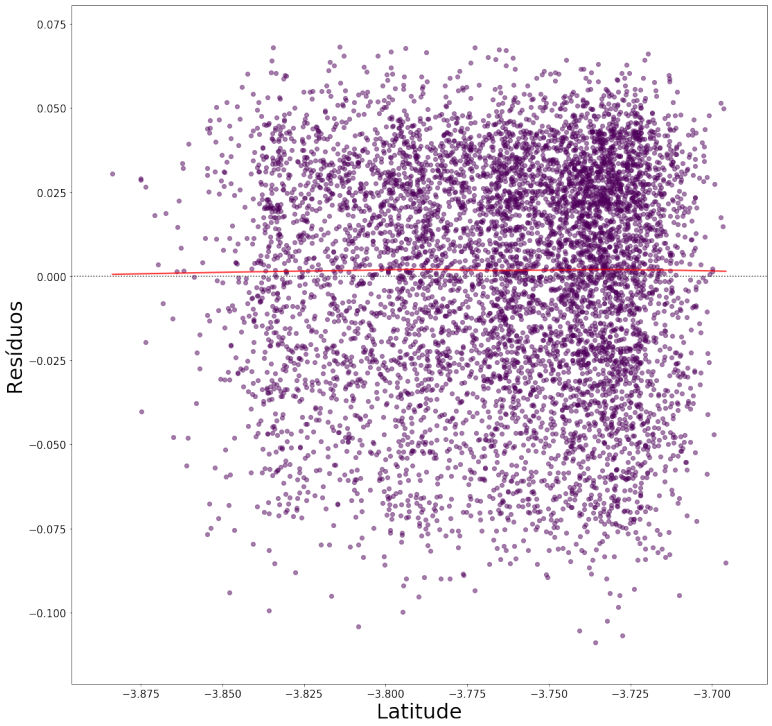
Fonte: Autoria própria.

Figura 43: Resíduos para Longitude de Decision Tree Regressor



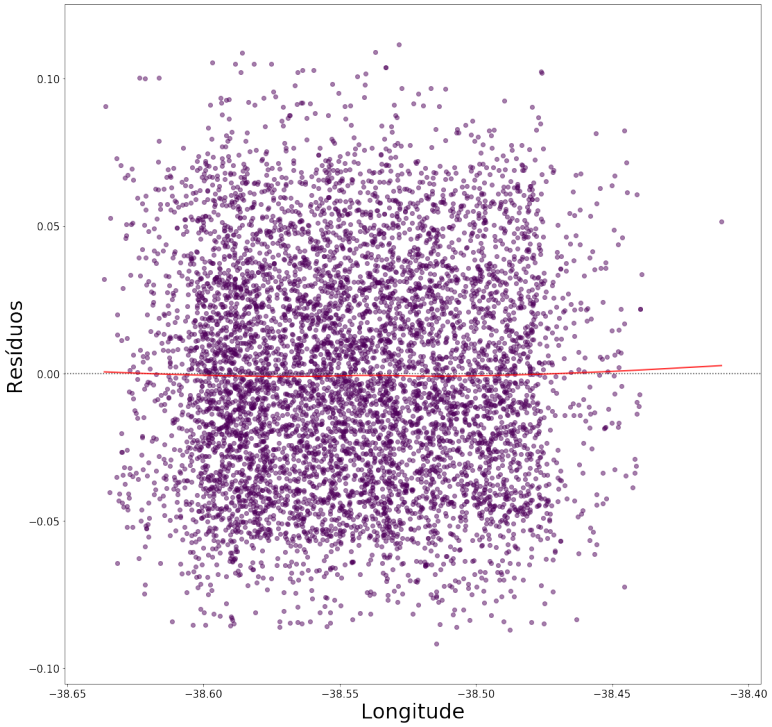
Fonte: Autoria própria.

Figura 44: Resíduos para Latitude de Bagging Regressor



Fonte: Autoria própria.

Figura 45: Resíduos para Longitude de Bagging Regressor

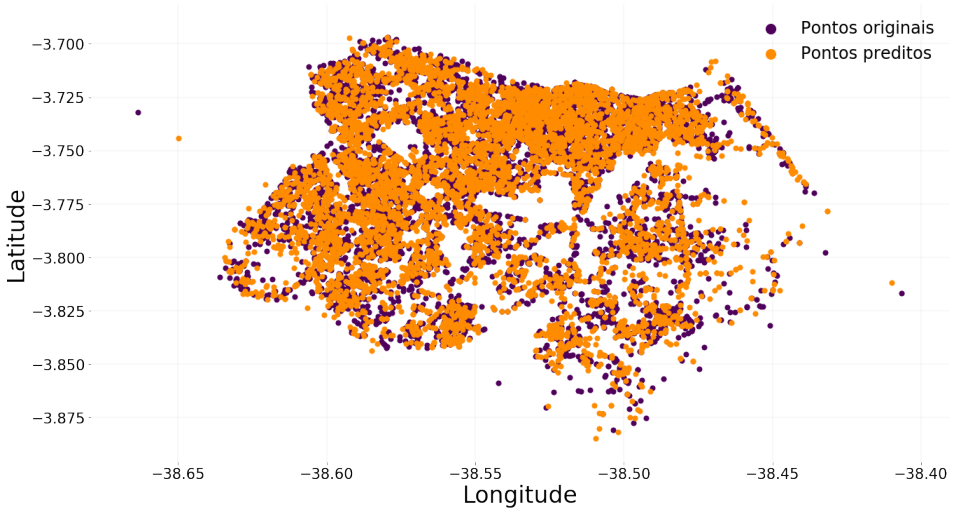


Fonte: Autoria própria.

Os erros de MSE e RMSE apresentados na Tabela 8 estão próximos um do outro, com diferenças mínimas. Essas diferenças são notáveis nos gráficos residuais em que as linhas vermelhas estão próximas ou estão sobre a posição 0 no eixo X, que é o

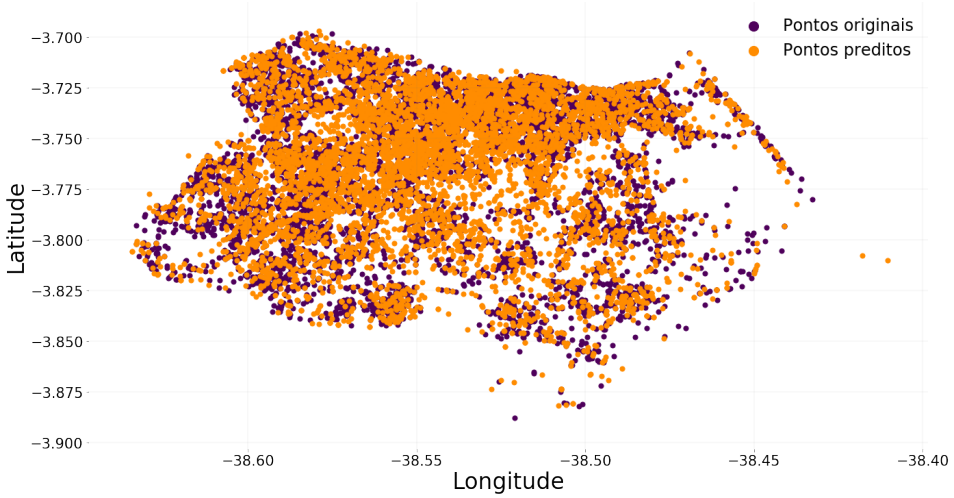
resultado mais favorável para o cenário de ótimas predições. Comparando cada um dos resíduos, diferença entre os valores esperados e os valores preditos, encontrados de forma individual, latitude e longitude, é notável que para métodos como *Extra Trees* e *Bagging Regressor* a linha vermelha, que exibe a diferença entre os pontos originais e os pontos preditos, está mais próxima de 0 para os resíduos de Longitude, e isto ocorre pois os valores residuais são calculados de forma individual, diferentemente dos valores de MSE e RMSE que calculam os valores do conjunto latitude e longitude para a geração do erro. O valor da latitude do método *Random Forest*, que detém o menor erro, possui uma curva mais acentuada para valores entre -3.775 e -3.725. Foi gerado também o resultado referente aos pontos preditos em um gráfico de dispersão, assim como foi feito com os dados da cidade da Filadélfia, Estados Unidos. As Figuras 46 a 50 apresentam esses resultados.

Figura 46: Pontos preditos para K-Nearest Neighbor Regressor



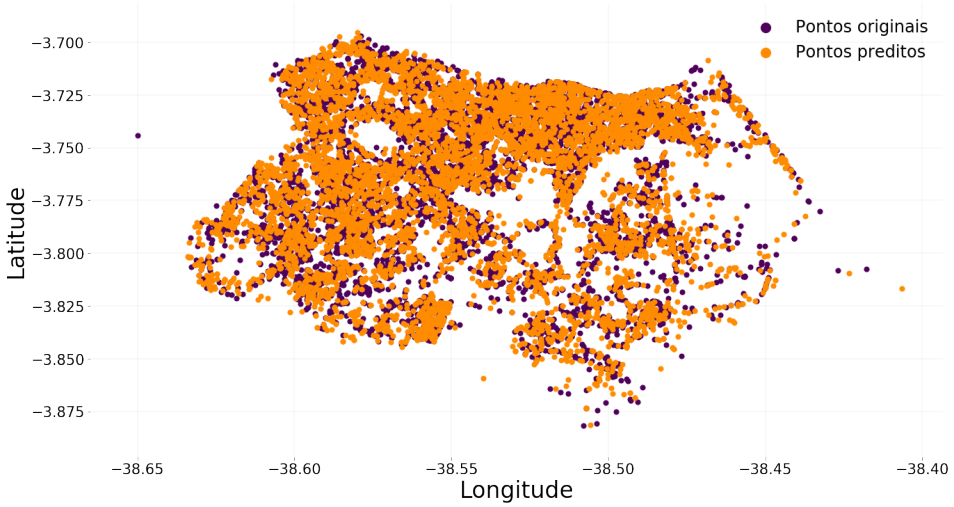
Fonte: Autoria própria.

Figura 47: Pontos preditos para Random Forest Regressor



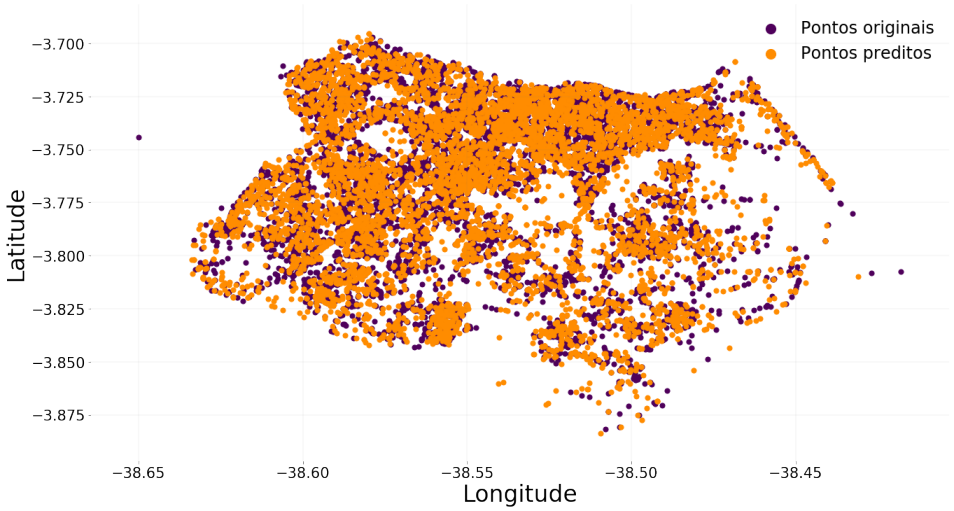
Fonte: Autoria própria.

Figura 48: Pontos preditos para Extra Trees Regressor



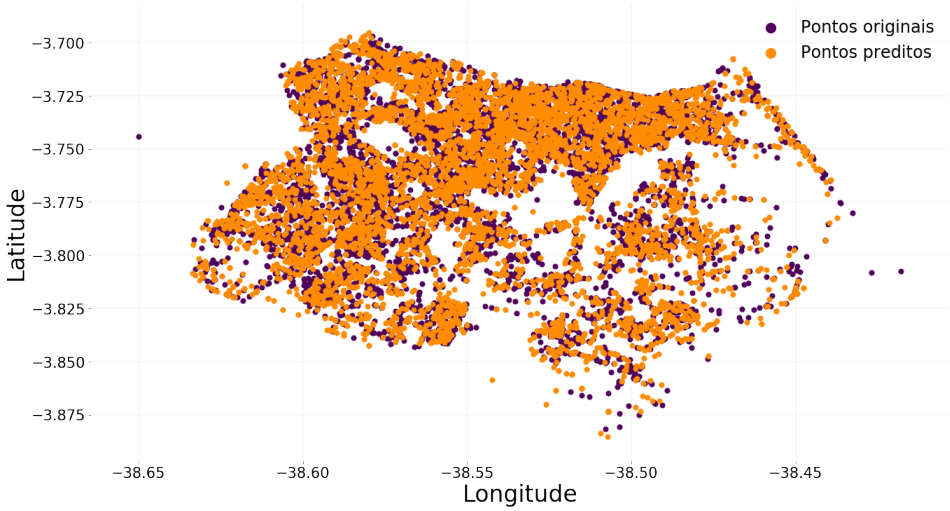
Fonte: Autoria própria.

Figura 49: Pontos preditos para Decision Tree Regressor



Fonte: Autoria própria.

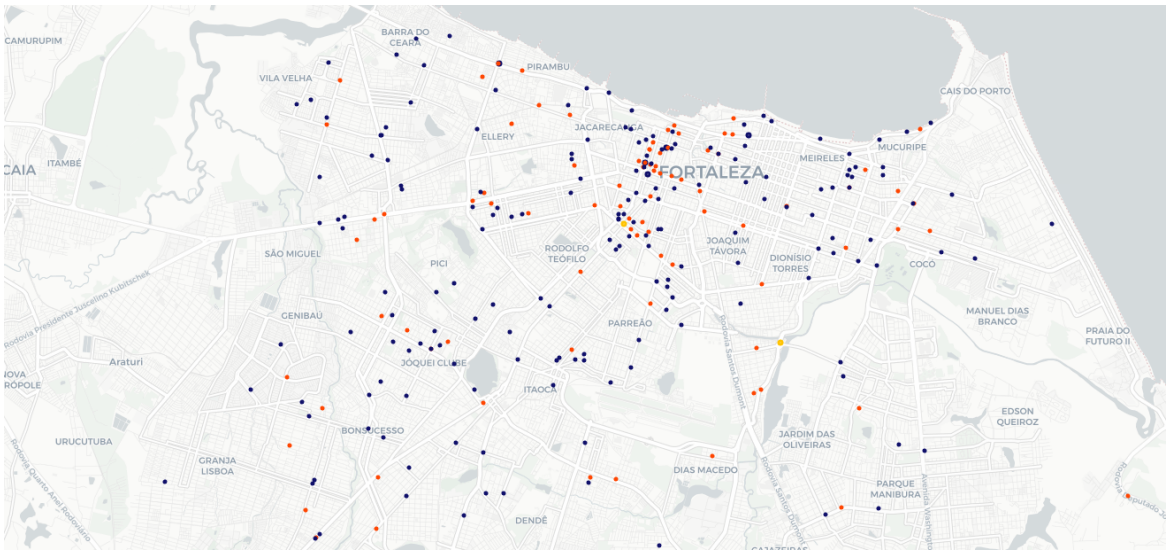
Figura 50: Pontos preditos para Bagging Regressor



Fonte: Autoria própria.

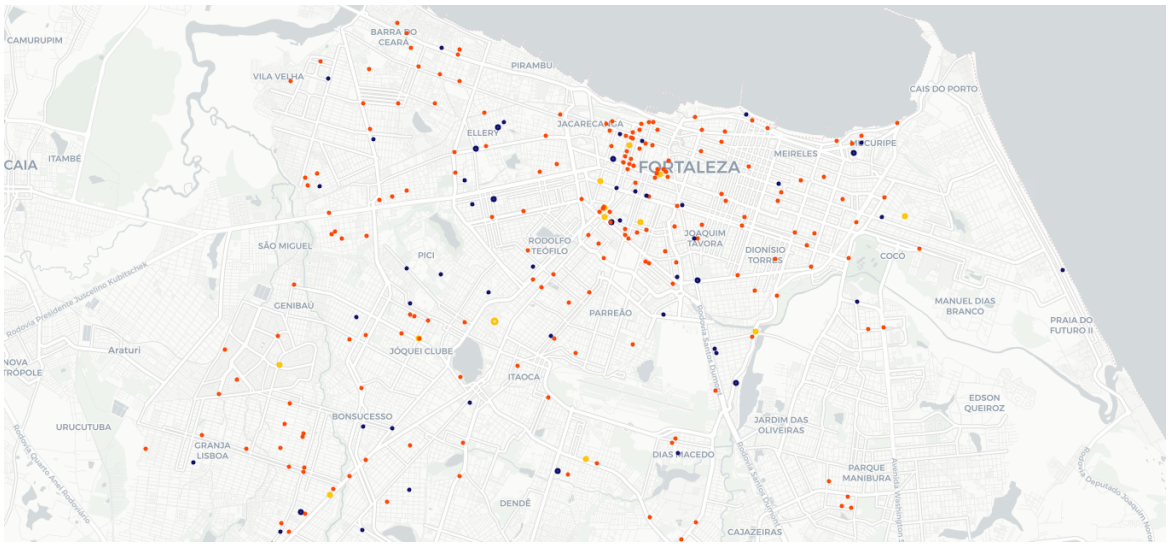
As Figuras 46 a 50 exibem os pontos preditos juntamente com os pontos originais, permitindo uma melhor comparação dos resultados obtidos. Assim como os resultados apresentados na Tabela 8, os gráficos de dispersão também apresentam resultados bastante semelhantes. Outro resultado importante, foi a plotagem desses pontos, de dispersão, em uma mapa da cidade de Fortaleza. Esse resultado é apresentado nas Figuras 51 a 56.

Figura 51: Pontos originais - Mapa da cidade de Fortaleza



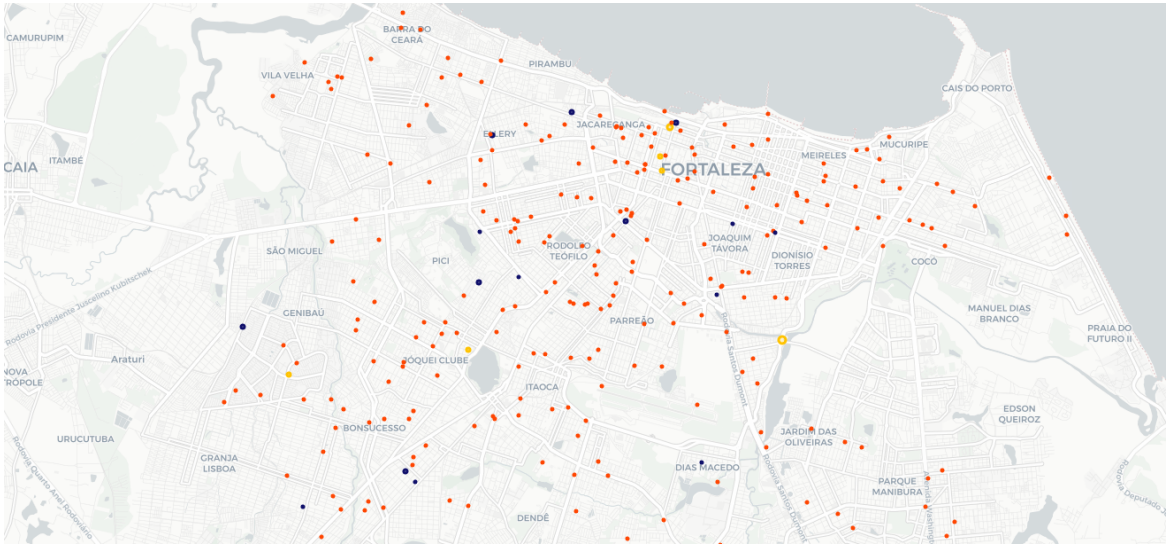
Fonte: Autoria própria.

Figura 52: Crimes preditos - K-Nearest Neighbor Regressor



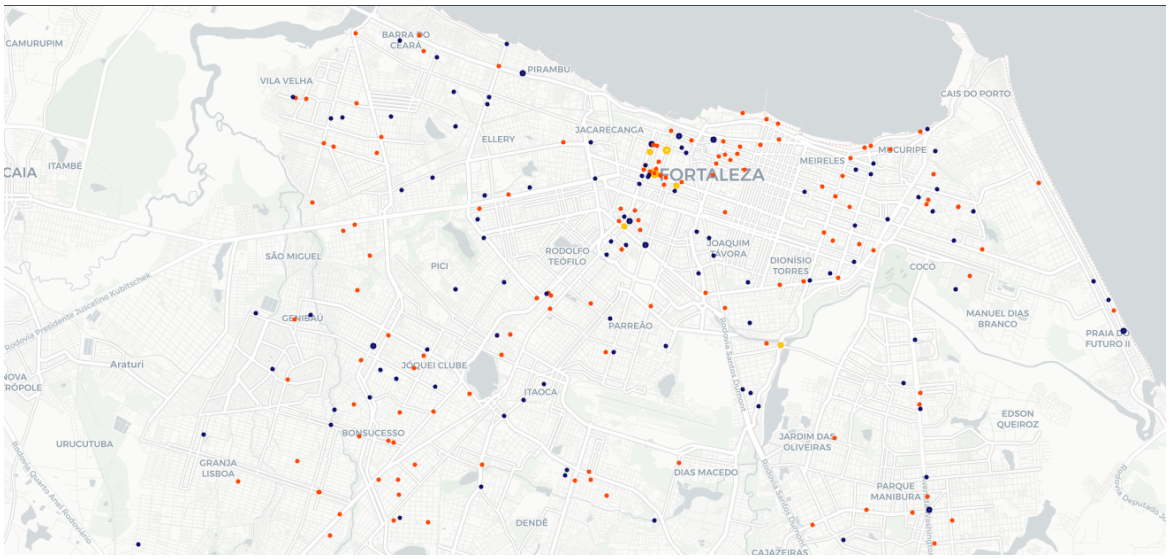
Fonte: Autoria própria.

Figura 53: Crimes preditos - Random Forest Regressor



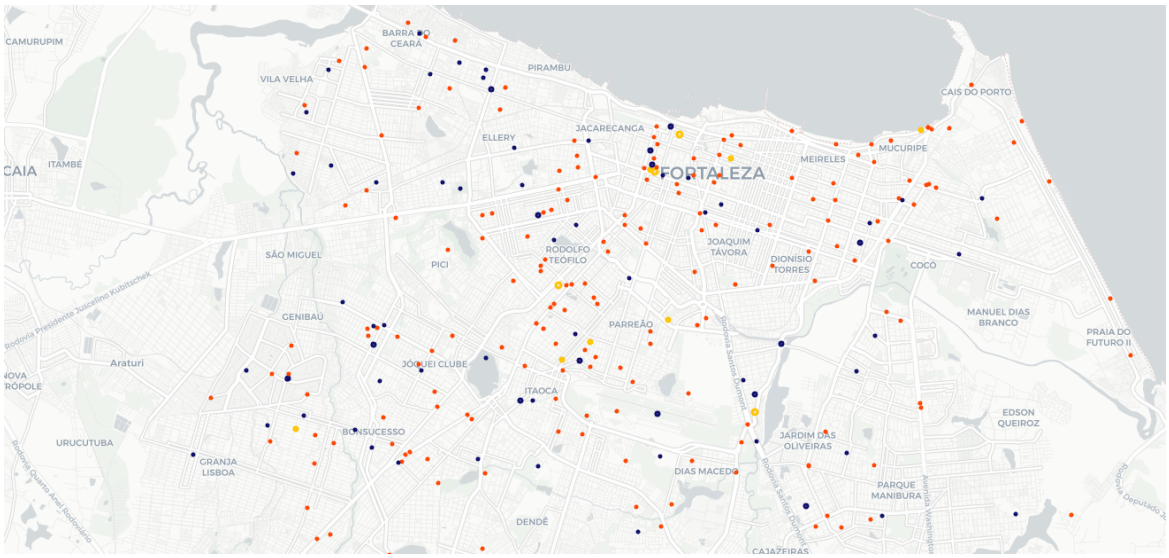
Fonte: Autoria própria.

Figura 54: Crimes preditos - Extra Trees Regressor



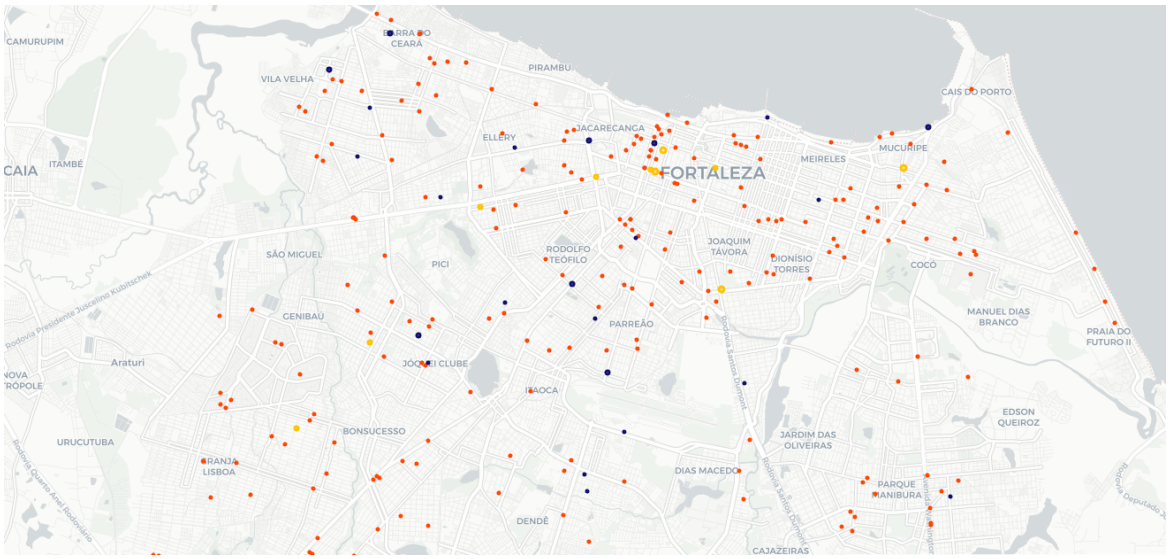
Fonte: Autoria própria.

Figura 55: Crimes preditos - Decision Tree Regressor



Fonte: Autoria própria.

Figura 56: Crimes preditos - Bagging Regressor



Fonte: Autoria própria.

Além de permitir a comparação entre os pontos preditos e os pontos originais, os mapas das figuras acima permitem visualizar onde ocorre a maior quantidade de crimes na cidade de Fortaleza. Os pontos preditos também estão divididos em frequência de ocorrência, sendo o azul para baixa quantidade de crimes, o vermelho para média quantidade de crimes, e o amarelo para a alta quantidade de crimes naquela área. Considera-se 5 ocorrências ou mais de crimes em um mesmo ponto como uma alta quantidade, entre 3 e 5 ocorrências em um mesmo local, considera-se como média e menos que 3, é classificada como baixa. Comparando os valores preditos, percebe-se que os mapas gerados para os métodos de *Decision Tree* e *Extra Trees* estão se aproximando mais dos valores presentes no mapa original da Figura 51.

4.2.2 EXPERIMENTO 1 - CATEGORIA 2

Na segunda parte do experimento 1, chamada também de categoria 2, considera-se que os crimes ainda ocorrem em diferentes partes de uma rua, latitude e longitude possuem valos distribuídos. Entretanto, o modelo agora só gera um valor de saída, um *hash*. Esse valor *hash* representa os valores de latitude e longitude. A Tabela 9 apresenta os valores utilizados para treinar os métodos e também os valores de erros obtidos para cada um deles.

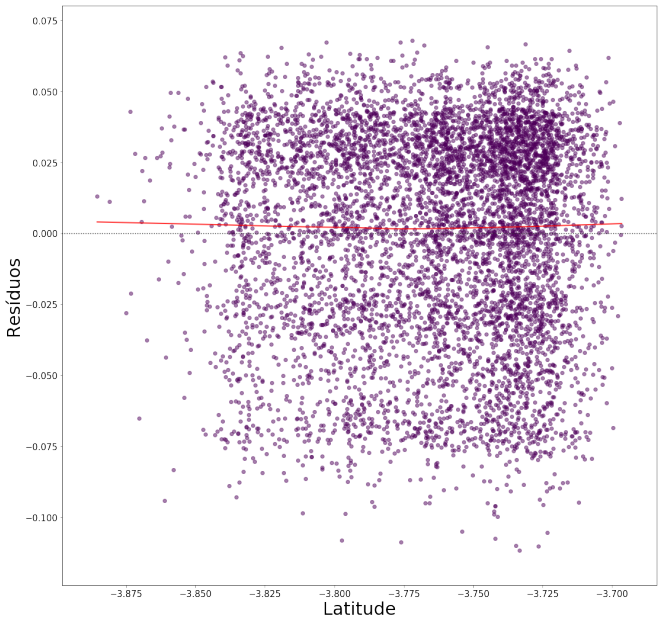
Tabela 9: Resultado experimento 1 - categoria 2

Método	Parâmetros			Erro	
	Número de estimadores	**Profundidade máxima**	**Número de vizinhos**	**MSE**	**RMSE**
K-Nearest Neighbor Regressor	**-**	**-**	**1**	**305190698.80**	**17469.70**
Random ForestRegressor	1	-	-	308368892.18	17560.43
Extra Trees Regressor	1	37	-	319818890.82	17883.48
Decision Tree Regressor	-	30	-	312382628.73	17674.34
Bagging Regressor	1	-	-	312914903.97	17689.40

Fonte: Autoria própria.

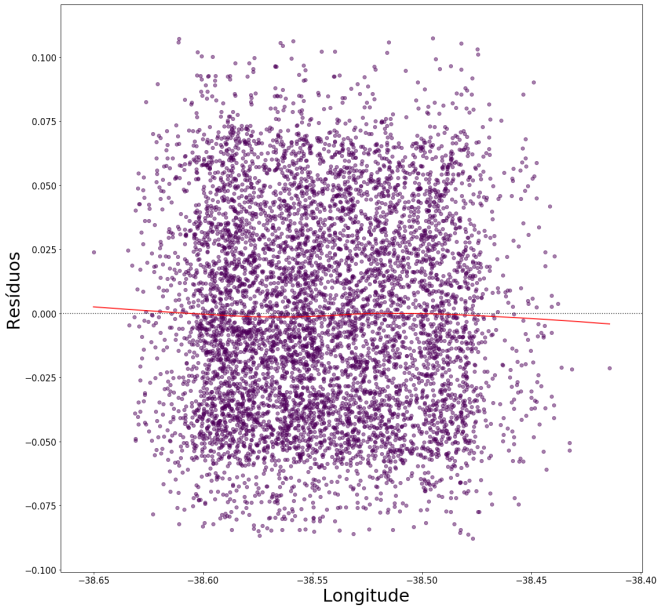
Nesse experimento, o método *K-Nearest Neighbor Regressor* apresenta um menor erro de MSE e RMSE. Como saída, o valor *hash*, está em uma escala diferente do que é apresentado por valores de latitude e longitude, e o valor de erro também é apresentado em uma escala diferente. O método *Random Forest* apresentou o segundo menor erro, diferentemente do experimento anterior no qual apresentou o menor erro. O método *Extra Trees* acabou apresentando o maior erro entre os métodos, assim como no experimento anterior. Os próximos resultados dizem respeito aos valores residuais calculados para esse experimento. As Figuras 57 a 66 exibem os resíduos.

Figura 57: Resíduos para Latitude de K-Nearest Neighbor Regressor



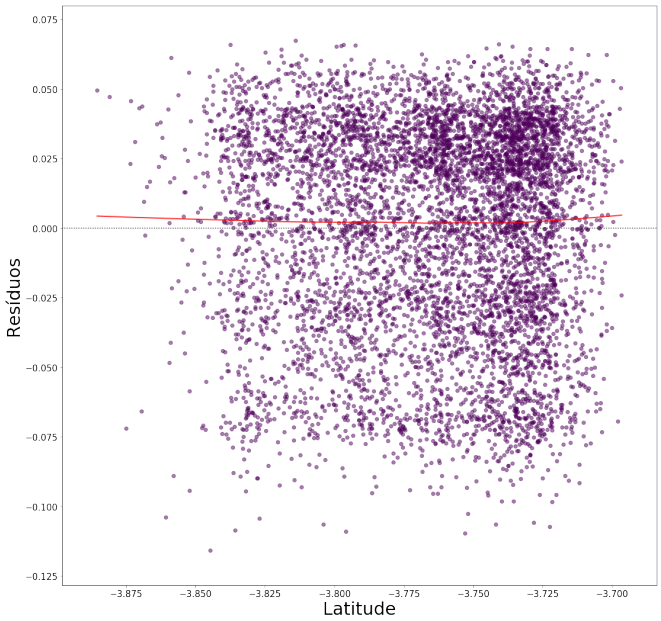
Fonte: Autoria própria.

Figura 58: Resíduos para Longitude de K-Nearest Neighbor Regressor



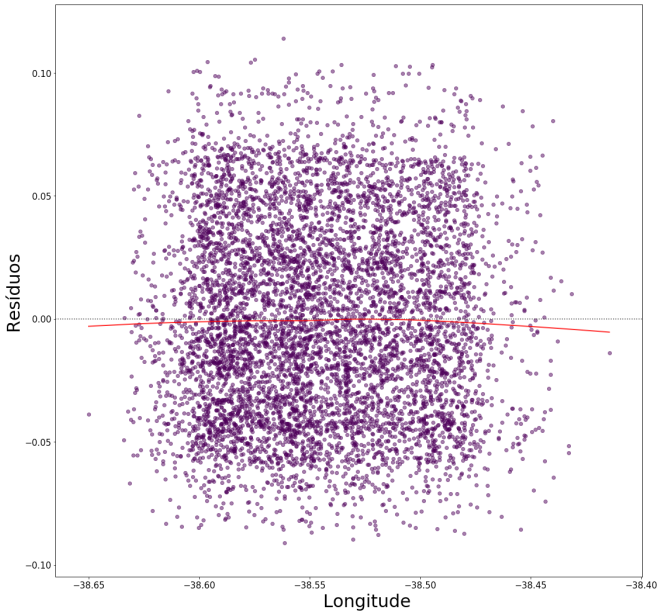
Fonte: Autoria própria.

Figura 59: Resíduos para Latitude de Random Forest Regressor



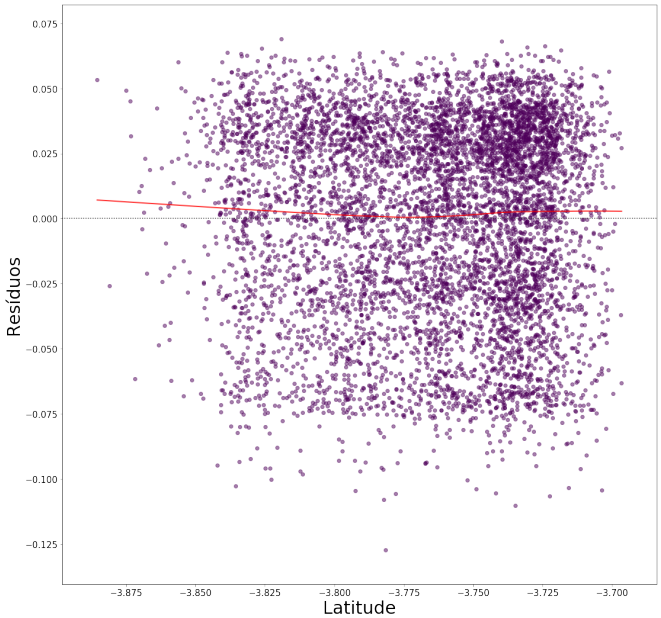
Fonte: Autoria própria.

Figura 60: Resíduos para Longitude de Random Forest Regressor



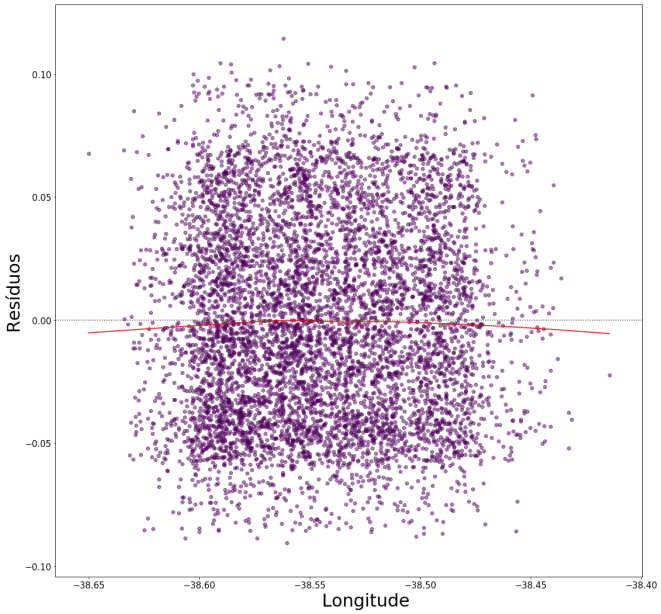
Fonte: Autoria própria.

Figura 61: Resíduos para Latitude de Extra Trees Regressor



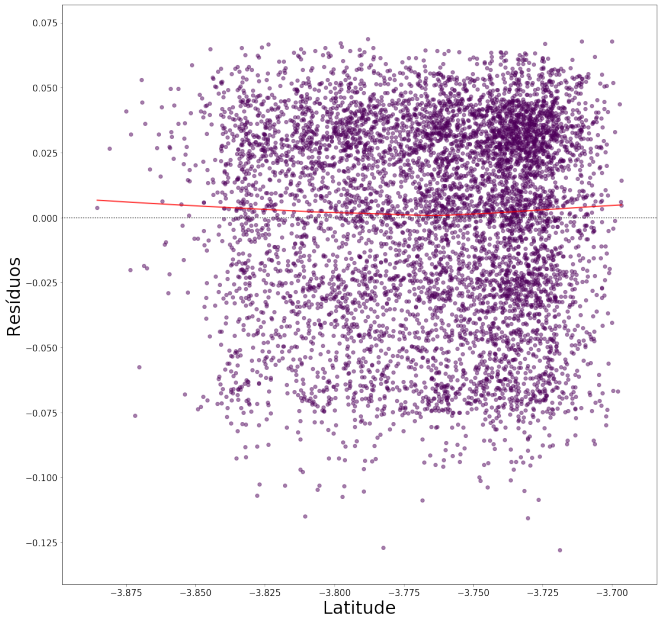
Fonte: Autoria própria.

Figura 62: Resíduos para Longitude de Extra Trees Regressor



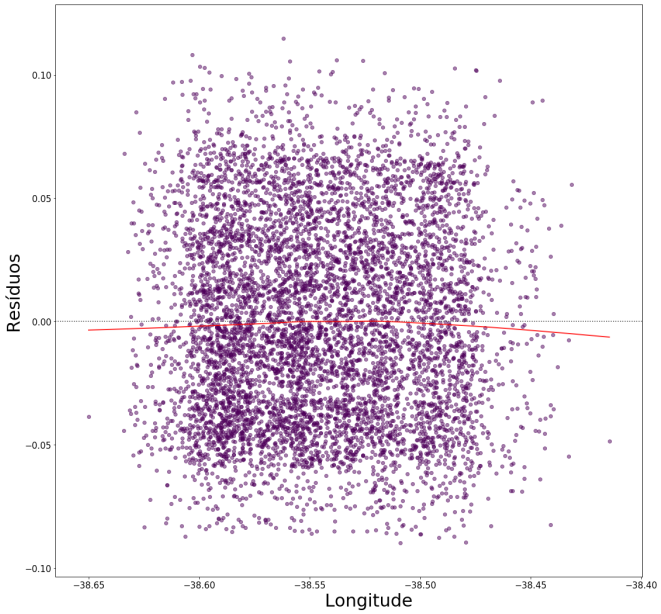
Fonte: Autoria própria.

Figura 63: Resíduos para Latitude de Decision Tree Regressor



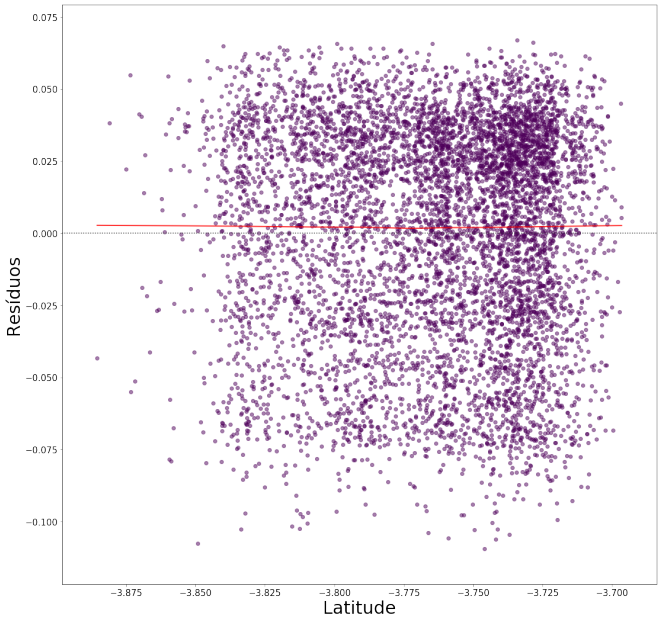
Fonte: Autoria própria.

Figura 64: Resíduos para Longitude de Decision Tree Regressor



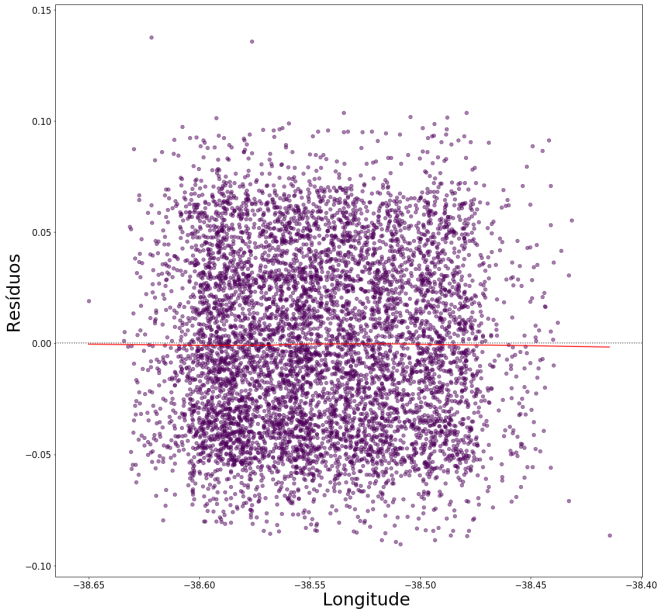
Fonte: Autoria própria.

Figura 65: Resíduos para Latitude de Bagging Regressor



Fonte: Autoria própria.

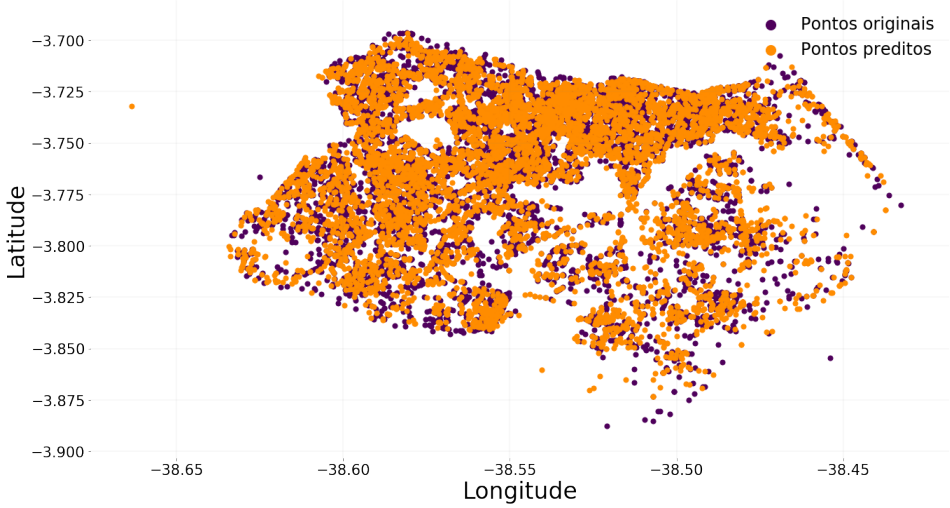
Figura 66: Resíduos para Longitude de Bagging Regressor



Fonte: Autoria própria.

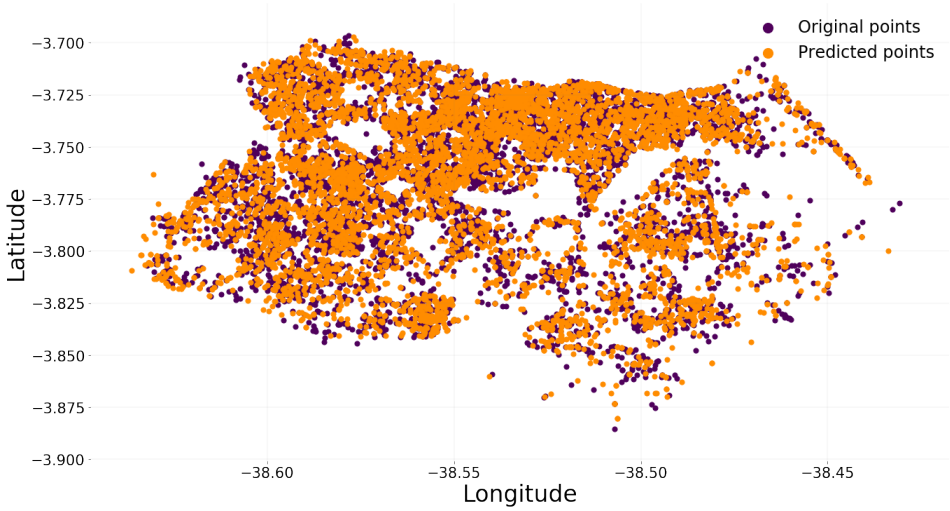
O resíduo, diferença entre os valores reais e os valores preditos, calculado para o valor de latitude do método *K-Nearest Neighbor Regressor*, que gerou o menor erro, apresenta uma reta de relação num valor um pouco acima de zero, o mesmo ocorre para o valor de *Random Forest*. Os outros resíduos de latitude possuem uma variação com decaimento na curva. Já o valor de resíduo de longitude para *K-Nearest Neighbor* apresentou pouca variação sobre o eixo X. Os valores residuais de longitude para *Random Forest* e *Extra Trees* apresentam um resultado semelhante ao *K-Nearest Neighbor*. O método *Bagging Regressor* gerou o melhor resultado residual para longitude, ficando praticamente em zero. Nos próximos resultados serão discutidos os crimes que foram preditos. Os resultados foram plotados em gráficos de dispersão e podem ser visualizados nas Figuras 67 a 71.

Figura 67: Pontos preditos para K-Nearest Neighbor Regressor



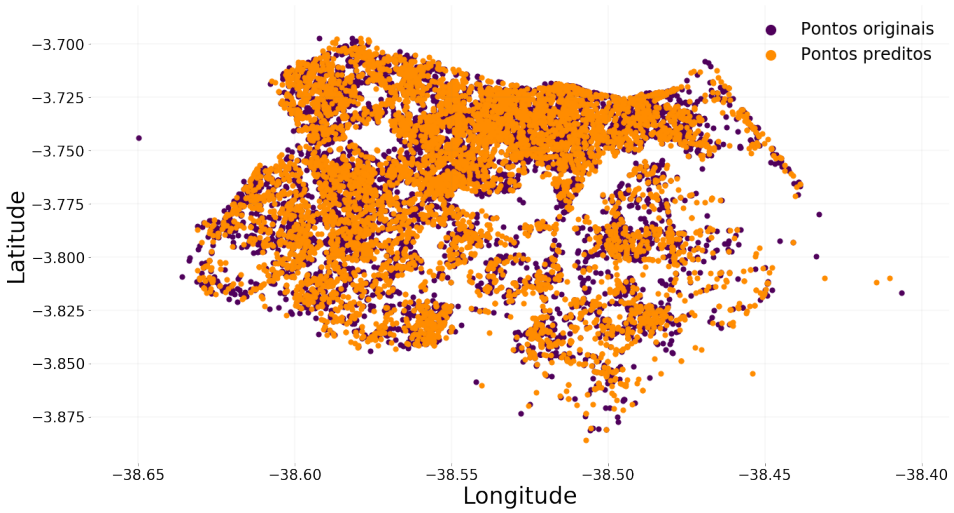
Fonte: Autoria própria.

Figura 68: Pontos preditos para Random Forest Regressor



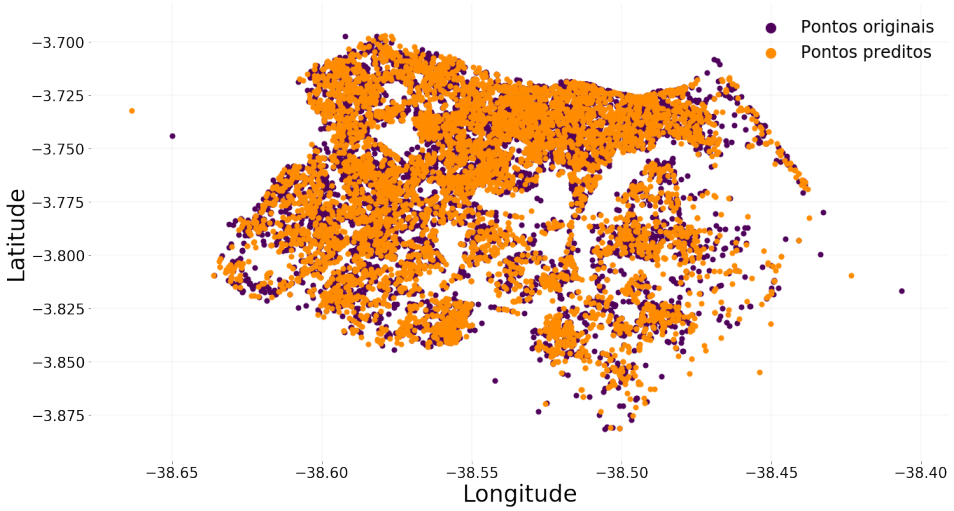
Fonte: Autoria própria.

Figura 69: Pontos preditos para Extra Trees Regressor



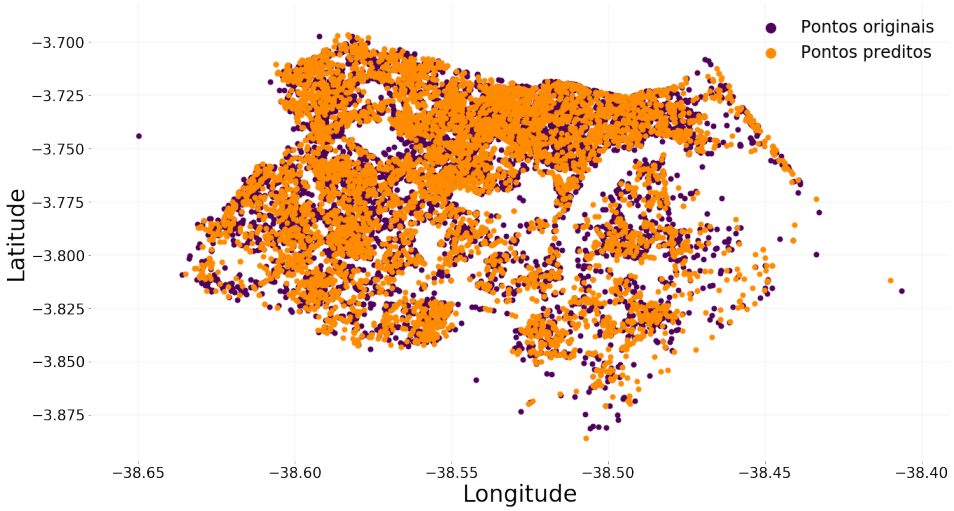
Fonte: Autoria própria.

Figura 70: Pontos preditos para Decision Tree Regressor



Fonte: Autoria própria.

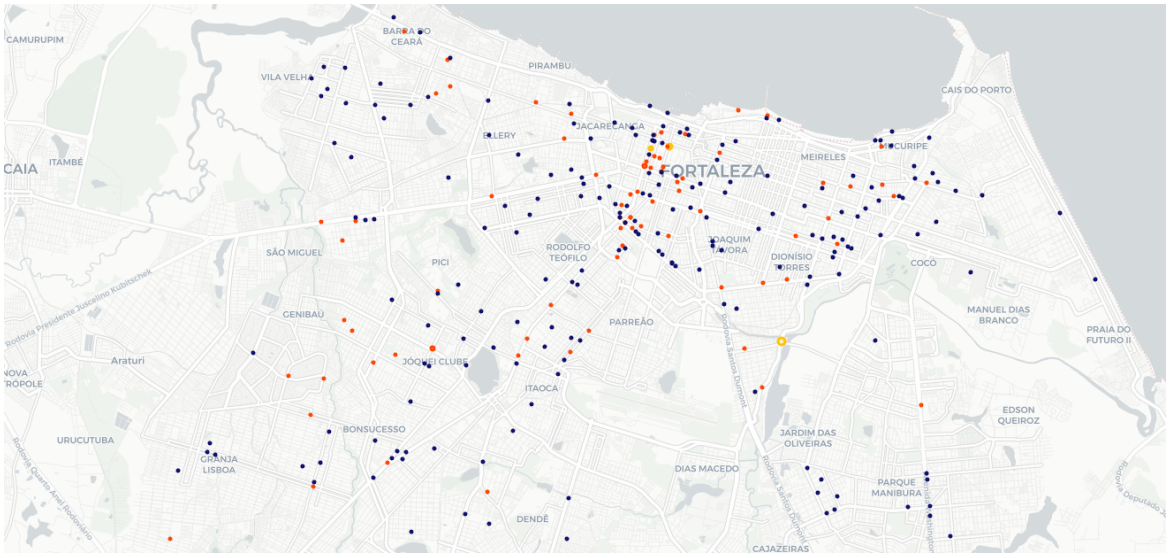
Figura 71: Pontos preditos para Bagging Regressor



Fonte: Autoria própria.

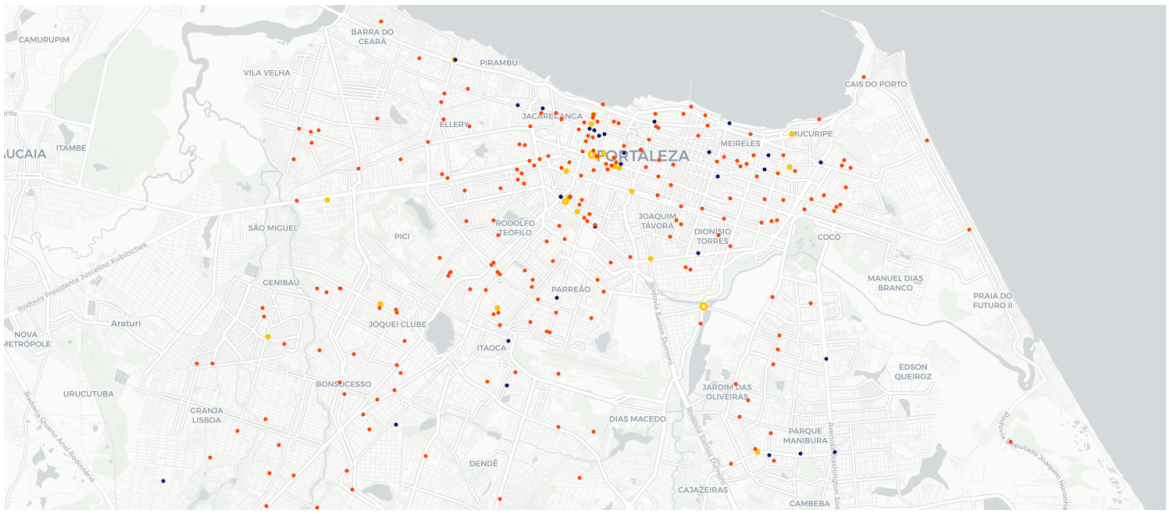
As Figuras acima apresentam os crimes preditos juntamente com os dados originais para a cidade de Fortaleza. Assim como os valores da Tabela 9, os pontos preditos para cada um dos métodos também estão aproximados um do outro, como destaque para o método que obteve o menor erro, o *K-Nearest Neighbor Regressor*. Para o menor que obteve o maior erro, *Extra Trees*, não é possível, visualmente, aponta muitas diferenças em relação ao *K-Nearest Neighbor*, os pontos preditos estão similares. E assim como anteriormente, foi gerado também o mapa dos pontos preditos. Esses mapas são exibidos nas Figuras 73 a 77, além dos pontos originais que são exibidos na Figura 72.

Figura 72: Pontos originais - Mapa da cidade de Fortaleza



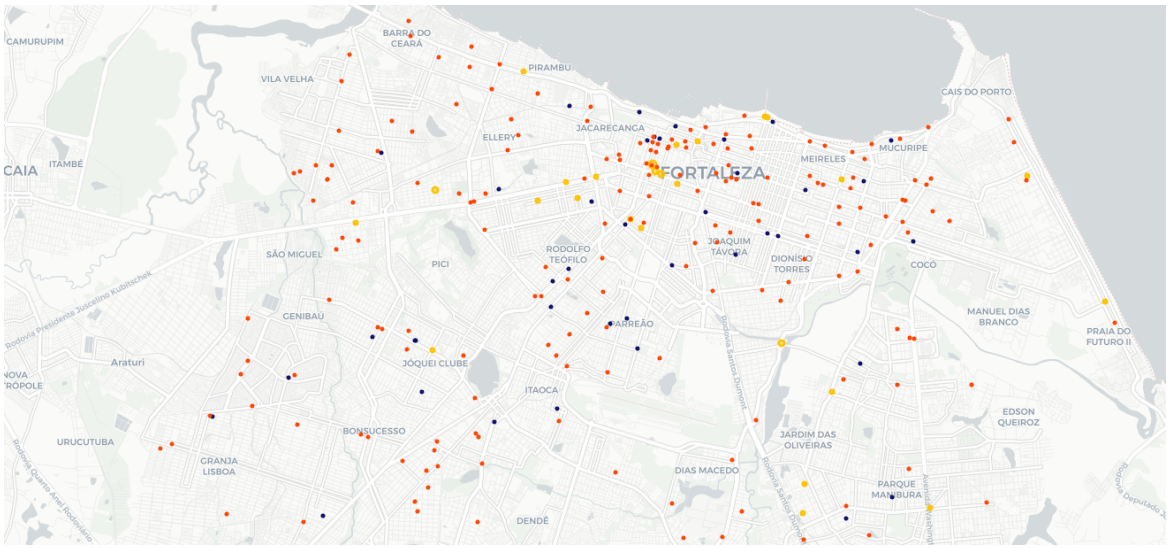
Fonte: Autoria própria.

Figura 73: Crimes preditos - K-Nearest Neighbor Regressor



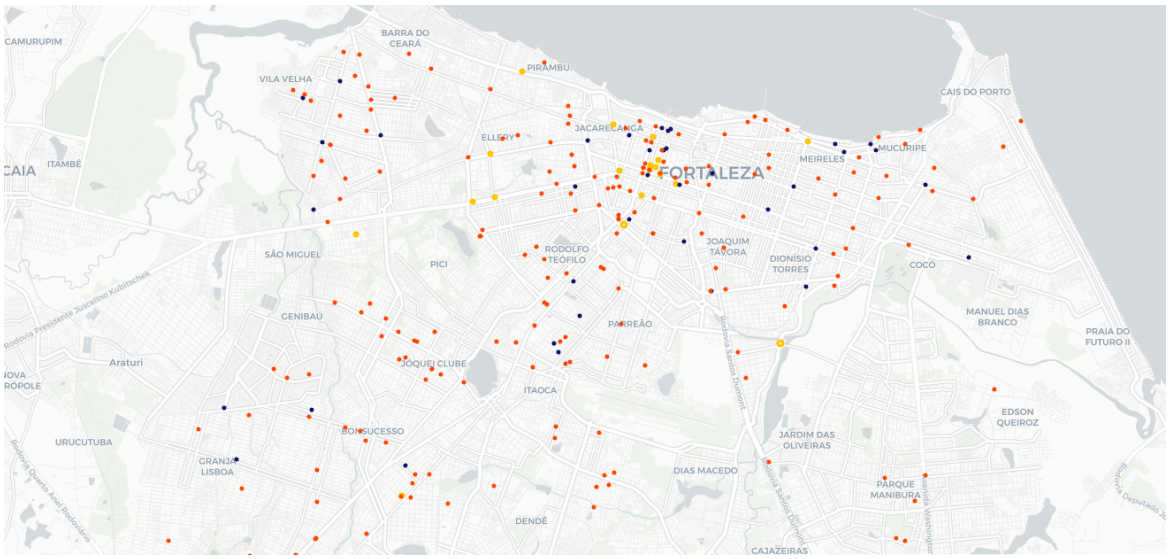
Fonte: Autoria própria.

Figura 74: Crimes preditos - Random Forest Regressor



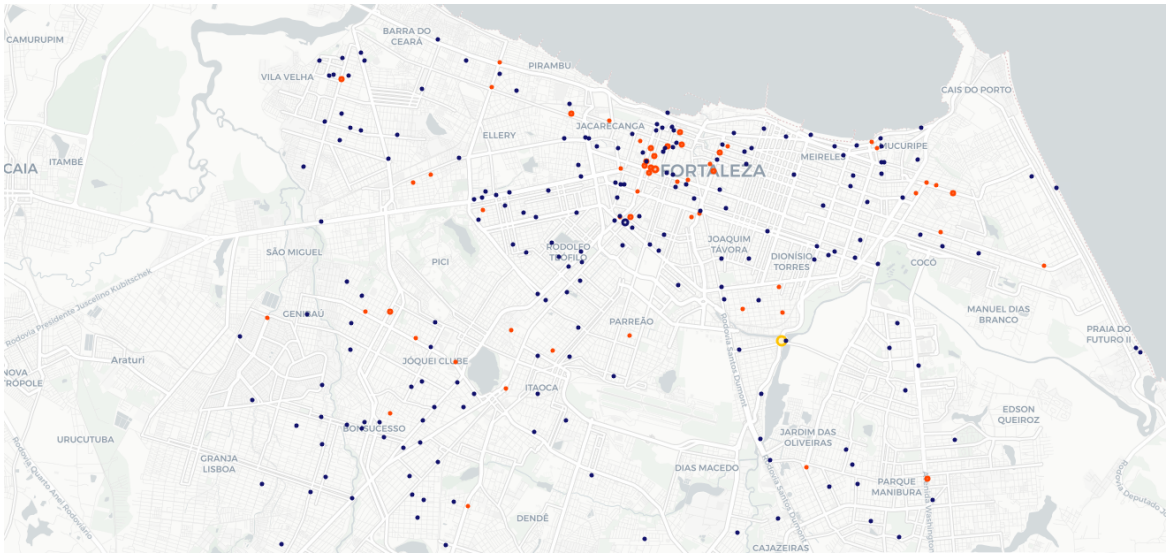
Fonte: Autoria própria.

Figura 75: Crimes preditos - Extra Trees Regressor



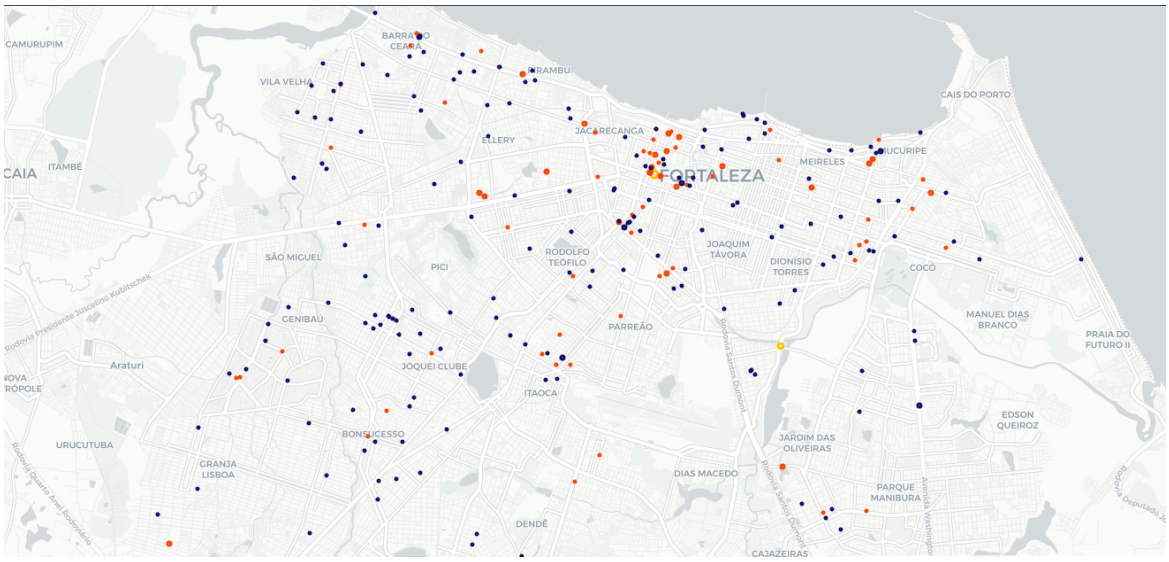
Fonte: Autoria própria.

Figura 76: Crimes preditos - Decision Tree Regressor



Fonte: Autoria própria.

Figura 77: Crimes preditos - Bagging Regressor



Fonte: Autoria própria.

Assim como os resultados anteriores, os crimes preditos nos mapas também estão divididos em frequência de ocorrência, sendo azul o mais baixo e amarelo o mais alto. Como explicado anteriormente, considera-se 5 ocorrências ou mais de crimes em um mesmo ponto como uma alta quantidade. Entre 3 e 5 ocorrências em um mesmo local, considera-se como média e menos que 3, é classificada como baixa. Os pontos preditos para o método *K-Nearest Neighbor* não se assemelham muito aos pontos originais, diferentemente, por exemplo, dos pontos preditos pelo método *Random Forest*, *Extra Trees* e *Decision Tree*, esse último possui também uma quantidade semelhante em relação a frequência de poucos crimes cometidos, os pontos azuis. O valor de *Bagging Regressor* possui também um resultado semelhante aos pontos originais, exceto por alguns locais em que os pontos estão mais distribuídos.

4.2.3 EXPERIMENTO 2 - CATEGORIA 1

Os próximos resultados que serão discutidos, dizem respeito ao experimento 2. Nesse experimento será considerado que os crimes em uma rua acontecem somente em um ponto específico. Ou seja, agora não há mais distribuição dos valores de latitude e longitude, esses valores estão concentrados em apenas um local. Na categoria 1 desse novo experimento, foi considerado também que a saída seria dois valores, latitude e longitude. A Tabela 10 apresenta os parâmetros utilizados para treinamento e os valores de erro resultante para cada um dos métodos utilizados.

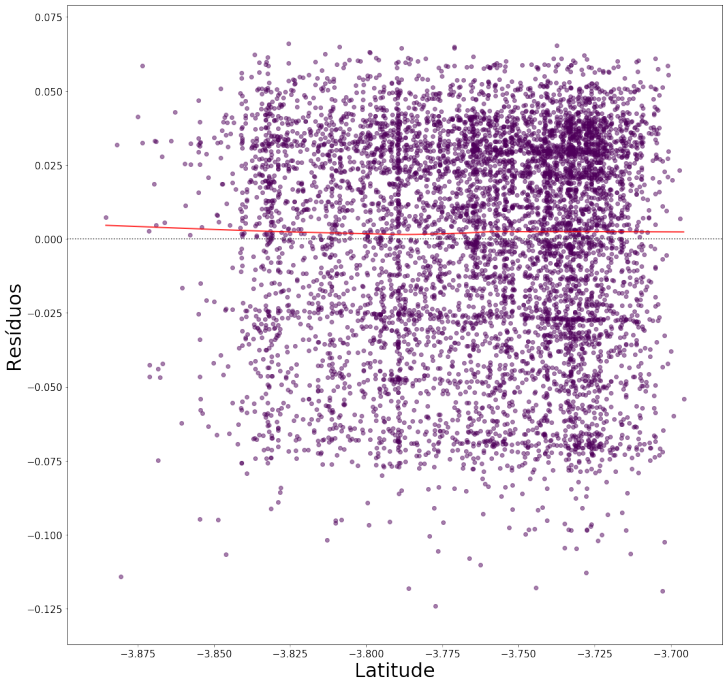
Tabela 10: Resultado experimento 2 - categoria 1

Método	Parâmetros			Erro	
	Número de estimadores	**Profundidade máxima**	**Número de vizinhos**	**MSE**	**RMSE**
K-Nearest Neighbor Regressor	**-**	**-**	**1**	**0.00279**	**0.05286**
RandomForestRegressor	1	-	-	0.00293	0.05418
ExtraTreesRegressor	1	43	-	0.00293	0.05417
DecisionTreeRegressor	-	60	-	0.00296	0.05445
BaggingRegressor	1	-	-	0.00301	0.05489

Fonte: Autoria própria.

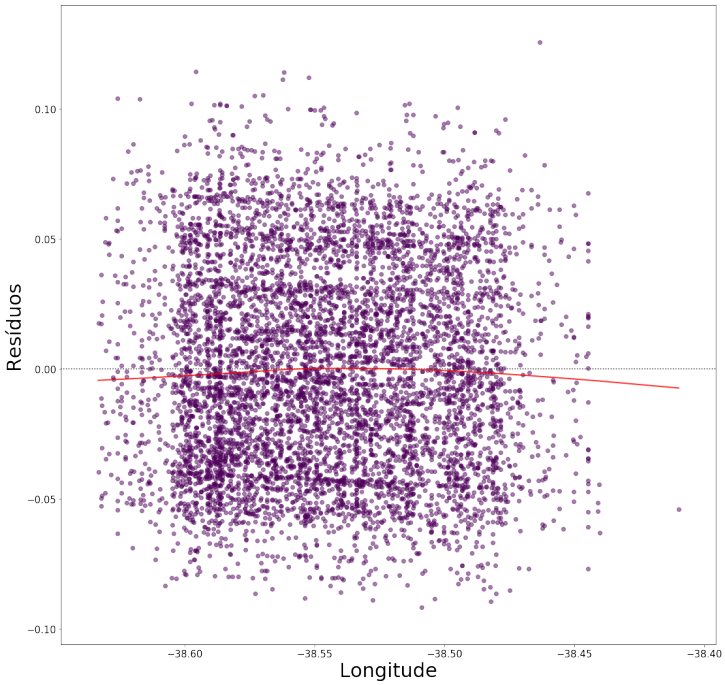
Pela tabela nota-se que o método de *K-Nearest Neighbor Regressor* obteve o menor erro entre as técnicas utilizadas, bem como também precisou de apenas um parâmetro para ser treinado. Os métodos de *Random Forest* e *Extra Trees* obtiveram erros MSE idênticos, mas obtiveram RMSE diferentes, sendo que *Extra Trees* obteve o menor. Os métodos de *Bagging Regressor* e *Decision Tree* obtiveram o primeiro maior e o segundo maior erro, respectivamente. Os próximos resultados analisados tratam-se dos valores residuais calculados. As Figuras 78 a 87 exibem os resíduos calculados.

Figura 78: Resíduos para Latitude de K-Nearest Neighbor Regressor



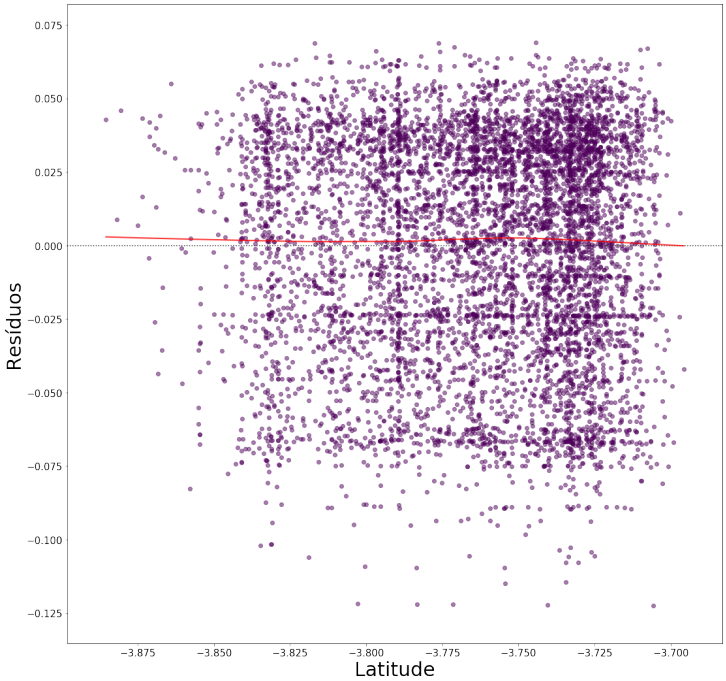
Fonte: Autoria própria.

Figura 79: Resíduos para Longitude de K-Nearest Neighbor Regressor



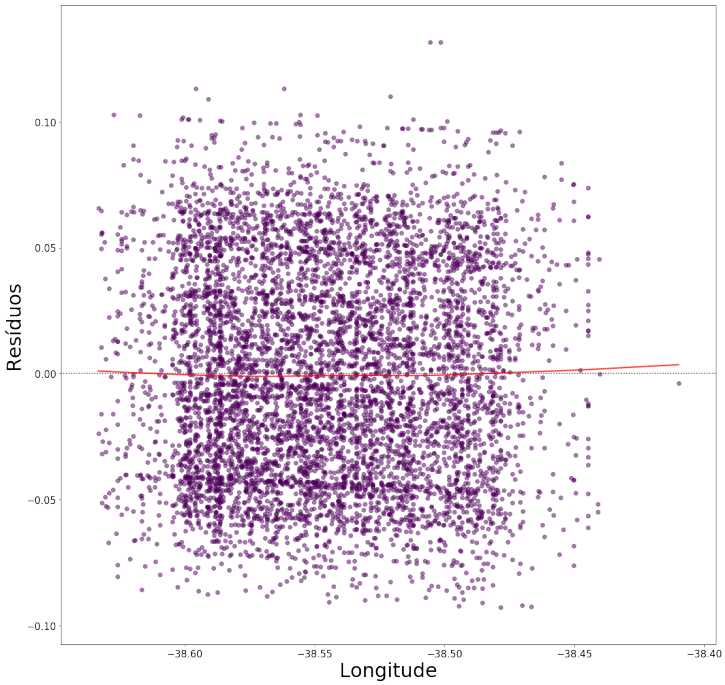
Fonte: Autoria própria.

Figura 80: Resíduos para Latitude de Random Forest Regressor



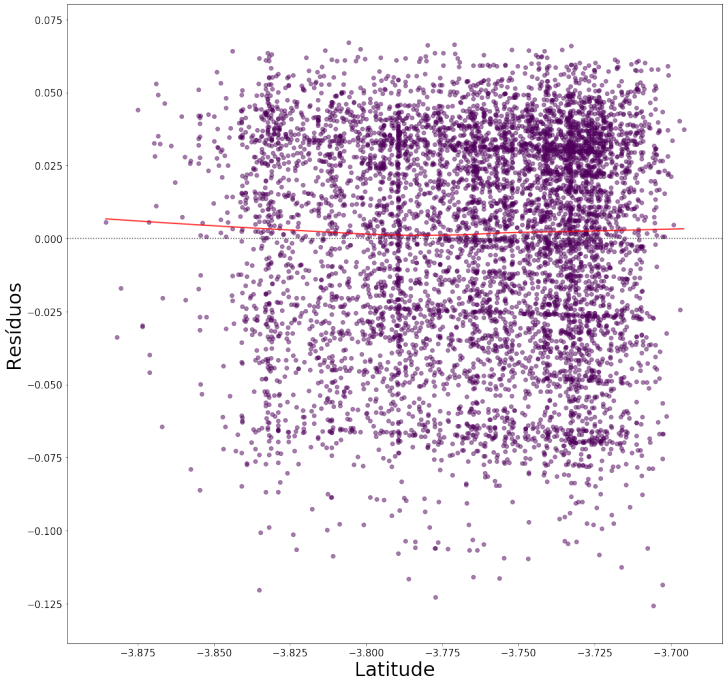
Fonte: Autoria própria.

Figura 81: Resíduos para Longitude de Random Forest Regressor



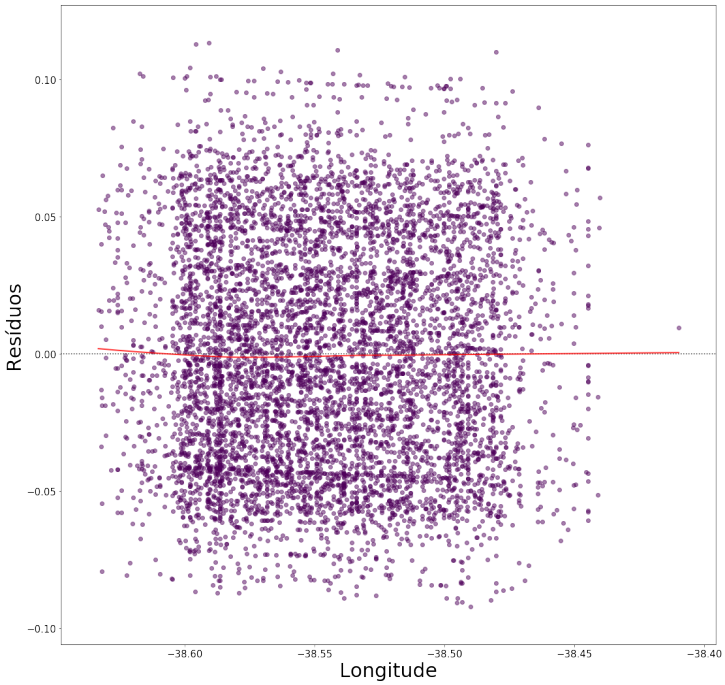
Fonte: Autoria própria.

Figura 82: Resíduos para Latitude de Extra Trees Regressor



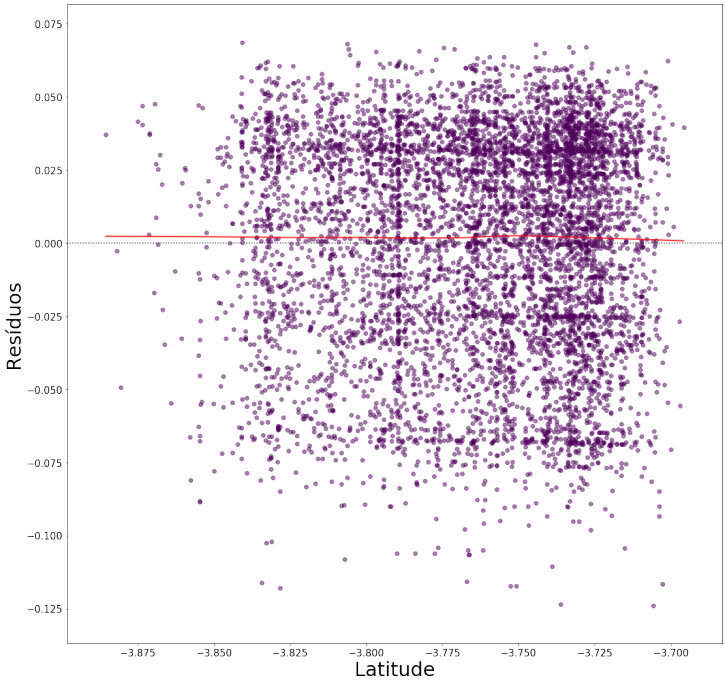
Fonte: Autoria própria.

Figura 83: Resíduos para Longitude de Extra Trees Regressor



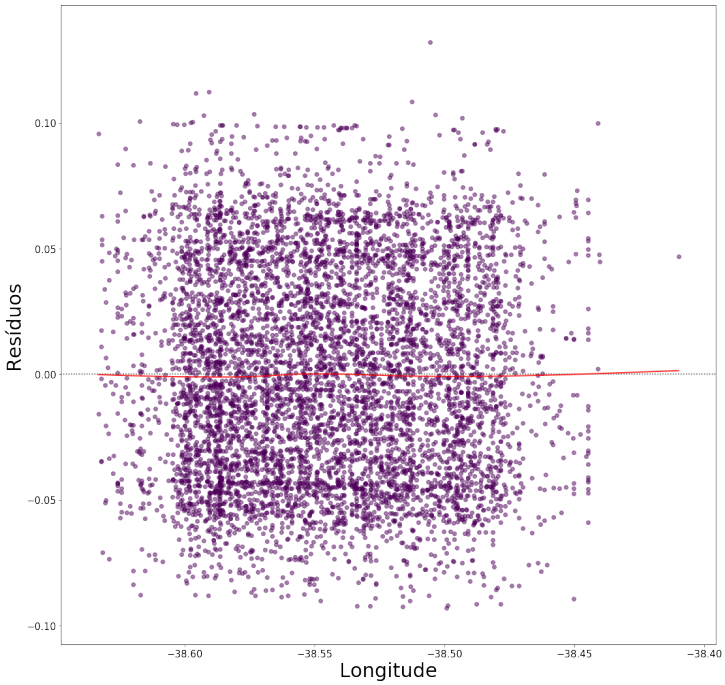
Fonte: Autoria própria.

Figura 84: Resíduos para Latitude de Decision Tree Regressor



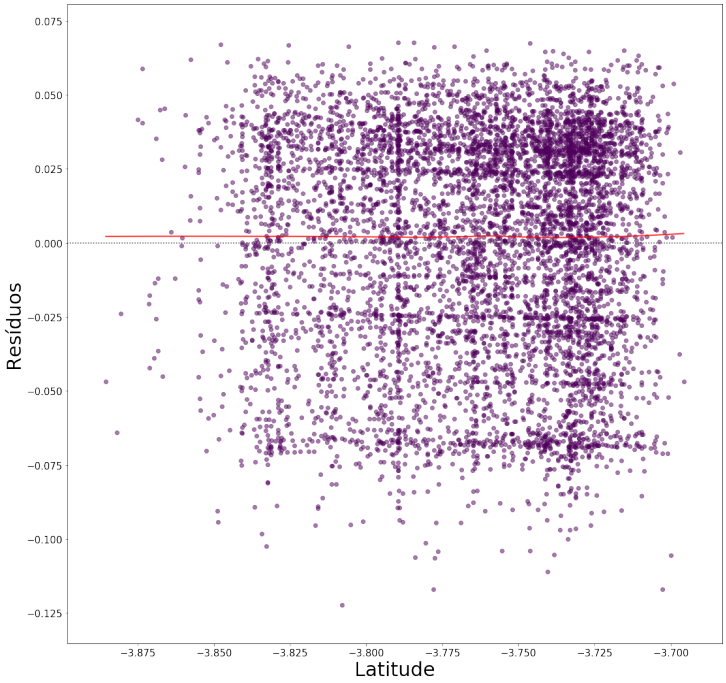
Fonte: Autoria própria.

Figura 85: Resíduos para Longitude de Decision Tree Regressor



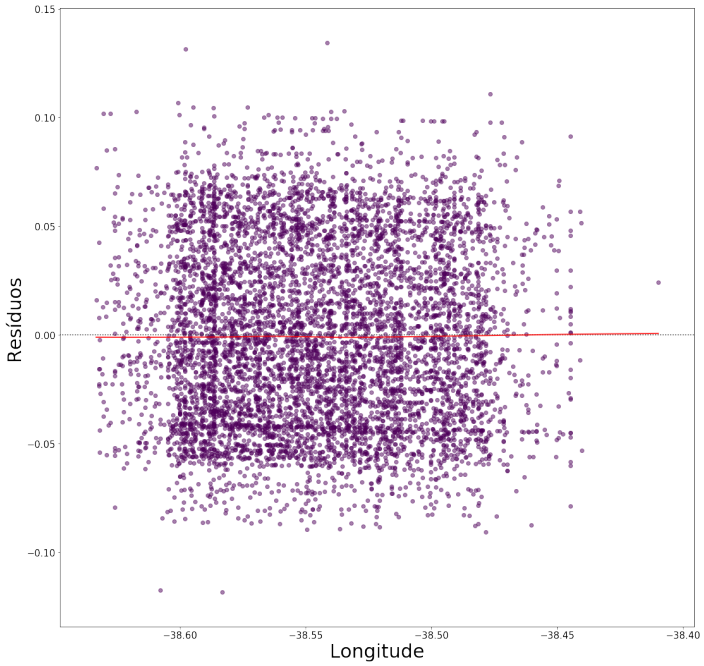
Fonte: Autoria própria.

Figura 86: Resíduos para Latitude de Bagging Regressor



Fonte: Autoria própria.

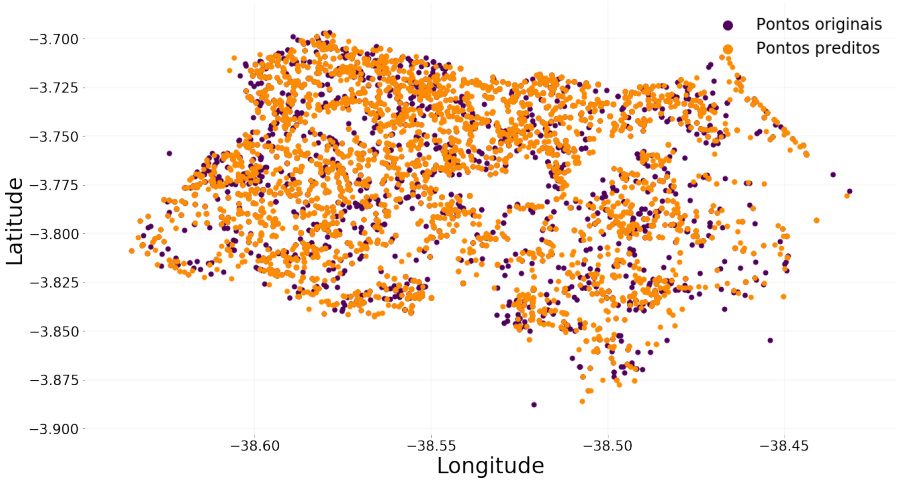
Figura 87: Resíduos para Longitude de Bagging Regressor



Fonte: Autoria própria.

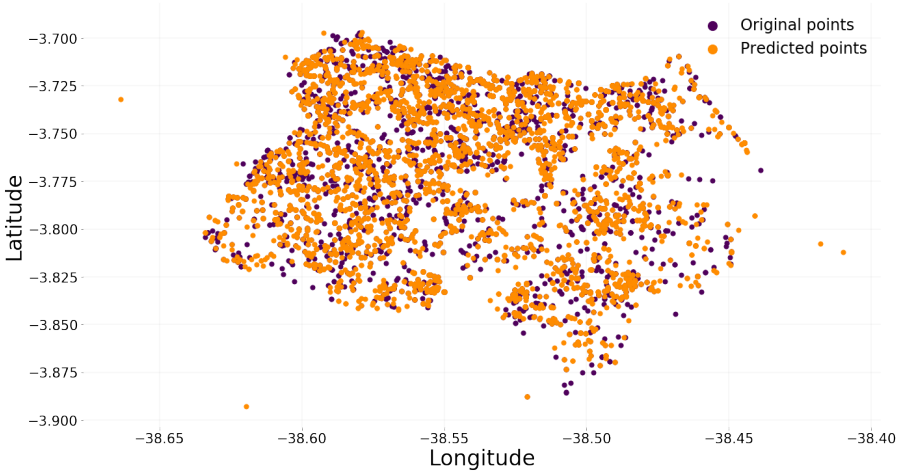
Para os cálculos residuais da longitude do método *K-Nearest Neighbor*, nota-se em alguns que há uma tendência negativa dos valores apresentados, o que representa que alguns pontos preditos foram maiores que os pontos originais. Para o valor de latitude dos mesmos métodos há uma tendência positiva, ainda assim os dois valores, latitude e longitude estão próximos de zero. Os métodos de *Random Forest*, *Extra Trees*, *Decision Tree* e *Bagging Regressor* possuem valores residuais semelhantes, tanto para latitude quanto para longitude, valores com tendência quase zero. Já para o método de *Random Forest*, o valor residual para longitude mostra uma pequena variação positiva mas que também está próxima de zero, assim como valor para latitude. A seguir são apresentados os pontos preditos e plotados em um gráfico de dispersão.

Figura 88: Pontos preditos para K-Nearest Neighbor Regressor



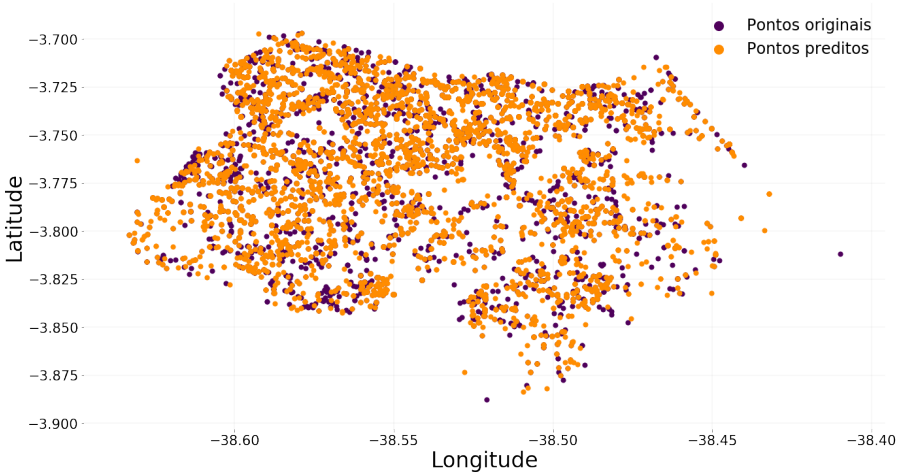
Fonte: Autoria própria.

Figura 89: Pontos preditos para Random Forest Regressor



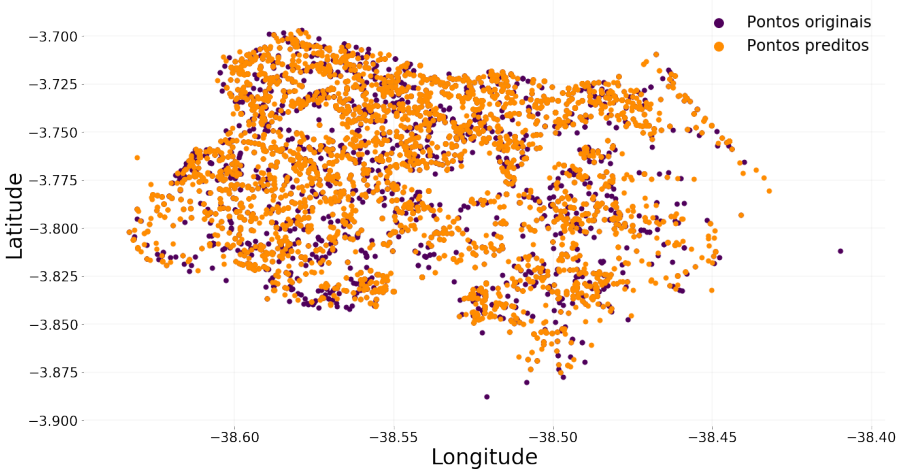
Fonte: Autoria própria.

Figura 90: Pontos preditos para Extra Trees Regressor



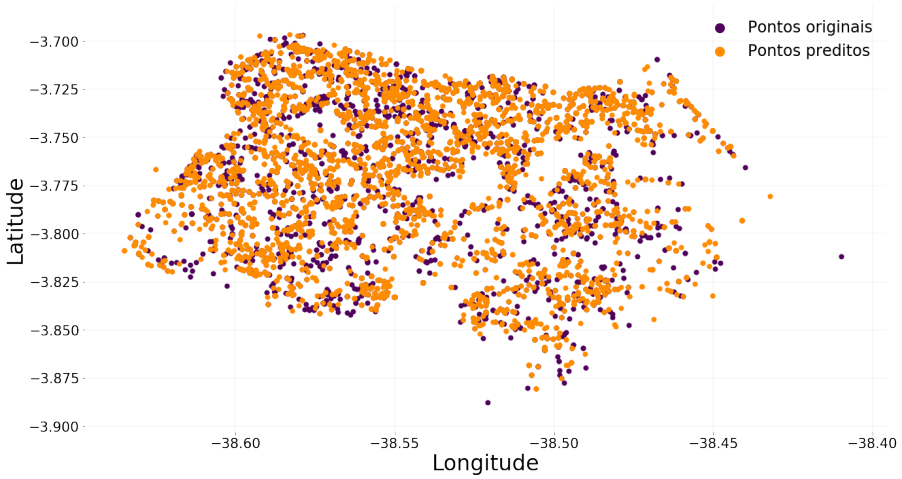
Fonte: Autoria própria.

Figura 91: Pontos preditos para Decision Tree Regressor



Fonte: Autoria própria.

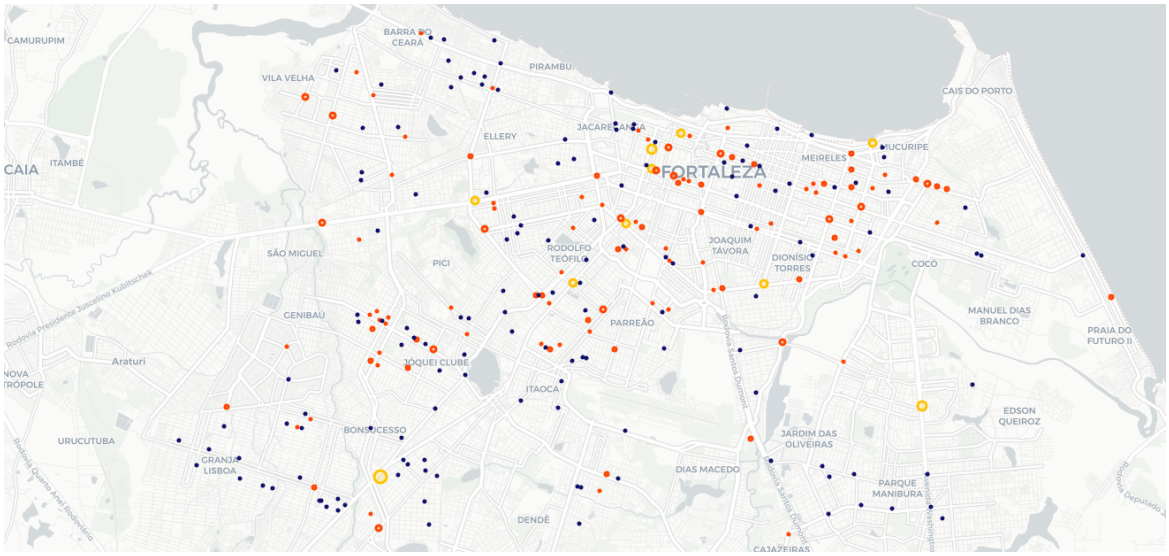
Figura 92: Pontos preditos para Bagging Regressor



Fonte: Autoria própria.

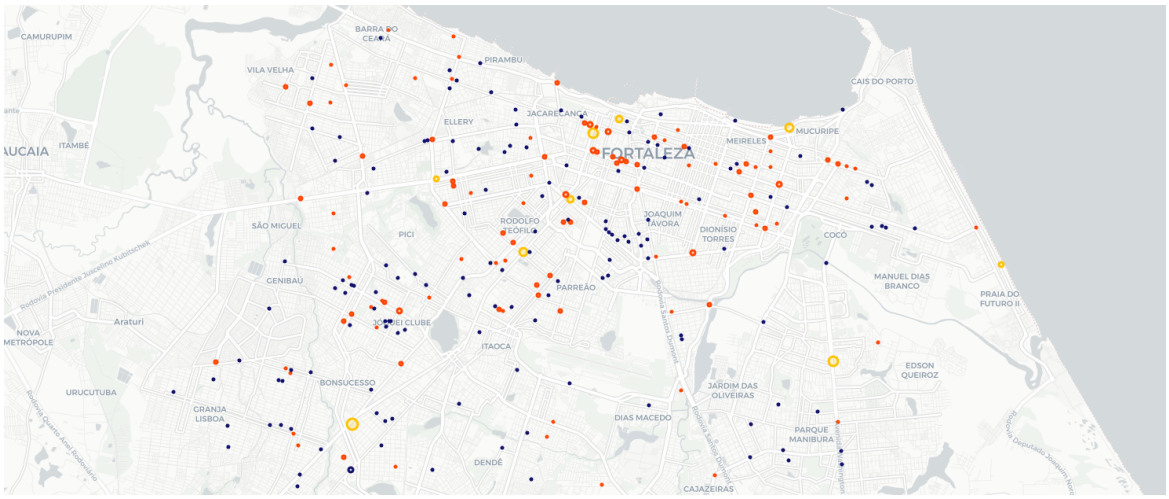
Pelas figuras 88 a 92 é possível notar que há uma diferença entre os pontos preditos para esse experimento comparado ao experimento 1. É percebível que os dados preditos estão mais dispersos, isso é possível pelo fato de os crimes, agora, ocorrerem em apenas um ponto de uma rua. Os pontos preditos para o método *K-Nearest Neighbor* apresentam melhores dispersões em relação aos pontos dos outros métodos. Os crimes preditos para *Random Forest*, *Extra Trees*, *Decision Tree* e *Bagging Regressor* são bastante similares, o que mostra também a eficiência dos modelos mesmo com parâmetros diferentes sendo utilizados. Foi gerado também um mapa para os pontos preditos pelos métodos. Esses resultados são exibidos a seguir.

Figura 93: Pontos originais - Mapa da cidade de Fortaleza



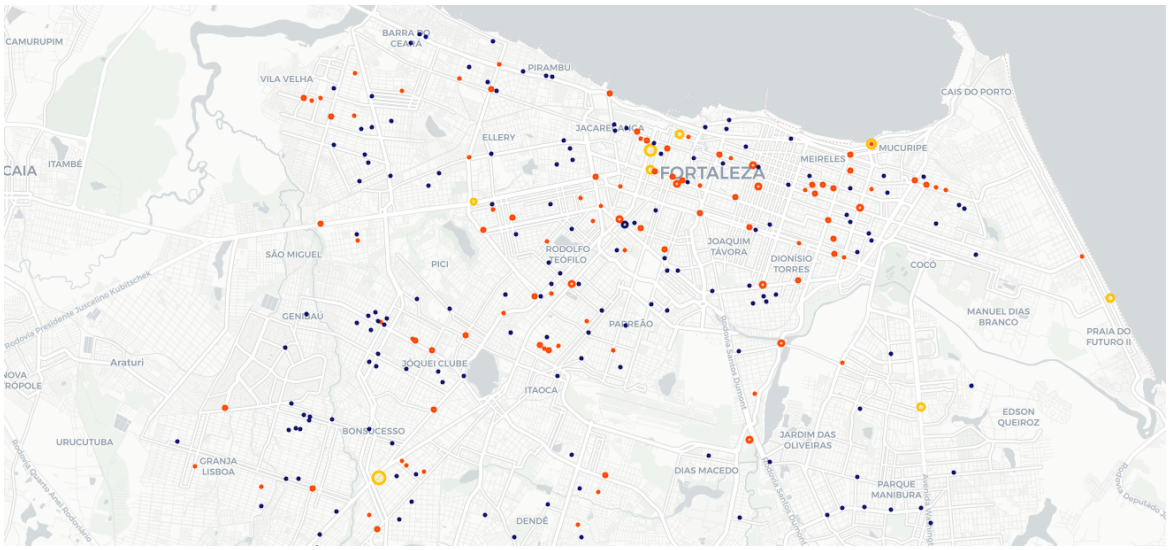
Fonte: Autoria própria.

Figura 94: Crimes preditos - K-Nearest Neighbor Regressor



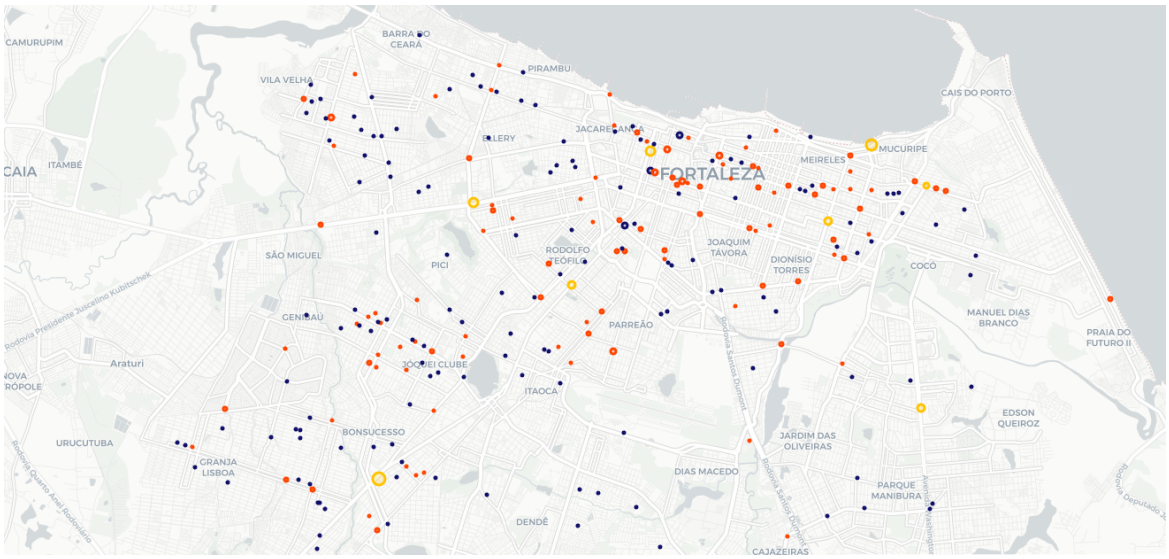
Fonte: Autoria própria.

Figura 95: Crimes preditos - Random Forest Regressor



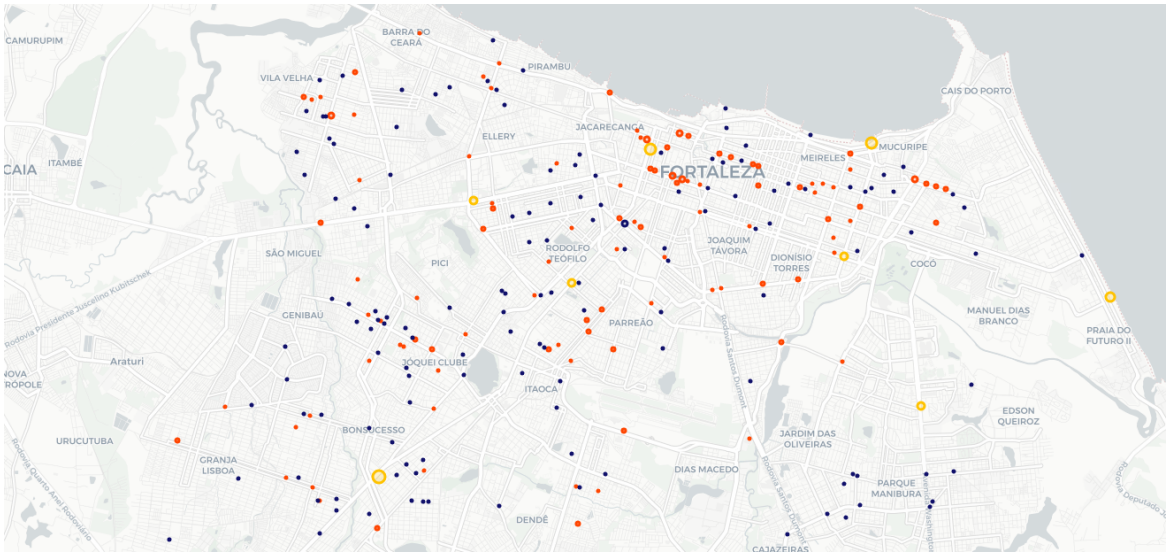
Fonte: Autoria própria.

Figura 96: Crimes preditos - Extra Trees Regressor



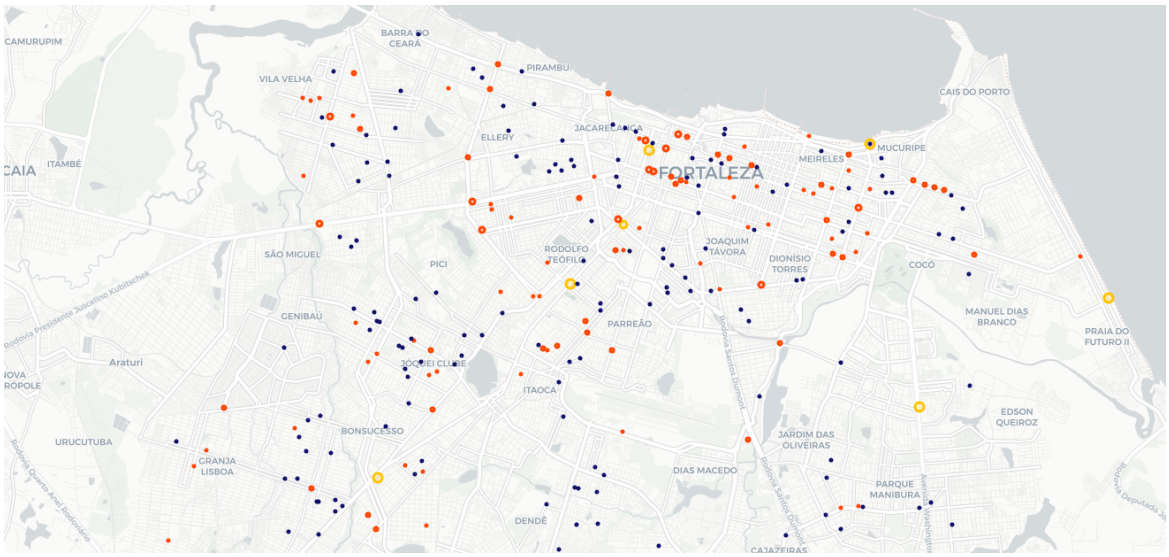
Fonte: Autoria própria.

Figura 97: Crimes preditos - Decision Tree Regressor



Fonte: Autoria própria.

Figura 98: Crimes preditos - Bagging Regressor



Fonte: Autoria própria.

Os crimes preditos para *K-Nearest Neighbor* apresentam bastante similaridades com os pontos originais, a frequência da ocorrência de crimes também é parecida, isso é notado na Figura 94. Os resultados gerados para os métodos *Random Forest*, *Extra Trees*, *Decision Tree* e *Bagging Regressor* também possuem uma similaridade com os pontos originais. Nesse experimento ficou evidente que cada um dos resultados visuais apresentados está condizente com os valores apresentados na Tabela 10.

4.2.4 EXPERIMENTO 2 - CATEGORIA 2

Na categoria 2 do segundo experimento, ainda é considerado que os crimes ocorrem em apenas um ponto de uma rua, entretanto, a saída será apenas um valor, o *hash*. Esse valor representa latitude e longitude. A Tabela 11 apresenta os parâmetros utilizados para cada um dos métodos e também os erros resultantes dos modelos.

Tabela 11: Resultado experimento 2 - categoria 2

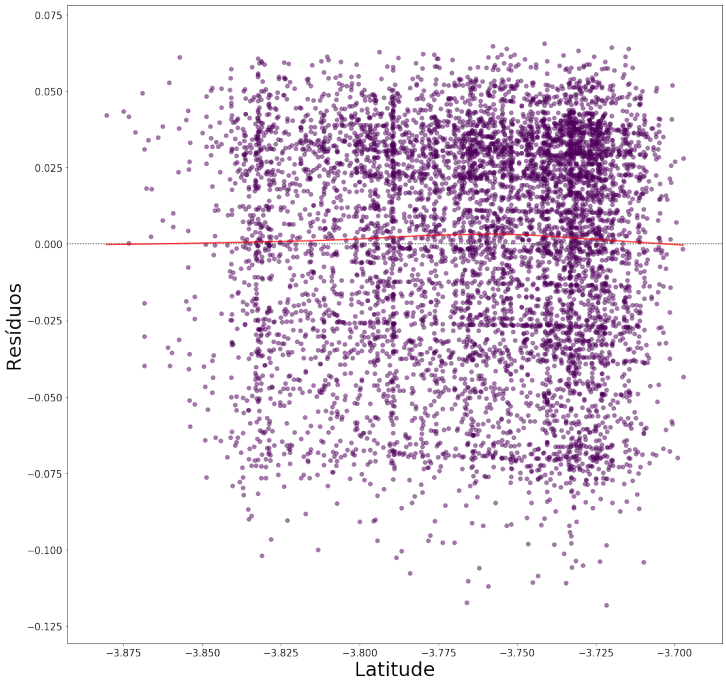
	Parâmetros			**Erro**	
Método	**Número de estimadores**	**Profundidade máxima**	**Número de vizinhos**	**MSE**	**RMSE**
K-Nearest Neighbor Regressor	**-**	**-**	**1**	**3498087.80**	**1870.31**
Random Forest Regressor	1	-	-	3658895.33	1912.82
Extra Trees Regressor	10	60	-	3606293.78	1899.02
Decision Tree Regressor	-	30	-	3641033.68	1908.14
Bagging Regressor	1	-	-	3635271.33	1906.63

Fonte: Autoria própria.

O método *K-Nearest Neighbor* novamente apresentou o menor erro, seguido pelo *Extra Trees Regressor*. O método de *Random Forest* e *Decision Tree* apresentaram o primeiro e o segundo maior erro, respectivamente. O próximo resultado apresentado

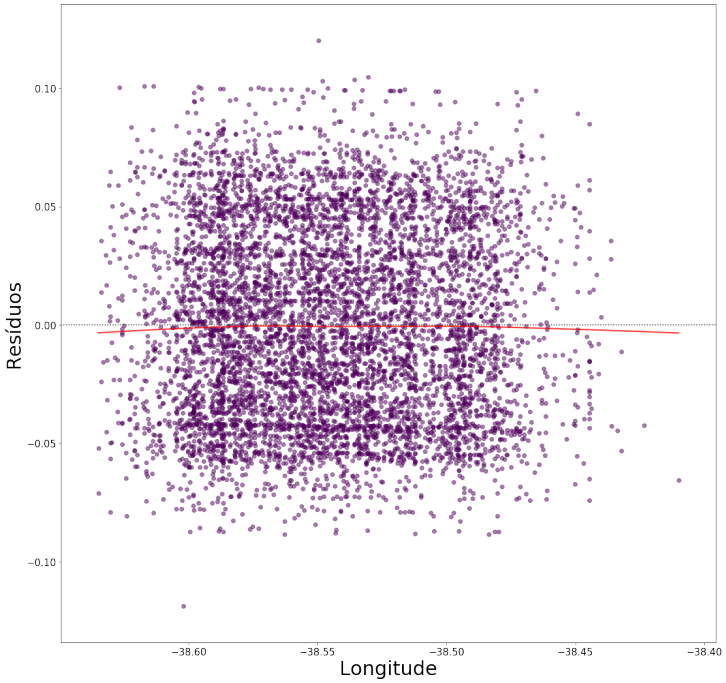
exibe os valores residuais calculados para latitude e longitude de cada um dos métodos. Esses resultados são apresentados a seguir.

Figura 99: Resíduos para Latitude de K-Nearest Neighbor Regressor



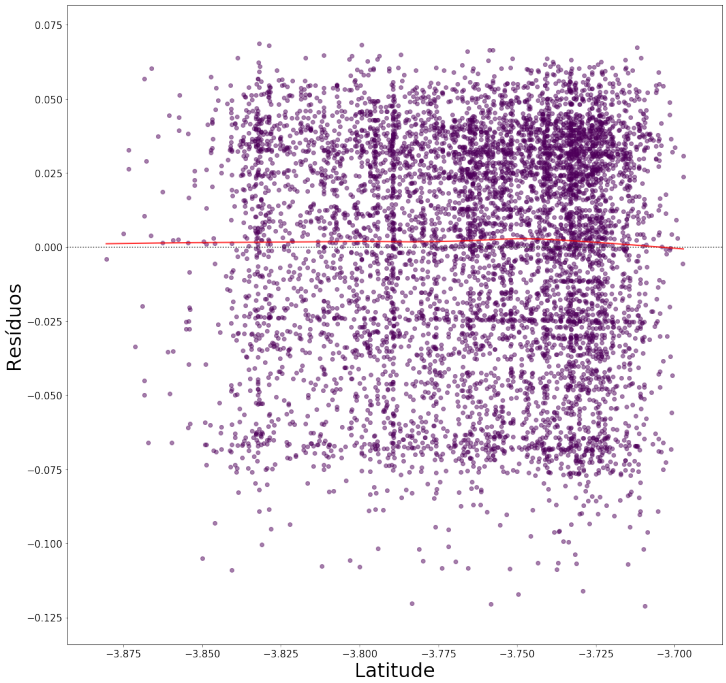
Fonte: Autoria própria.

Figura 100: Resíduos para Longitude de K-Nearest Neighbor Regressor



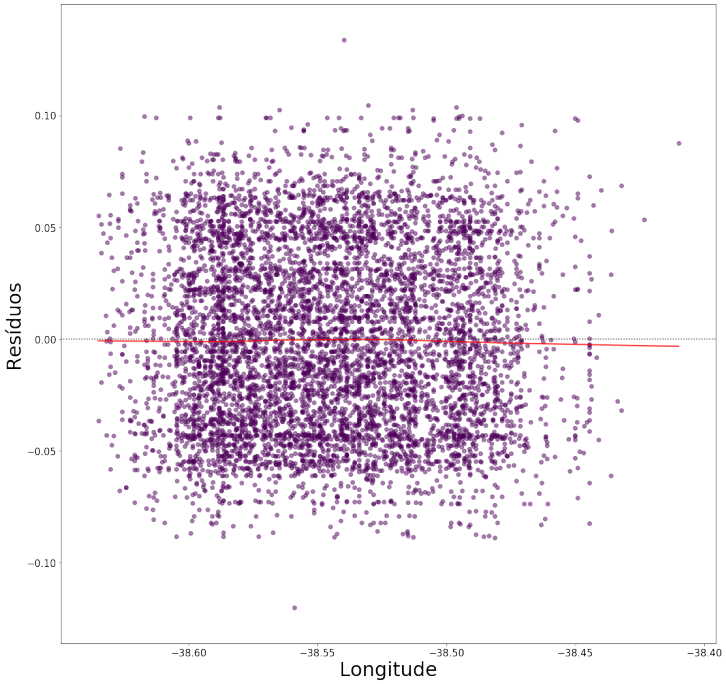
Fonte: Autoria própria.

Figura 101: Resíduos para Latitude de Random Forest Regressor



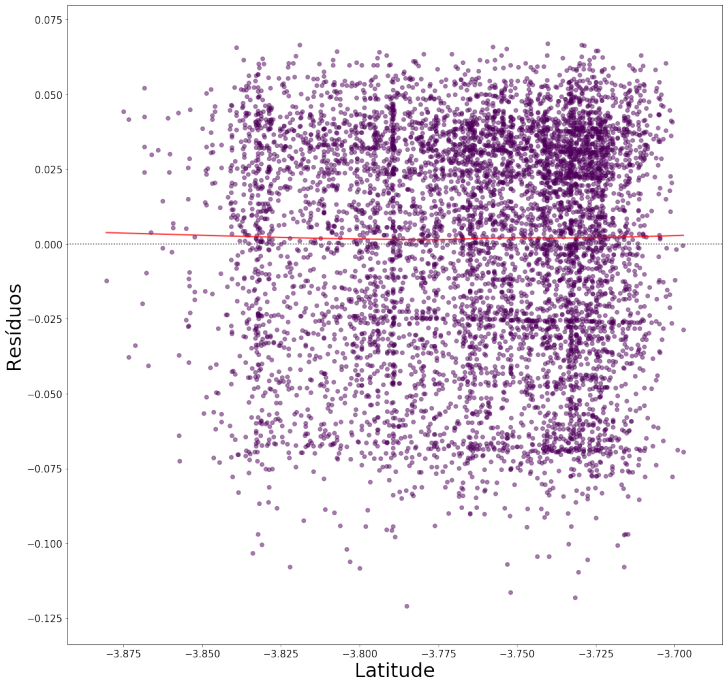
Fonte: Autoria própria.

Figura 102: Resíduos para Longitude de Random Forest Regressor



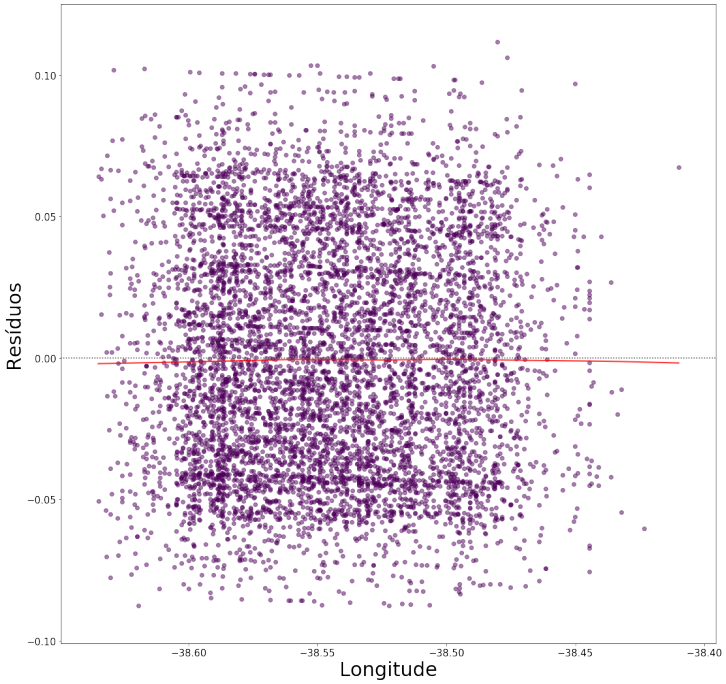
Fonte: Autoria própria.

Figura 103: Resíduos para Latitude de Extra Trees Regressor



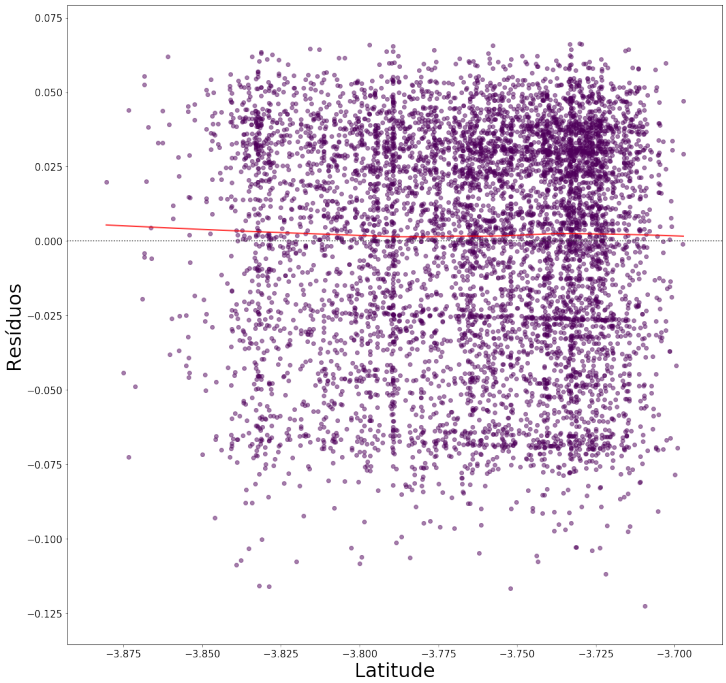
Fonte: Autoria própria.

Figura 104: Resíduos para Longitude de Extra Trees Regressor



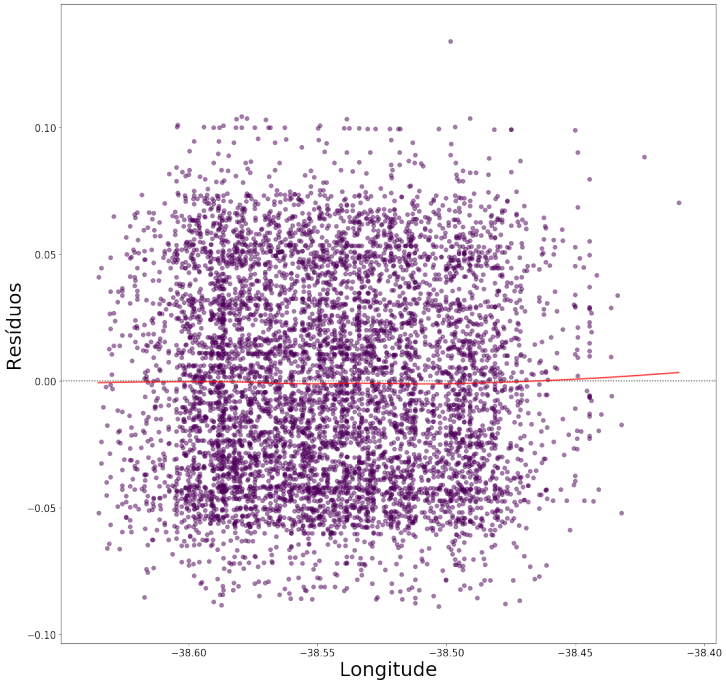
Fonte: Autoria própria.

Figura 105: Resíduos para Latitude de Decision Tree Regressor



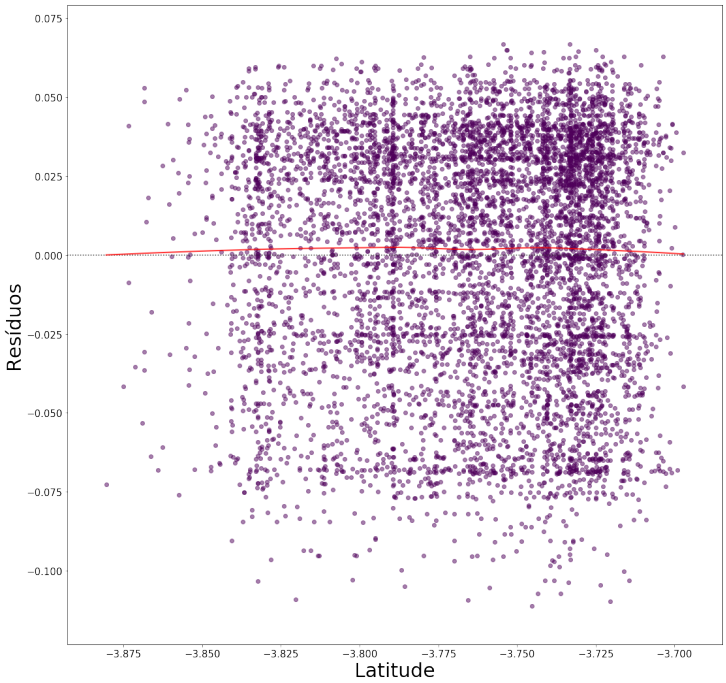
Fonte: Autoria própria.

Figura 106: Resíduos para Longitude de Decision Tree Regressor



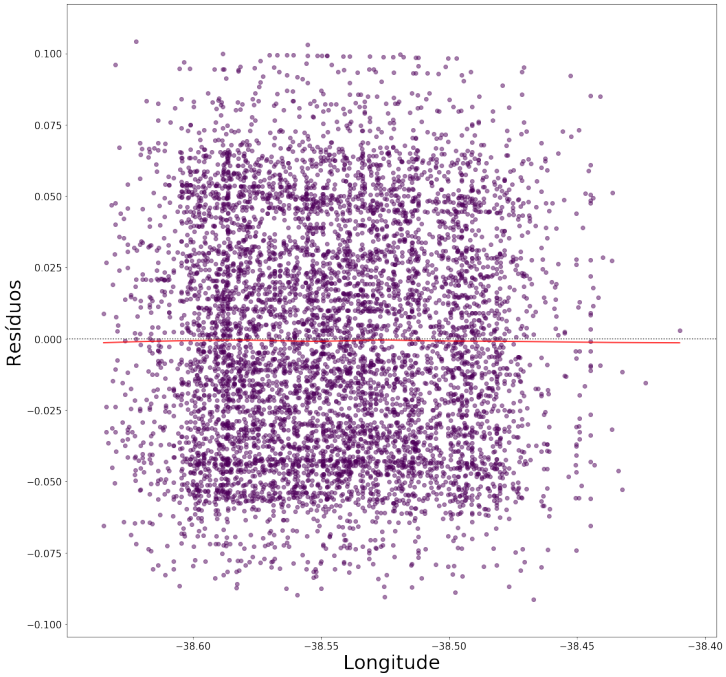
Fonte: Autoria própria.

Figura 107: Resíduos para Latitude de Bagging Regressor



Fonte: Autoria própria.

Figura 108: Resíduos para Longitude de Bagging Regressor

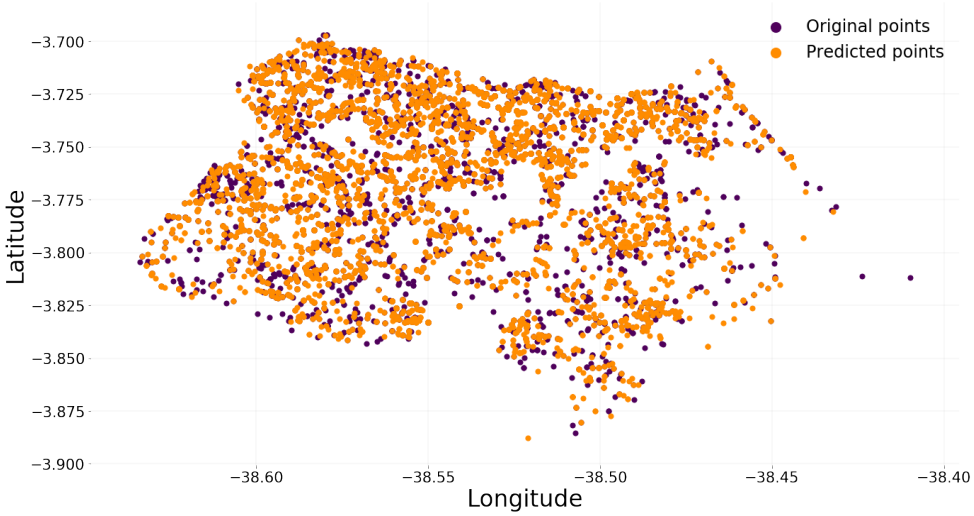


Fonte: Autoria própria.

As Figuras 99 a 108 apresentam os valores residuais dos métodos utilizados nesse trabalho. Para *K-Nearest Neighbor* os resíduos estão próximos de zero, tanto para latitude quanto longitude, com uma pequena variação para esse último. Para o método de *Random Forest* e *Decision Tree*, os valores obtidos são semelhantes ao do *K-Nearest Neighbor*. O método *Extra Trees* gerou um resíduo bem próximo a zero para o valor de

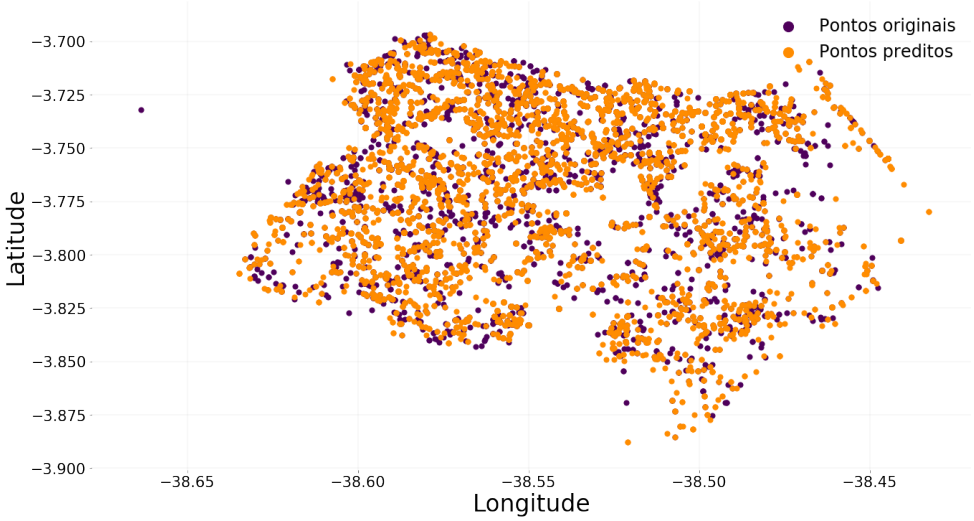
longitude, para latitude o valor residual também segue uma linha reta, o que é um bom resultado. O método de *Bagging Regressor* apresentou também ótimos valores residuais, sendo os dois próximos a zero e sem muita variação. Os próximos resultados apresentados dizem respeito aos pontos preditos que foram plotados em um gráfico de dispersão. As figuras a seguir apresentam esses dados.

Figura 109: Pontos preditos para K-Nearest Neighbor Regressor



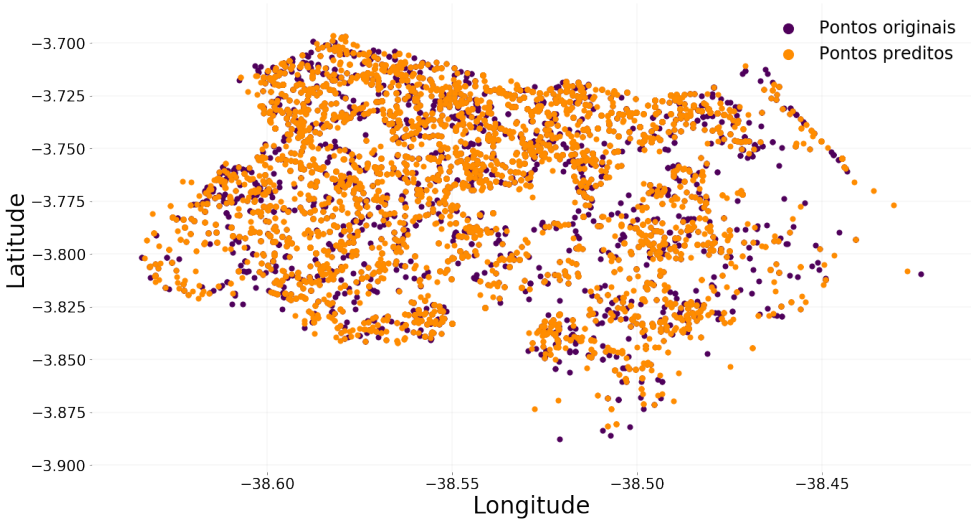
Fonte: Autoria própria.

Figura 110: Pontos preditos para Random Forest Regressor



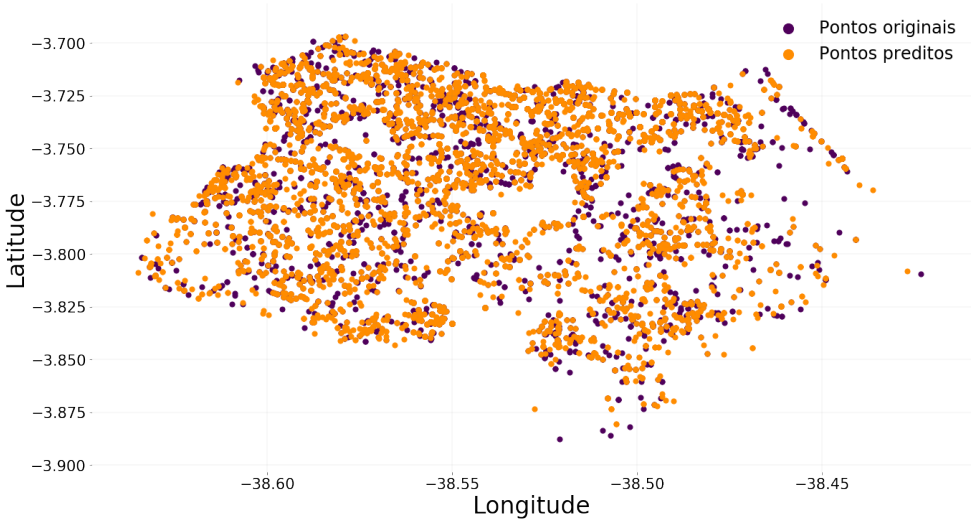
Fonte: Autoria própria.

Figura 111: Pontos preditos para Extra Trees Regressor



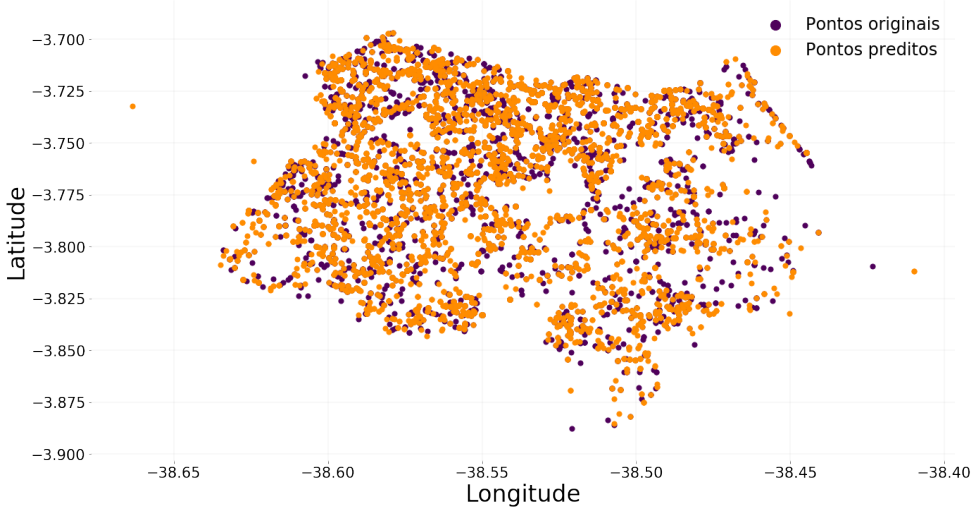
Fonte: Autoria própria.

Figura 112: Pontos preditos para Decision Tree Regressor



Fonte: Autoria própria.

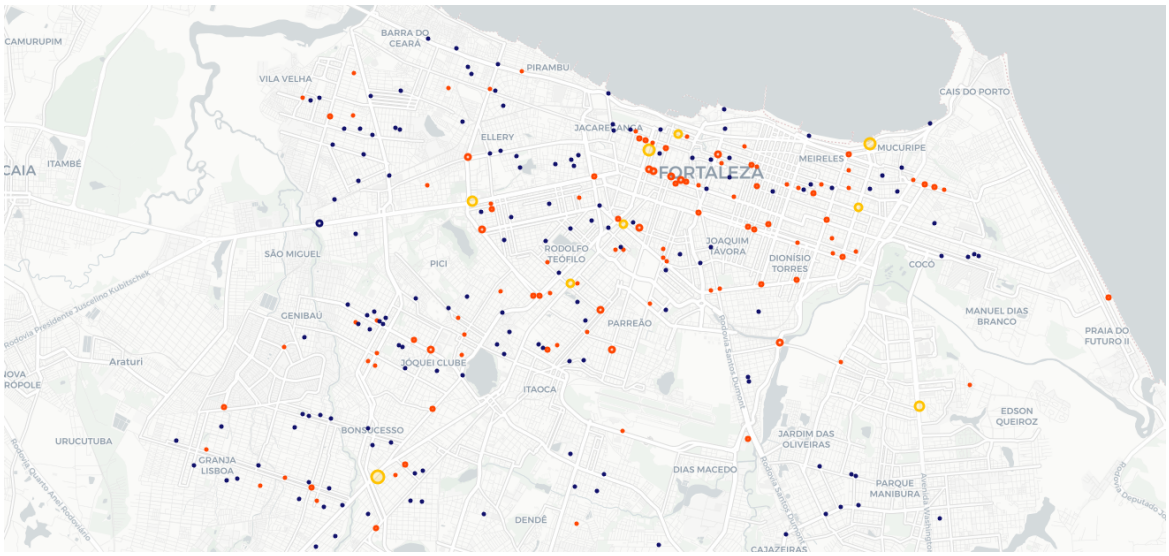
Figura 113: Pontos preditos para Bagging Regressor



Fonte: Autoria própria.

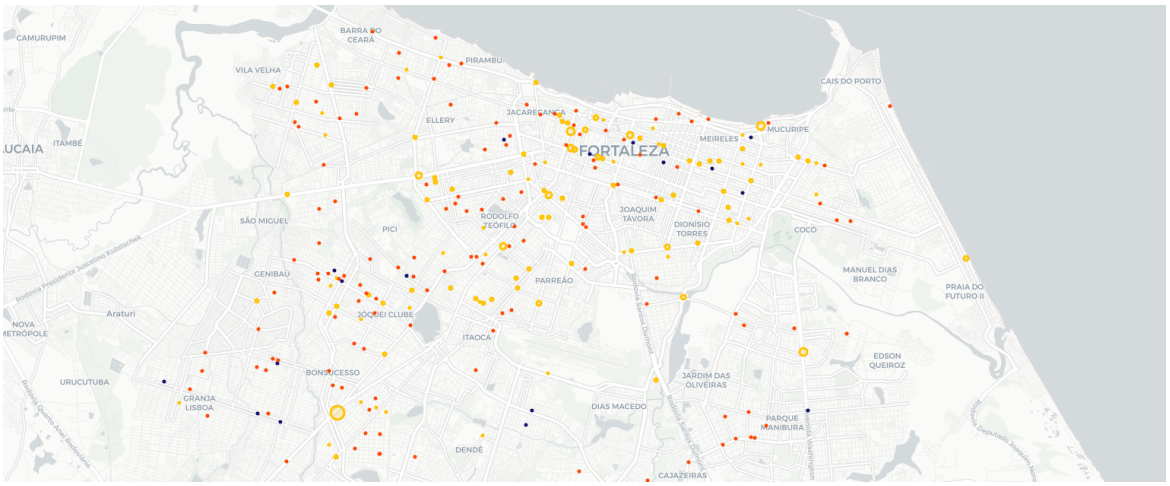
Assim como na categoria 1 desse segundo experimento, na categoria 2 os valores preditos também estão bem distribuídos nos gráficos de dispersão. O método *K-Nearest Neighbor* apresenta as melhores predições, a quantidade de pontos roxos, pontos originais, que estão visíveis é menor que os dos outros métodos. Os pontos preditos para os métodos *Random Forest*, *Extra Trees*, *Decision Tree* e *Bagging Regressor* apresentam predições similares, com poucas variações. O próximo resultado apresentado trata-se dos pontos preditos plotados agora em um mapa da cidade de Fortaleza. Esses mapas são apresentados nas Figuras 115 a 119.

Figura 114: Pontos originais - Mapa da cidade de Fortaleza



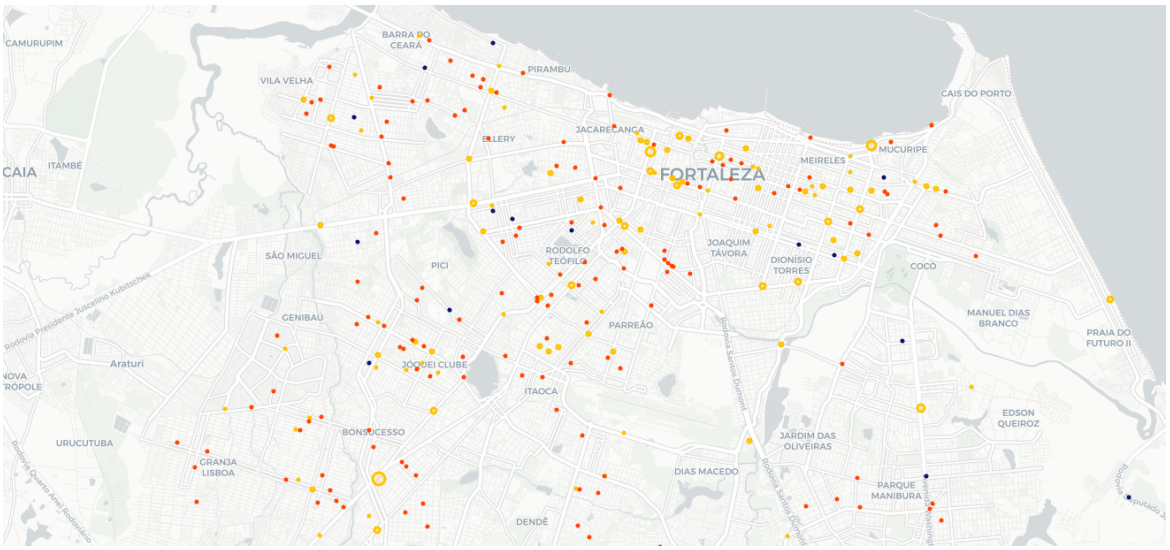
Fonte: Autoria própria.

Figura 115: Crimes preditos - K-Nearest Neighbor Regressor



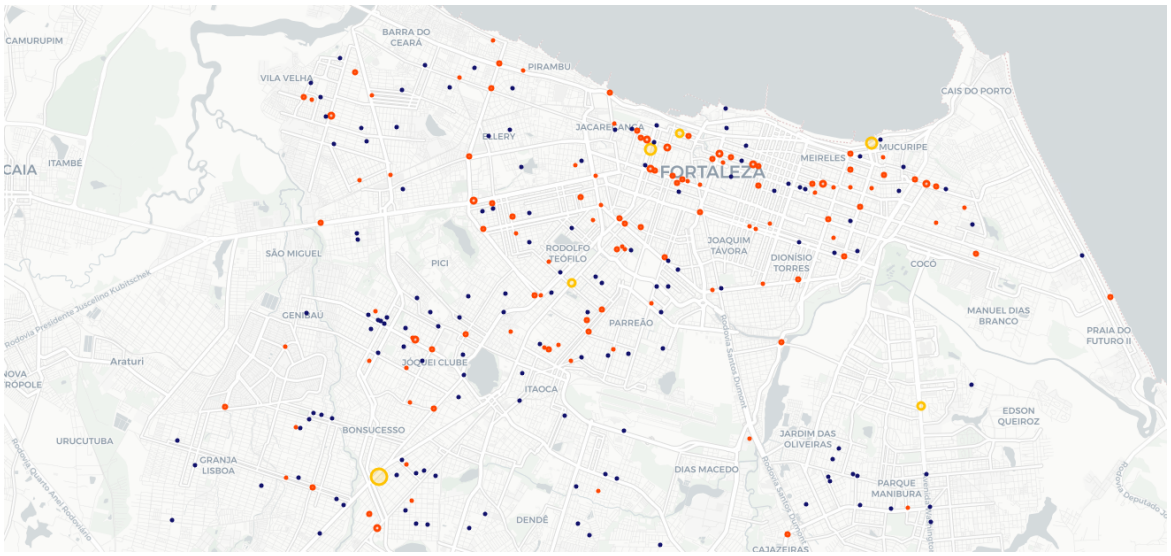
Fonte: Autoria própria.

Figura 116: Crimes preditos - Random Forest Regressor



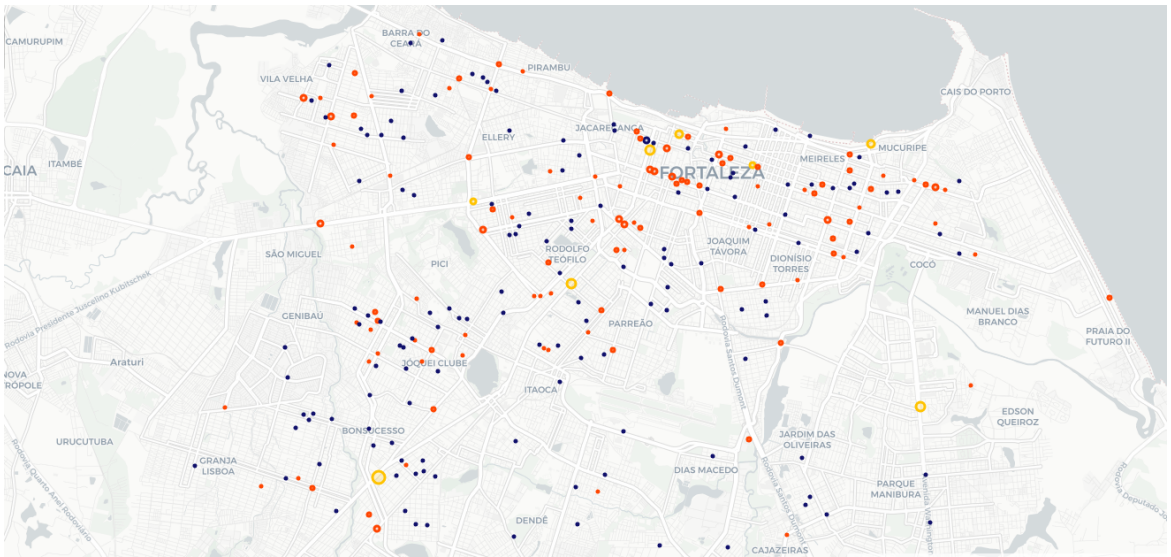
Fonte: Autoria própria.

Figura 117: Crimes preditos - Extra Trees Regressor



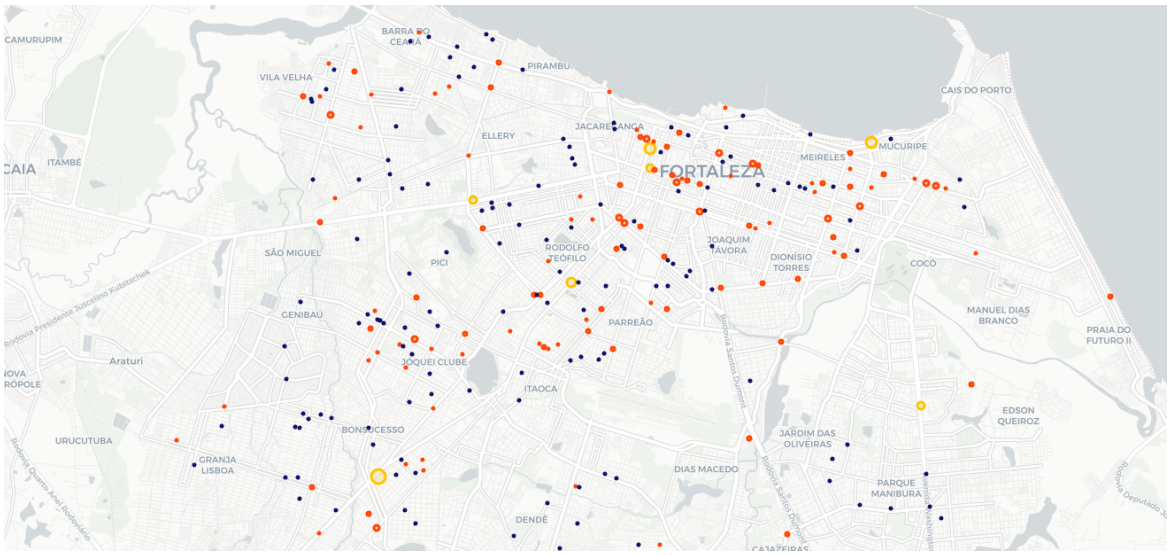
Fonte: Autoria própria.

Figura 118: Crimes preditos - Decision Tree Regressor



Fonte: Autoria própria.

Figura 119: Crimes preditos - Bagging Regressor



Fonte: Autoria própria.

Os métodos *K-Nearest Neighbor* e *Random Forest* obtiveram uma maior quantidade de pontos onde ocorre crimes com maior frequência, as Figuras 115 e 116 exibem esses pontos, de cor amarela. Já os métodos *Extra Trees*, *Decision Tree* e *Bagging Regressor* apresentaram resultados de predição semelhantes entre si e também semelhantes com os pontos originais.

Para a cidade da Filadélfia, Estados Unidos, foi considerado que os crimes ocorreram em diversos pontos de uma rua, também foi considerado que os valores preditos, as saídas do modelo, seriam duas, latitude e longitude. Os resultados apresentados mostram que o método *Decision Tree Regressor* obteve os menores erros, portanto, pode ser considerado como o melhor método para ser utilizado com esses dados. É importante salientar de que o fato de *Decision Tree* ter obtido bons resultados com essa base de dados, não necessariamente significa que também apresentará bons resultados com outras bases.

O uso da base de dados da cidade de Fortaleza foi um pouco mais abrangente. Foram feitos quatro experimentos com esse *dataset*. No experimento 1, é levado em questão que os crimes ocorrem em vários pontos de uma rua e que o modelo deve realizar a predição de dois valores, a latitude e longitude. Os resultados obtidos mostram que o método *Random Forest* obteve as menores taxas de erros e consequentemente obteve as melhores predições. No experimento 2, os crimes ainda ocorrem em vários pontos, entretanto, o modelo deve realizar a predição de apenas um valor, o valor hash, valor esse que representa latitude e longitude. Nesse experimento o método *K-Nearest Neighbor* acaba se destacando, obtendo os menores erros. Comparando os dois experimentos, o primeiro obteve um melhor destaque, principalmente em relação aos pontos preditos no mapa, sendo dessa forma o *Random Forest* o regressor com o melhor resultado para o experimento 1.

A base de dados da cidade de Fortaleza também foi utilizada para realizar os terceiro e quarto experimentos. Na terceira experimentação é considerado que os crimes ocorrem em apenas um ponto da rua e que o modelo deverá realizar a predição de dois valores, latitude e longitude. Nesse experimento, o regressor *K-Nearest Neighbor* também apresenta o menor erro, sendo dessa forma o que apresenta as melhores predições. No quarto experimento, os crimes ainda ocorrem em apenas um ponto da rua, mas agora, o modelo realiza a predição de apenas um valor, novamente o valor *hash*. Para esse novo experimento, o regressor *K-Nearest Neighbor* também obtém os menores erros. Realizando uma comparação entre o terceiro e quarto experimento, nota-se que os resultados obtidos no experimento três foram melhores do que os obtidos no resultado do experimento quatro, principalmente os pontos preditos no mapa.

Comparando-se o experimento 1 com o experimento 3, nota-se que os valores preditos para o terceiro experimento estão mais condizentes com os pontos originais, isso é evidenciado principalmente pelo mapa plotado. Portanto, temos que o regressor *K-Nearest Neighbor* apresenta os melhores valores preditos para esse problema onde os crimes ocorrem em apenas um ponto de uma rua e também onde os valores de saída são latitude e longitude.

4.3 CONSIDERAÇÕES FINAIS

Este capítulo tratou sobre os resultados obtidos para os experimentos utilizando a base de dados da Filadélfia, Estados Unidos, e também da cidade de Fortaleza. Os resultados obtidos permitiram avaliar qual experimento e também quais métodos apresentaram melhores desempenhos.

O próximo capítulo irá tratar sobre as considerações finais do trabalho e também irá apresentar algumas sugestões de trabalhos futuros que podem ser desenvolvidos a partir do que foi mostrado neste.

5 CONCLUSÃO

Neste trabalho foi utilizado uma base de dados da cidade americana Filadélfia, Estados Unidos, e de uma cidade brasileira, Fortaleza. O objetivo principal deste trabalho, que foi a realização de análise de dados criminais e realizar a predição do crime de roubo a pessoa, em região urbana, utilizando técnicas de regressão foi alcançado. Os resultados obtidos mostraram que alguns métodos podem ser utilizados para a criação de modelos preditivos que podem ser colocados em situações reais.

A análise dos dados criminais permitiu verificar padrões em que ocorriam crimes, principalmente em relação a horários e dias da semana. Esse tipo de padrão permite que, por exemplo, um patrulhamento mais inteligente seja criado para determinadas regiões. Apesar de o modelo ter acertado algumas predições, é necessário haver sempre uma periodicidade na obtenção de novos dados, visto que o padrão dos crimes sempre irá mudar.

As técnicas de regressão utilizadas neste trabalho obtiveram ótimos resultados na realização da predição dos crimes. Diferente de outros métodos, como *deep learning* e *boosting*, as técnicas de *K-Nearest Nighbor*, *Random Forest*, *Extra Trees*, *Decision Tree* e *Bagging* não precisaram de muitos parâmetros para serem treinadas, isso permitiu maior tempo na modelagem dos dados do que propriamente no treinamento dos modelos.

Com os vários resultados gerados e computados, houve a necessidade de torná-los mais visuais, facilitando dessa forma para qualquer público interpretar e visualizar de forma mais rápida o que aqueles dados representam. Neste trabalho foi gerado gráficos de plotagem dos pontos preditos, também foram gerados mapas com as predições de crimes para as cidade da Filadélfia, Estados Unidos, e de Fortaleza. Os mapas permitem verificar em quais áreas há uma maior incidência de crimes.

5.1 TRABALHOS FUTUROS

Para projetos futuros, algumas sugestões são válidas. Deve-se utilizar novas variações de dados para grupos de treino e teste para as técnicas de regressão selecionadas. Outro ponto que deverá ser explorado futuramente trata sobre a avaliação métodos *boosting* com a base de dados da cidade de Fortaleza. Somado a esses novos métodos, há também a possibilidade da análise de regressão com outros tipos de crimes, como homicídios, por exemplo. Outra medida a ser avaliada como tratativa futura é a utilização de *hot points*, como bancos, *shopping*, farmácias, praças para verificar se esses locais estão ligados a ocorrências de crimes.

REFERÊNCIAS

Abbass, Z.; Ali, Z.; Ali, M.; Akbar, B.; Saleem, A. A Framework to Predict Social Crime through Twitter Tweets By Using Machine Learning. *IEEE 14th International Conference on Semantic Computing (ICSC)*, 2020.

Araújo, A.; Borges, J.; Bezerra, L.; Vieira, C. Towards a Crime Hotspot Detection Framework for Patrol Planning. *20th International Conference on High Performance Computing e Communications*, 2018.

Bravo, S.; Moreno, A. A Random Forest Approach for Predicting the Microwave Drying Process of Amaranth Seed. *2nd International Conference on Information e Computer Technologies*, 2019.

Breiman, L. Random Forests. 2001.

Buhlmann, P. Bagging, Boosting e Ensemble Methods. *Hebook of Computational Statistics: Concepts e Methods*, 2012.

Capellán, E.; Otero, M.C. Using Artificial Intelligence on Crime Prediction: PREDPOL. 2017.

Cerqueira, D.; Bueno, S.; Lima, R.; Neme, C.; Ferreira, H.; Alves, P.; Marque, D.; Reis, M.; Cypriano, O.; Sobral, I.; Pacheco, D.; Lins, G.; Armstrong, K. Atlas da Violência. *Forúm Brasileiro de Segurança Pública*, 2019.

Chakure, A. Random Forest Regression. 2020. Disponível em: <https://towardsdatascience.com/random-forest-e-its-implementation-71824ced454f>. Acesso em: 9 de maio de 2020.

Cutler, A.; Cutler, D.; Stevens, J. Random Forests. *Machine Learning*, 2011.

Dash, S.; Safro, I.; Sakrepatna, R. Spatio-temporal prediction of crimes using network analytic approach. *IEEE International Conference on Big Data (Big Data)*, 2018.

Dezhic, E. Understeing Decision Trees. 2020. Disponível em: <https://becominghuman.ai/understeing-decision-trees-43032111380f>. Acesso em: 9 de maio de 2020.

Dineva, K.; Atanasova, T. OSEMN Process for working over data acquired by IOT devices mounted in beehives. *Current Trends in Natural Sciences*, 2018.

Dutt, R.; Krishna, V. Forecasting the Grant Duration of a Patent using Predictive Analytics. *International Journal of Computer Applications*, 2019.

GeoPy. GeoLocator. 2020. Disponível em: <https://pypi.org/project/geopy/>. Acesso em: 10 de maio de 2020.

Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. 2005.

Google. Google Maps Platform. 2020. Disponível em: <https://developers.google.com/maps/documentation/geolocation/overview>. Acesso em: 10 de maio de 2020.

Igarapé, Instituto. CrimeRadar. 2020. Disponível em: <https://rio.crimeradar.org/about>. Acesso em: 10 de maio de 2020.

Ingilevich, V.; Ivanov, S. Crime rate prediction in the urban environment using social factors. *International Young Scientist Conference on Computational Science*, 2018.

Jaiswal, J.; Samikannu, R. Application of Random Forest Algorithm on Feature Subset Selection e Classification e Regression. *World Congress on Computing e Communication Technologies*, 2017.

John, V.; John, N., Karunakaran; Guo, C. Free Space, Visible e Missing Lane Marker Estimation using the PsiNet e Extra Trees Regression. *24th International Conference on Pattern Recognition*, 2018.

Kaggle. Philadelphia Crime Data. 2020. Disponível em: <https://www.kaggle.com/mchirico/philadelphiacrimedata>. Acesso em: 16 de maio de 2020.

Kim, S.; Joshi, P.; Singh, P.; Taheri, P. Crime Analysis Through Machine Learning. *IEEE 9th Annual Information Technology, Electronics e Mobile Communication Conference (IEMCON)*, 2018.

Kumar, T. Solution of linear e non linear regression problem by K Nearest Neighbour approach. *IEEE International Conference on Computational Intelligence & Communication Technology*, 2015.

Madhuri, R.; G, A.; Pujitha, M. House Price Prediction Using Regression Techniques: A Comparative Study. *6th International Conference on smart structures e systems*, 2019.

Ortiz-Bejar, J.; Graff, M.; Tellez, E.; Ortiz-Bejar, J. K-Nearest Neighbor Regressors Optimized by using Random Search. *International Autumn Meeting on Power, Electronics e Computing*, 2018.

Pinto, J.; Kelur, S.; Shetty, J. Iris flower species identification using maching learning approach. *4th International Conference for Convergence in Technology*, 2018.

Rathan, K.; Sai, S.; Manikanta, T. Crypto-Currency price prediction using Decision Tree e Regression techniques. *3rd International Conference on Trends in Electronics e Informatics*, 2019.

Sathyadevan, S.; Gangadhara, S. Crime Analysis e Prediction Using Data Mining. *First International Conference on Networks e Soft Computing*, 2014.

SSP. Detecta. 2020. Disponível em: <http://www.ssp.sp.gov.br/acoes/leAcoes.aspx?id=33833>. Acesso em: 10 de maio de 2020.

Vidhya, Analytics. KNN Regressor. 2020. Disponível em: <https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/>. Acesso em: 10 de maio de 2020.

Xu, Y.; Fu, C.; Kennedy, E.; Jiang, S.; Owusu-Agyemang, S. The impact of street lights on spatial-temporal patterns of crime in Detroit. 2017.