Grupo 3

Lucas de Miranda Oliveira[1]
Pedro Henrique Sebe Rodrigues[2]
Renata Massami Hirota[3]
Rodolfo Riani Sundfeld[4]
Rubens Santos Andrade Filho[5]

Dezembro de 2020

Sumário

[1]Número USP: 11577209
[2]Número USP: 10256915
[3]Número USP: 7165654
[4]Número USP: 8535770
[5]Número USP: 10370336

1 Descrição do Problema

O Brasil tem a segunda maior taxa de homicídios da América do Sul, segundo relatório das Nações Unidas em 2017, e ocupa a 12ª posição no ranking global de violência, com uma taxa de 27,5% de homicídios a cada 100 000 habitantes. Em 2016, a Organização para a Cooperação e Desenvolvimento Econômico (OCDE) divulgou que as vítimas de assaltos no Brasil, nos últimos 12 meses da pesquisa, foram de 7,9% dos entrevistados, quase o dobro da média dos países em que a pesquisa foi realizada. Este preocupante quadro mostra a situação da violência e a necessidade de políticas públicas adequadas para segurança pública.

A partir de dados de boletins de ocorrência (BO) da Secretaria de Seguraça Pública (SSP) do estado de São Paulo, o objetivo deste estudo é prever o horário do crime de acordo com algumas características, por exemplo, localização, sexo da vítima, cor da vítma, entre outros. Visando, dessa forma, identificar quais os horários das ocorrências de acordo com a tipologia do crime e assim fornecer insights para o policiamento.

2 Conjunto de dados

A base dados que foi utilizada é a Crime Data in Brazil e pode ser acessada **aqui**.

A base de dados que será trabalhada refere-se a boletins de ocorrência na Grande São Paulo.

Variável	Descrição
ID_DELEGACIA	Código da delegacia responsável pelo registro da ocorrencia
NOME_DEPARTAMENTO	Departamento responsável pelo registro
NOME_SECCIONAL	Delegacia Seccional responsável pelo registro
NOME_DELEGACIA	Delegacia responsável pelo registro
CIDADE	Cidade de Registro
ANO_BO	Ano da ocorrencia
NUM_BO	Número do BO
NOME_DEPARTAMENTO_CIRC	Departamento de Circunscrição
NOME_SECCIONAL_CIRC	Seccional de Circunscrição
NOME_DELEGACIA_CIRC	Delegacia de Circunscrição
NOME_MUNICIPIO_CIRC	Município de Circunscrição
DESCR_TIPO_BO	Tipo de Documento
DATA_OCORRENCIA_BO	Data da Ocorrência
HORA_OCORRENCIA_BO	Hora da Ocorrência
DATAHORA_COMUNICACAO_BO	Data Hora da Comunicação da Ocorrência
FLAG_STATUS	Status da Ocorrência
RUBRICA	Natureza jurídica da ocorrência
DESCR_CONDUTA	Conduta na Ocorrência
DESDOBRAMENTO	Desdobramento na Ocorrência
DESCR_TIPOLOCAL	Tipo de Local
DESCR_SUBTIPOLOCAL	Descrição do subTipo de local
LOGRADOURO	Logradouro dos fatos
NUMERO_LOGRADOURO	Numero do Logradouro dos fatos
LATITUDE	Latitude da Ocorrência
LONGITUDE	Longitude da Ocorrência
DESCR_TIPO_PESSOA	Qualificação do envolvido na ocorrência
FLAG_VITIMA_FATAL	Condição do Autor / Vítma na corrência
SEXO_PESSOA	Sexo
IDADE_PESSOA	Idade
COR_CUTIS	Cor da Pele

2.1 Possíveis variáveis de interesse

- Tipo de local da ocorrência
- Localização (latitude e longitude) da ocorrência
- Sexo da vítima
- Idade da vítima
- Tipificação do crime

```
## PhantomJS not found. You can install it with webshot::install_phantomjs(). If it is installed, pleas
```

3 Avaliação do problema

A variável HORA_OCORRENCIA_BO será transformada de forma a produzir quatro classes de acordo com o horário do crime: manhã, tarde, noite e madrugada. O problema é de classificação.

Os métodos de aprendizagem supervisionada para classificação utilizados serão:

- Multinomial logistic regression;
- Support Vector Machine Multiclass;
- K-Nearest Neighbor(KNN) Classification;
- Feedforward Neural Network For Multiclass Classification;
- Random Forest;
- Boosting;
- Bagging;
- Gradient Boosting.

4 Análise Descritiva

4.0.1 Tipo de local da ocorrência

4.0.2 Localização

4.0.3 Sexo e idade das vítimas

4.0.4 Tipificação

4.0.5 Distribuição da idade por cor

4.0.6 Vítimas fatais

4.0.7 Horário e dia da semana

5 Métodos utilizados

- Regressão Logística Multinomial

- Árvore de decisão
- Random forest
- KNN - K-Nearest Neighbor Classification
- XGBoost

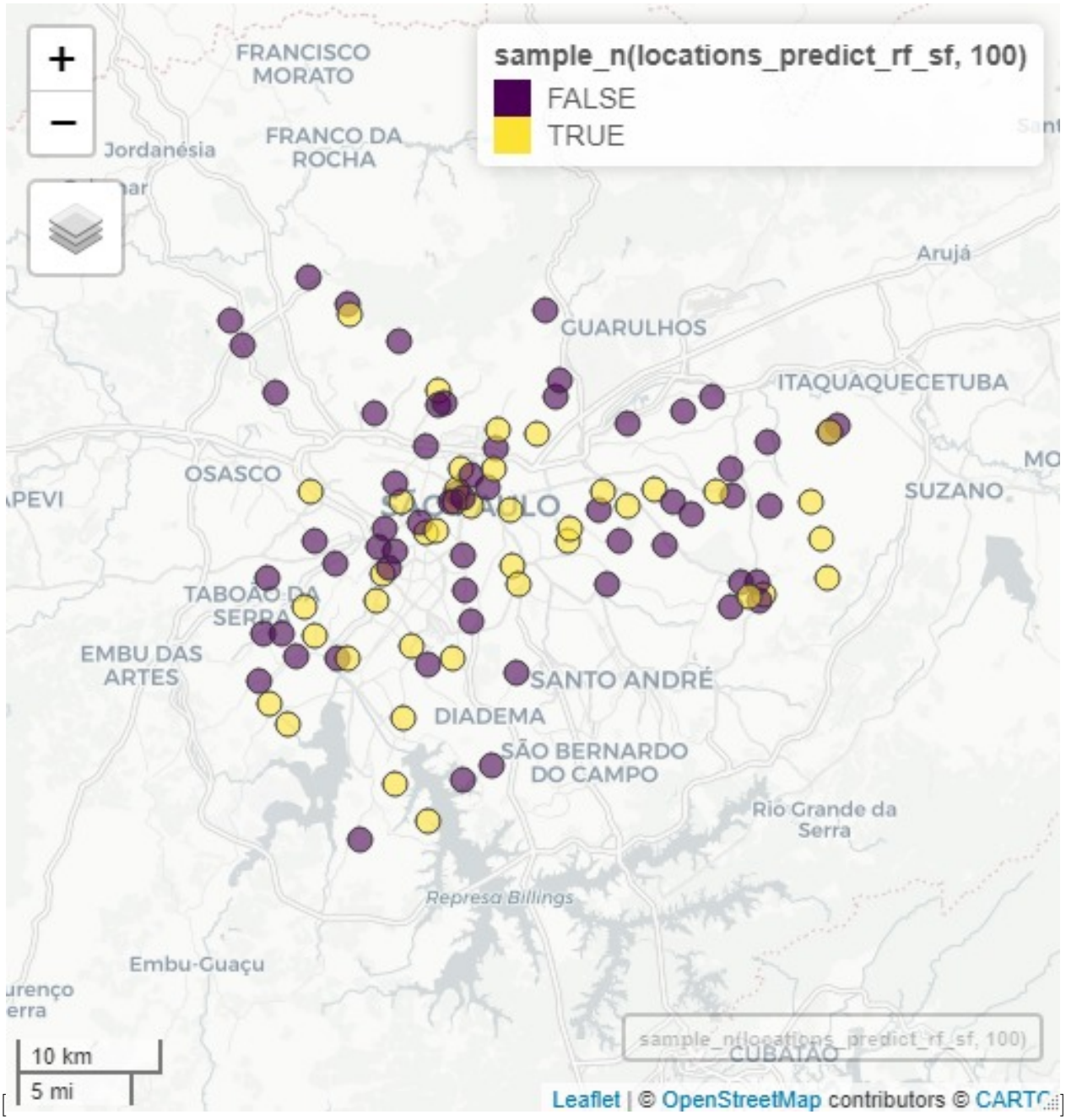
5.1 Regressão Logística Multinomial

- É usada para prever a categoria ou a probabilidade de uma categoria em uma variável dependente com base em várias variáveis independentes.
- As variáveis independentes podem ser dicotômicas (ou seja, binárias) ou contínuas (ou seja, intervalo ou razão em escala).
- É uma extensão da regressão logística binária que permite mais de duas categorias da variável dependente.

5.2 Random Forest

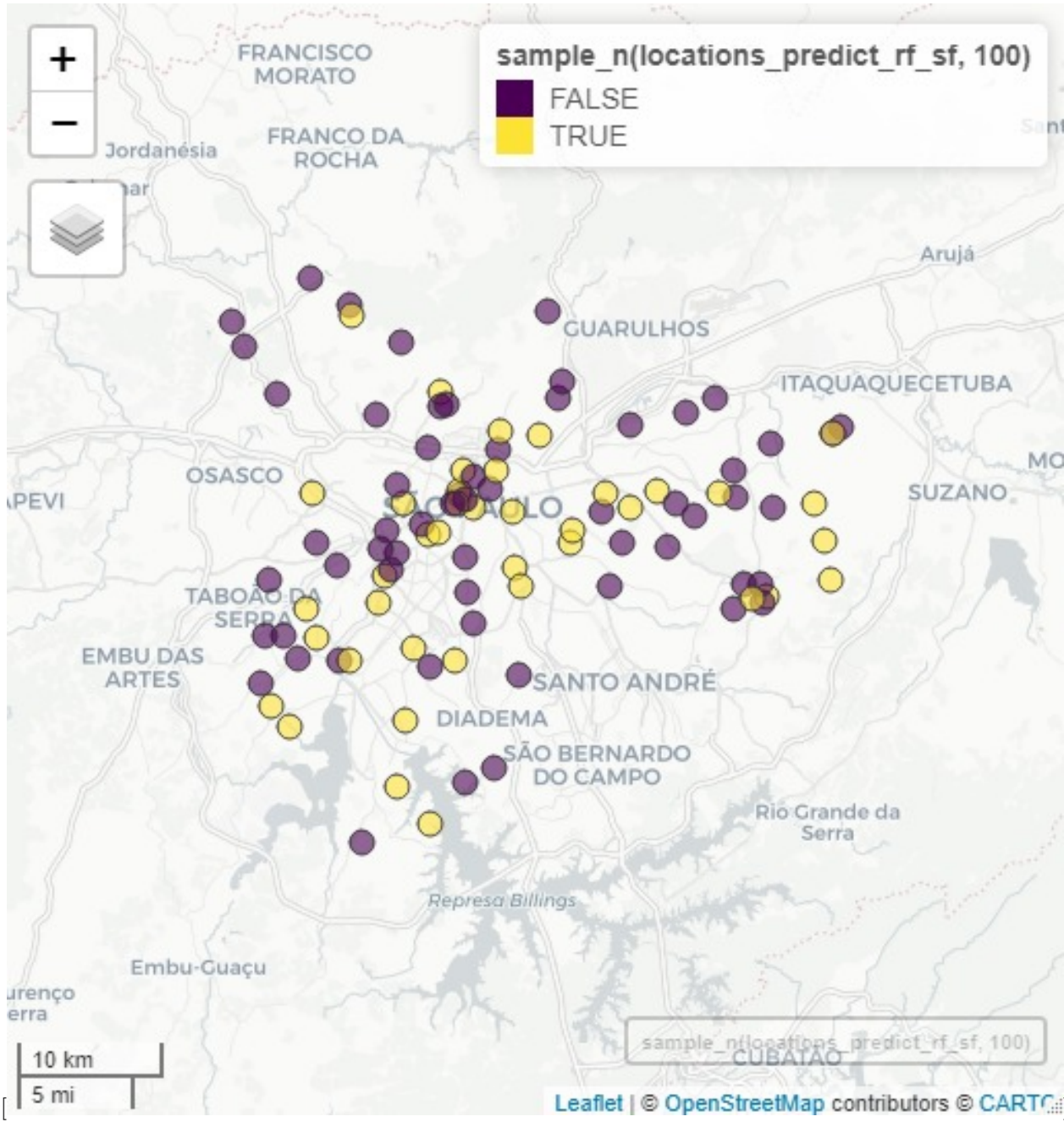
- É um método de aprendizagem de conjunto para classificação, regressão e outras tarefas.
- Constroi uma infinidade de árvores de decisão no momento do treinamento
- Gera a categorias que é a moda das categorias, no caso de classificação.

    – - Corrigem o hábito das árvores de decisão de sobreajustar os dados de treinamento.

.center[
]

5.3 KNN - K-Nearest Neighbor Classification

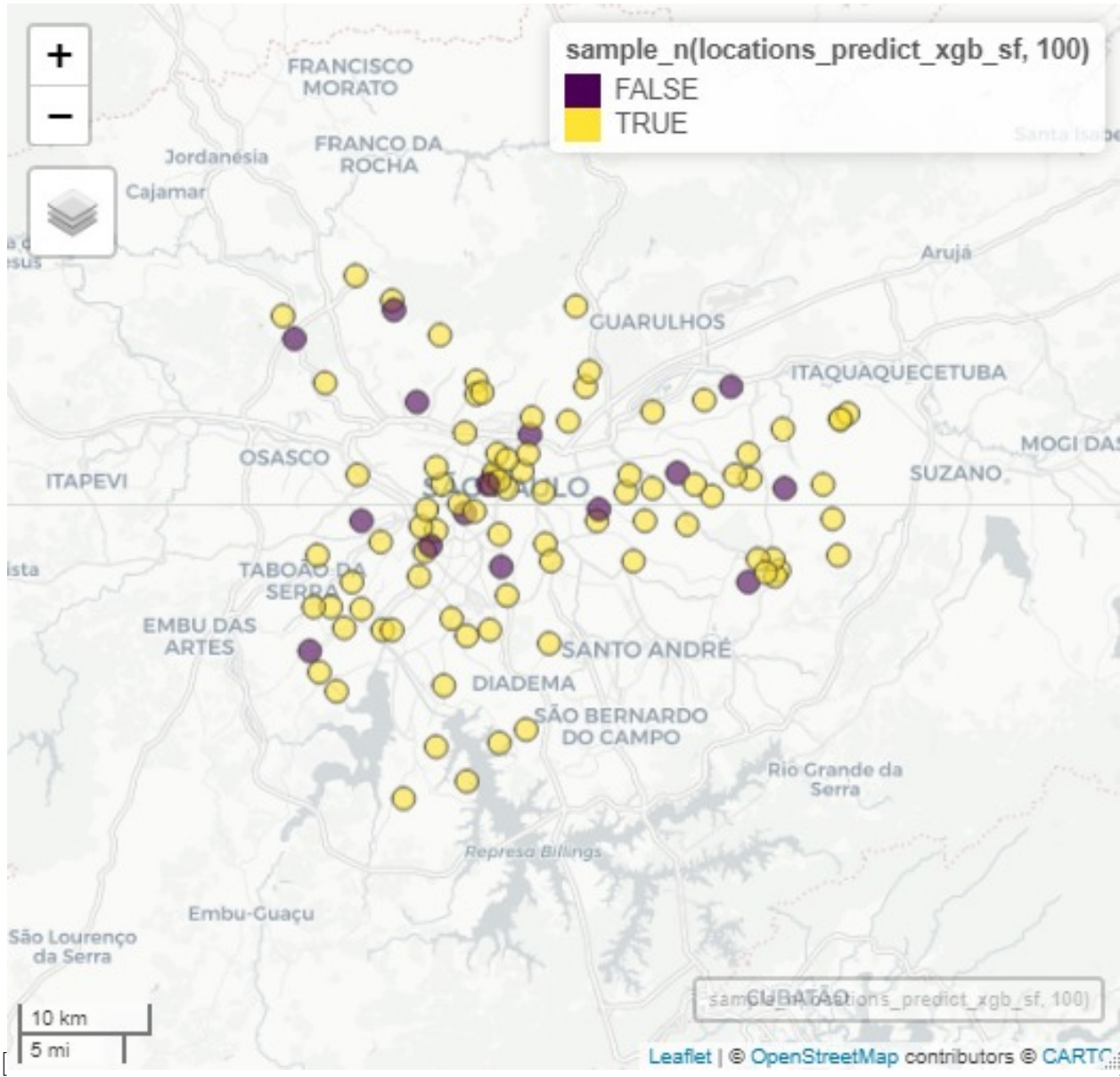
- Na classificação k-NN, a saída é uma associação de classe.
- Um objeto é classificado por uma pluralidade de votos de seus vizinhos, com o objeto sendo atribuído à classe mais comum entre seus k vizinhos mais próximos
- Se k = 1, então o objeto é simplesmente atribuído à classe daquele único vizinho mais próximo.
- Valores maiores de k reduzem o efeito do ruído na classificação, mas tornam os limites entre as classes menos distintos.



.center[

5.4 XGBoost

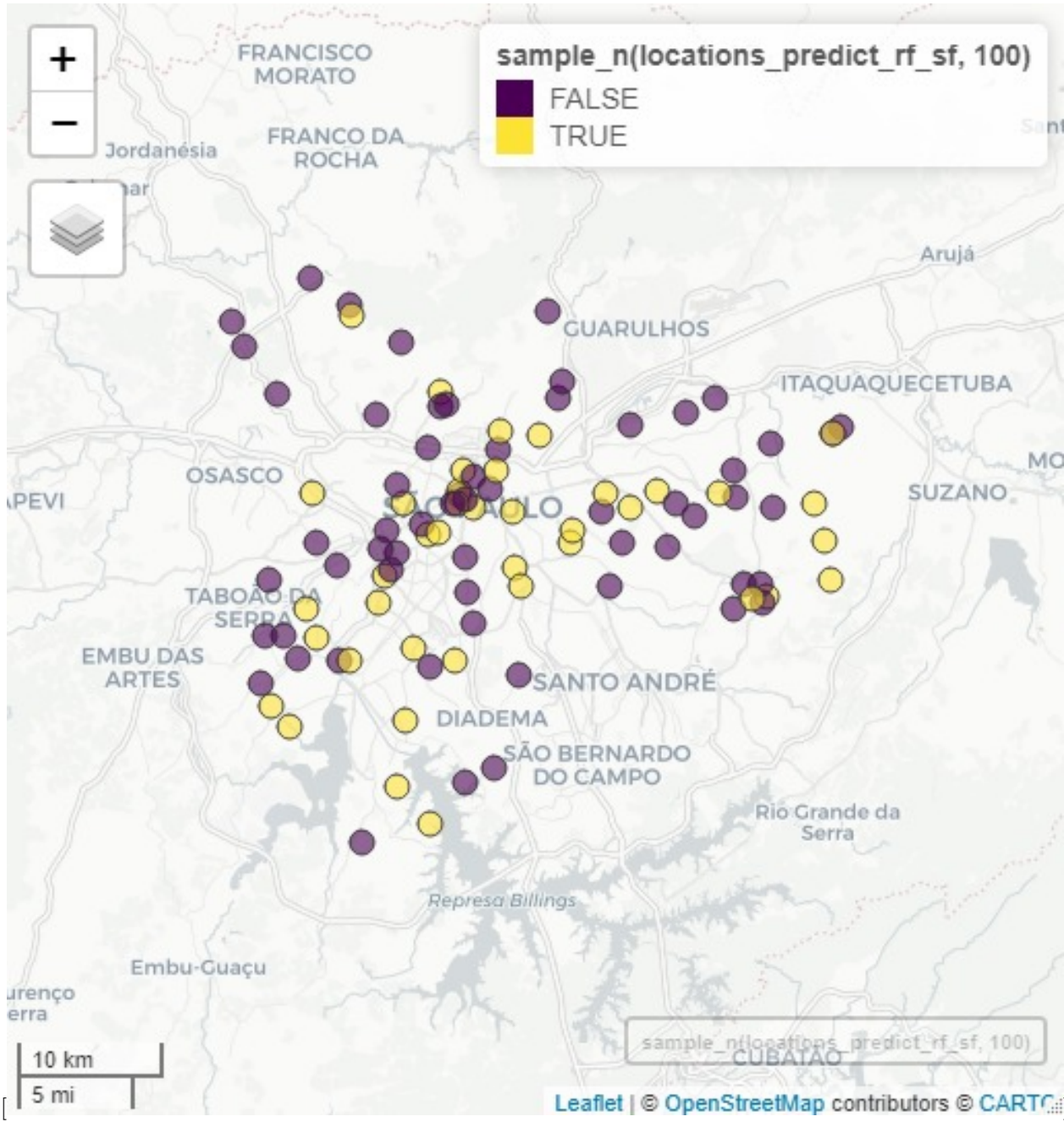
- XGBoost é um algoritmo de aprendizado de máquina baseado em árvore de decisão que usa uma estrutura de gradiente *boosting*.
- Regularização: penaliza modelos mais complexos por meio da regularização LASSO (L1) e Ridge (L2) para evitar *overfitting*.
- Validação cruzada: O algoritmo vem com método de validação cruzada embutido em cada iteração, eliminando a necessidade de programar explicitamente essa pesquisa e especificar o número exato de iterações de reforço necessárias em uma única execução.

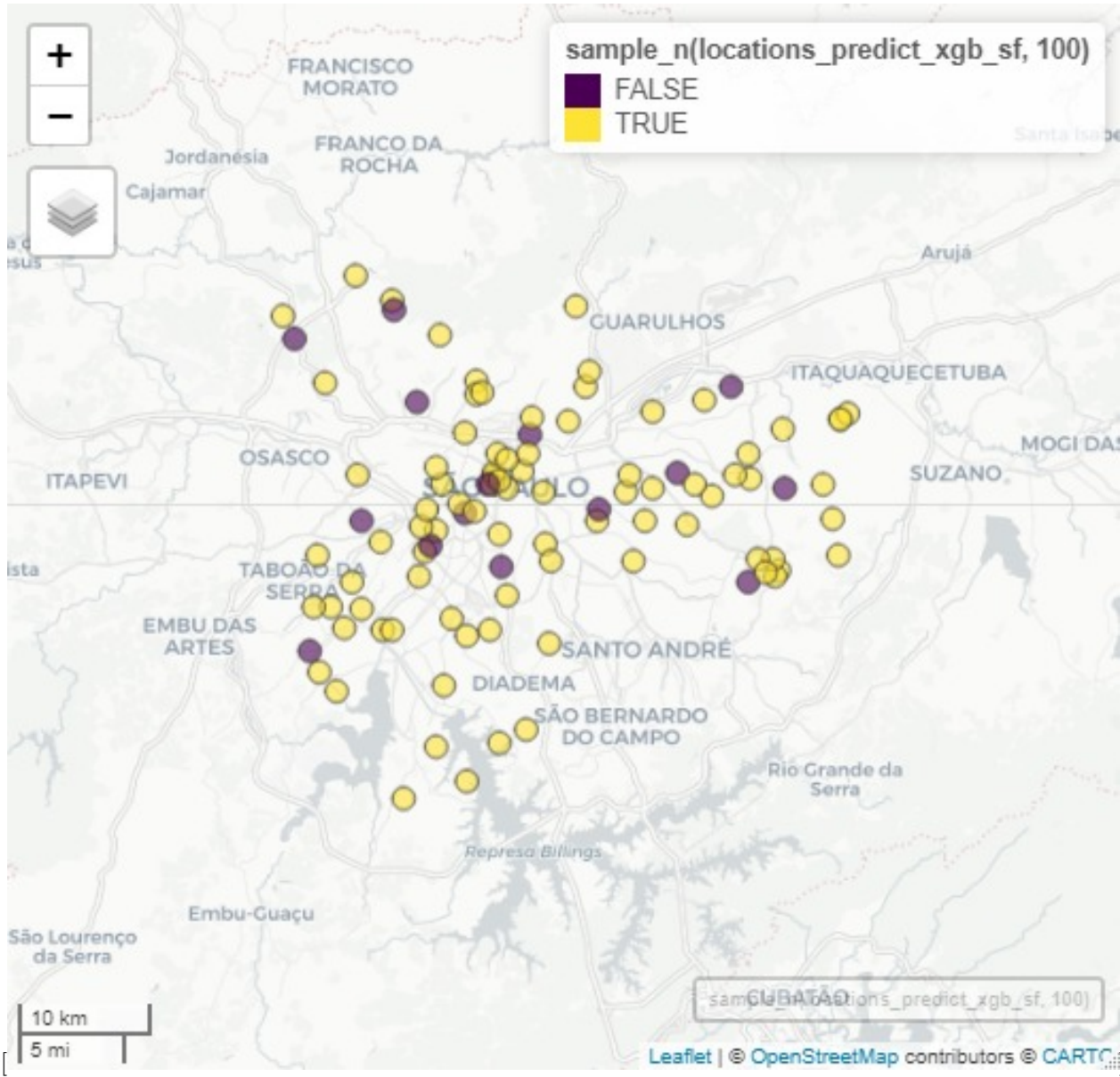


.center[

6 Resultados

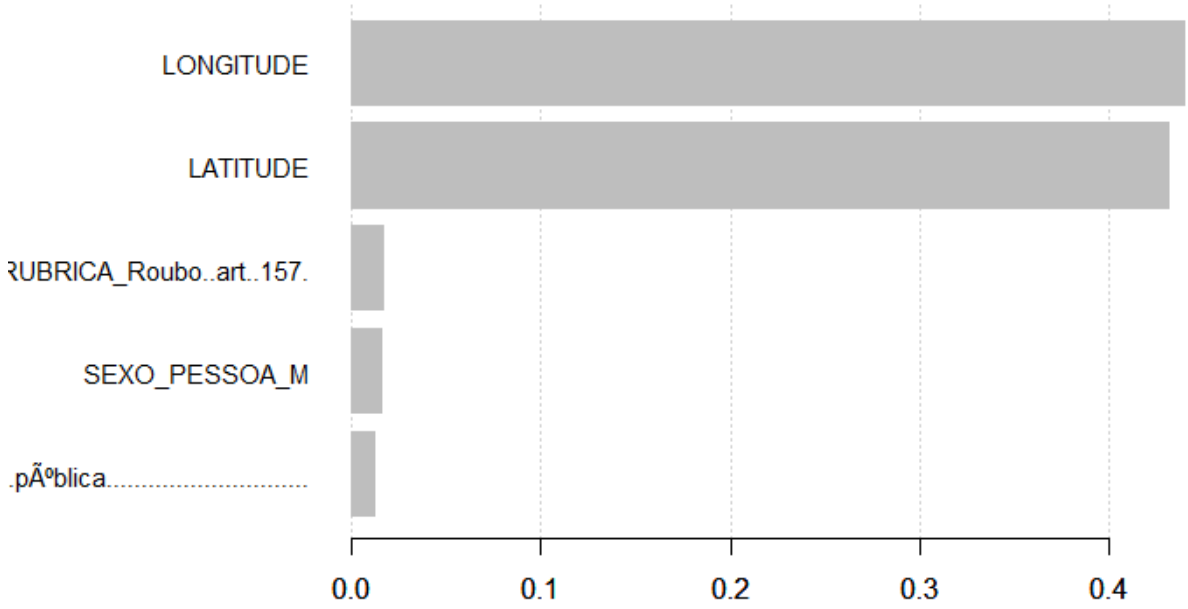
Método	Acurácia Treino	Acurácia Teste
Regressão Logística Multinomial	0.322	0.328
Random Forest	0.362	0.360
KNN	0.609	0.569
XGBoost	0.845	0.837

.center[
sample_n(locations_predict_rf_sf, 100)
FALSE
TRUE
FRANCISCO MORATO
FRANCO DA ROCHA
Jordanésia
Arujá
GUARULHOS
ITAQUAQUECETUBA
OSASCO
SUZANO
TABOÃO DA SERRA
EMBU DAS ARTES
SANTO ANDRÉ
DIADEMA
SÃO BERNARDO DO CAMPO
Rio Grande da Serra
Represa Billings
Embu-Guaçu
CUBATÃO
10 km
5 mi
Leaflet | © OpenStreetMap contributors © CARTO

.center[

7 Conclusão

- Variáveis espaciais, LAGITUDE e LONGITUDE, são as que mais apresentaram relevância nos modelos ajustados.



.center[

- Variáveis espaciais, LAGITUDE e LONGITUDE, são as que mais apresentaram relevância nos modelos ajustados.
- Dos métodos ajustados, XGBoost foi o que apresentou a melhor acurácia após otimização de hiperparâmetros, seguido pelo KNN.

8 Referências

BREIMAN, Leo. Bagging predictors. Machine learning, v. 24, n. 2, p. 123-140, 1996.

BREIMAN, Leo. Random forests. Machine learning, v. 45, n. 1, p. 5-32, 2001.

FRIEDMAN, Jerome H. Stochastic gradient boosting. Computational statistics & data analysis, v. 38, n. 4, p. 367-378, 2002.

HEARST, Marti A.. et al. Support vector machines. IEEE Intelligent Systems and their applications, v. 13, n. 4, p. 18-28, 1998

SVOZIL, Daniel; KVASNICKA, Vladimir; POSPICHAL, Jiri. Introduction to multi-layer feed-forward neural networks. Chemometrics and intelligent laboratory systems, v. 39, n. 1, p. 43-62, 1997.