Técnicas de Machine Learning aplicadas na predição da quantidade de leitos públicos nos municípios do Brasil

Bruno Almeida de Carvalho[1] **e Ricardo Ferreira da Rocha**[1]

[1] Departamento de Estatística, Universidade Federal da Bahia, Brasil

Resumo

O presente trabalho tem como objetivo a aplicação de técnicas de Aprendizado de Máquina para a predição da quantidade de leitos públicos nos municípios do Brasil. A base utilizada é composta por informações extraídas do CNES (Cadastro Nacional de Estabelecimentos de Saúde) e do IBGE (Instituto Brasileiro de Geografia e Estatistica). Os modelos de regressão linear aplicados no estudo foram: ”Linear Regression”, ”Ridge”, ”Lasso”, ”Elastic Net”, ”Decision Tree Regressor”, ”Random Forest Regressor” e ”KNN Regressor”.

Keywords: *Aprendizado de Máquina; Regressão.*

1. Introdução

A Idéia do trabalho surgiu após o contato com a base de dados do CNES (Cadastro Nacional de Estabelecimentos de Saúde). Explorando o seu conteúdo, percebeu-se que tratava-se de um interessante raio-x das Unidades Básicas de Saúde do Brasil, trazendo informações como: tipos de Unidades Básicas de Saúde, bem como a quantidade de leitos públicos de cada estabelecimento. A partir daí, como uma forma de incrementar a base, foram inseridas variáveis decorrentes da extração de dados do IBGE (Instituto Brasileiro de Geografia e Estatistica).

2. A proposta

O Estudo tem como grande desafio criar um modelo de aprendizado de máquina capaz de prever a quantidade de leitos em municípios onde não há leitos, utilizando como base-treino a porção da base de dados onde os municípios possuem leitos. A relevância do tema reside em estabelecer um modelo que possa prever a quantidade de leitos que determinados municípios deveriam ter com base em caraterísticas como: região, estado, população, área, densidade demográfica, latitude, longitude e tipos de Unidades Básicas de Saúde e suas quantificações.

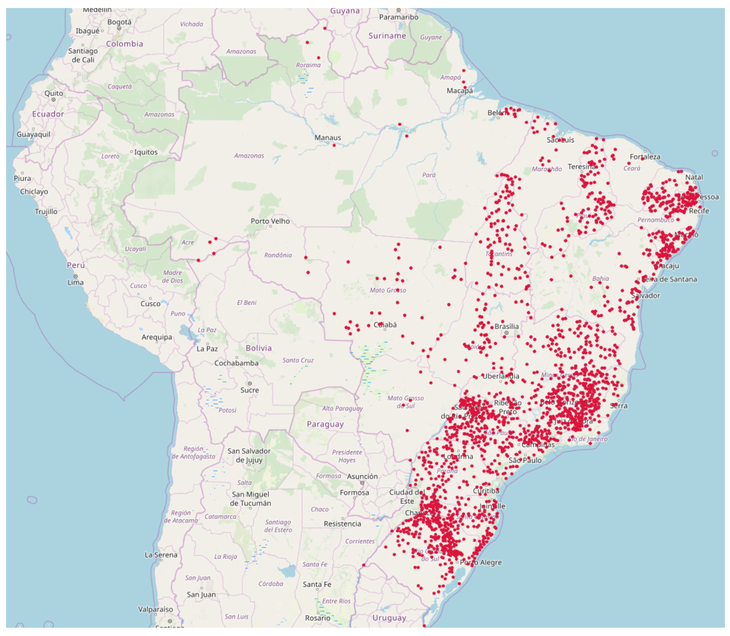

Figure 1: *Municípios do Brasil que não possuem leitos.*

Com a análise dos dados percebeu-se que dos 5.570 municípios analisados, 1.864 não possuíam sequer um leito. Na Figura 1 cada ponto vermelho representa um município sem leitos. É possível observar que a indisponibilidade de leitos está concentrada na porção sul e sudeste do país.

Na Figura 2 verifica-se a distribuição de leitos por região, bem como a linha média de leitos por região, a passo que na Figura 3 apresenta-se a distribuição de leitos por estado, bem como a respectiva linha média.

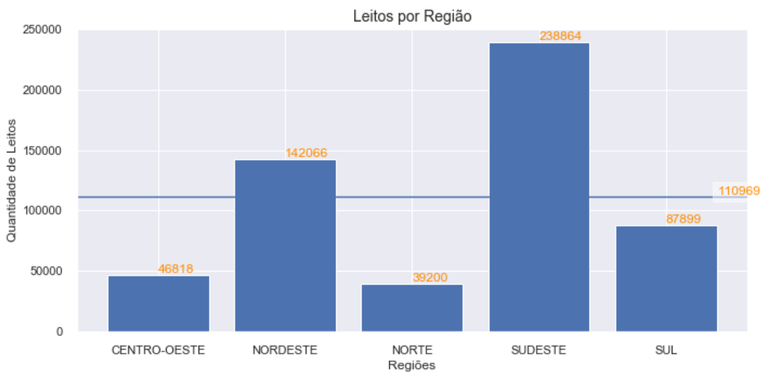

Figure 2: *Quantidade de leitos por Região.*

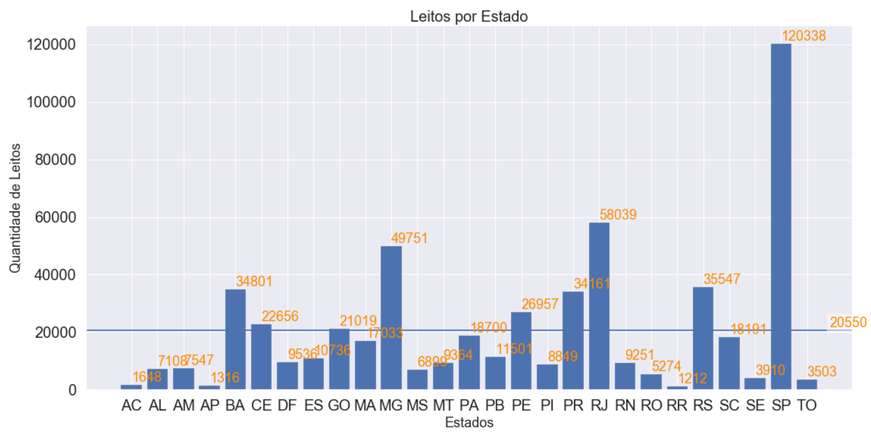

Figure 3: *Quantidade de leitos por Estado.*

3. Aplicação

DEVIDO a limitações de processamento pelo fato de existirem 50 variáveis preditoras, seria custoso computacionalmente falando analisar todas as possibilidades agrupamento entre as 50 variáveis e por tanto foi decidido considerá-las em sua totalidade. O próximo passo, então, seria a aplicação de modelos de regressão linear como: ”Linear Regression”, ”Ridge”, ”Lasso”, ”Elastic Net”, ”Decision Tree Regressor”, ”Random Forest Regressor” e ”KNN Regressor” (Figura 4). Foi utilizado como método de validação, a validação cruzada (cross validation), que é utilizada para detectar problema de conjuntos de testes selecionados incorretamente. O número de ”k-folds” (ou dobras) utilizado no estudo foi igual a 10 (sem repetições). A métrica de validadação para a escolha do melhores modelos de Machine Learning foi feita com cálculo do erro absoluto médio (MAE).

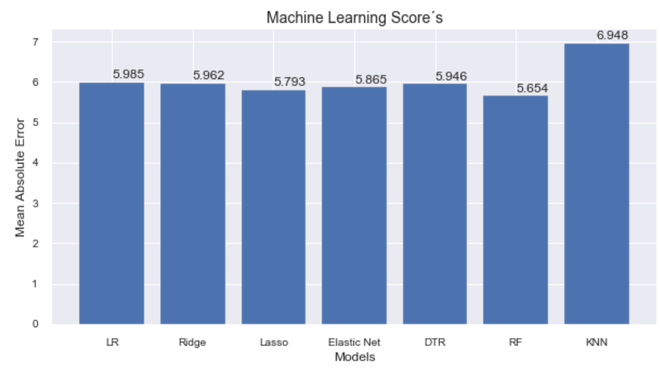

Figure 4: *Resultado final das regressões.*

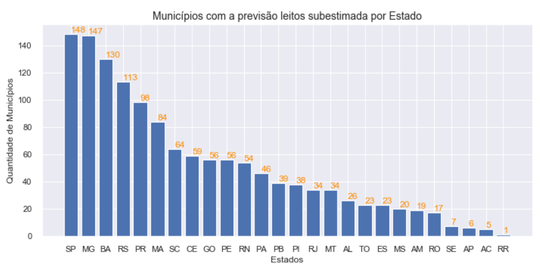

Figure 5: *Superestimações por estado*

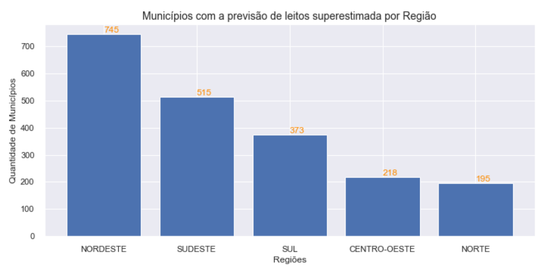

Figure 6: *Superestimações por região*

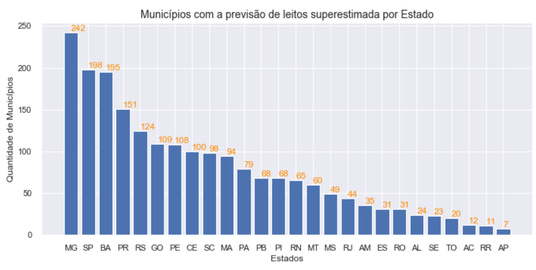

Figure 7: *Subestimações por estado*

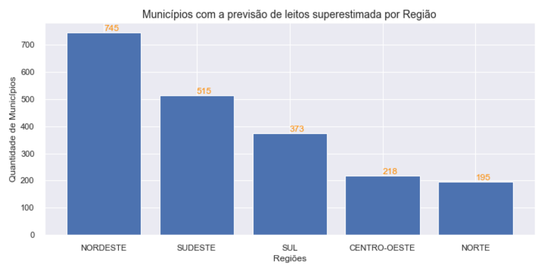

Figure 8: *Subestimações por região*

Conclusões

DIANTE da exploração dos dados, conclui-se que o modelo, em linhas gerais, considerou que diante das características físicas e estruturais os municípios ”deveriam” possuir mais leitos. O modelo em geral superestimou a quantidade de leitos dos municípios, possível que a discrepância na quantidade de leitos nos municípios de maior população e estrutura estejam influênciando os que possuem menor população e estrutura. Como foi também observado ao longo do presente trabalho, variáveis como: população estimada, centro de atenção psicossocial, clínica centro de especialidade, consultório isolado, hospital geral, unidade de apoio diagnose e terapia isolado e a quantidade de unidades básicas guardam forte correlação com a quantidade de leitos. Utilizando como ponto de partida o que já foi produzido no presente trabalho, seria interessante no futuro incrementar a base de dados com indicadores como IDH, escolaridade e renda, como forma de melhorar a performace do modelo e extrair mais conhecimento dos dados.