PREVISÃO DE NOTIFICAÇÕES DE GRIPE UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

Antônio Carlos e Gustavo Ferreira

1. Introdução

A gripe é uma infecção viral altamente transmissível que afeta milhões de pessoas todos os anos, representando um desafio significativo para os sistemas de saúde pública. Além de causar impactos diretos na saúde da população, como internações e complicações em grupos de risco, surtos gripais também geram consequências econômicas, afetando a produtividade e aumentando a demanda por recursos hospitalares. Diante desse cenário, a capacidade de prever a incidência de casos gripais é essencial para que governos possam planejar e otimizar suas respostas, garantindo a alocação eficiente de leitos, vacinas e medidas preventivas.

Nos últimos anos, avanços em inteligência artificial (IA) têm demonstrado grande potencial para auxiliar na tomada de decisão em saúde pública. Modelos de aprendizado de máquina e técnicas de previsão de séries temporais permitem a identificação de padrões em dados históricos, possibilitando estimativas mais precisas sobre o comportamento futuro de epidemias sazonais. Com essas ferramentas, gestores podem antecipar surtos e implementar políticas preventivas de forma mais estratégica, reduzindo os impactos da doença sobre a sociedade.

Este trabalho tem como objetivo desenvolver um modelo de previsão baseado em séries temporais para estimar o número de casos de notificação de sintomas gripais por mês. Utilizando abordagens estatísticas e de aprendizado de máquina, busca-se analisar tendências e padrões, fornecendo uma ferramenta útil para o monitoramento e a tomada de decisão no controle da gripe. Dessa forma, espera-se contribuir para um planejamento mais eficaz das ações de saúde pública, mitigando os efeitos de surtos e melhorando a capacidade de resposta dos sistemas de saúde.

2. Dados

2.1. Origem

Os dados utilizados no presente trabalho foram retirados da plataforma OpenDataSUS, que disponibiliza informações que podem servir para subsidiar análises objetivas da situação sanitária, tomadas de decisão baseadas em evidências e elaboração de programas de ações de saúde. A base de dados utilizada foi a Notificações de Síndrome Gripal - 2022, que possui notificações de casos leves e moderados suspeitos de covid-19. O conjunto de dados é composto de 28 bases de dados diferentes, sendo uma para cada estado, uma para o Distrito Federal e outra para dados com o estado não identificado. O foco deste trabalho será na base de dados de Minas com foco nas cidades com mais números de notificações mensais.

2.2. Pré-processamento

No pré-processamento encontramos problemas em relação ao tamanho da base de dados, que muitas vezes interrompia processos devido à quantidade de dados. Para solucionar este problema tivemos que reduzir nosso espaço amostral, ou seja, diminuir a quantidade de registros utilizados para a análise.

Para isso, utilizamos um recurso do Pandas que elimina linhas aleatoriamente.

Como não conseguimos garantir nenhum resultado a partir de um evento aleatório, geramos as análises tanto para a base inteira quanto para a base reduzida e verificamos que ambas apresentam comportamento semelhante.

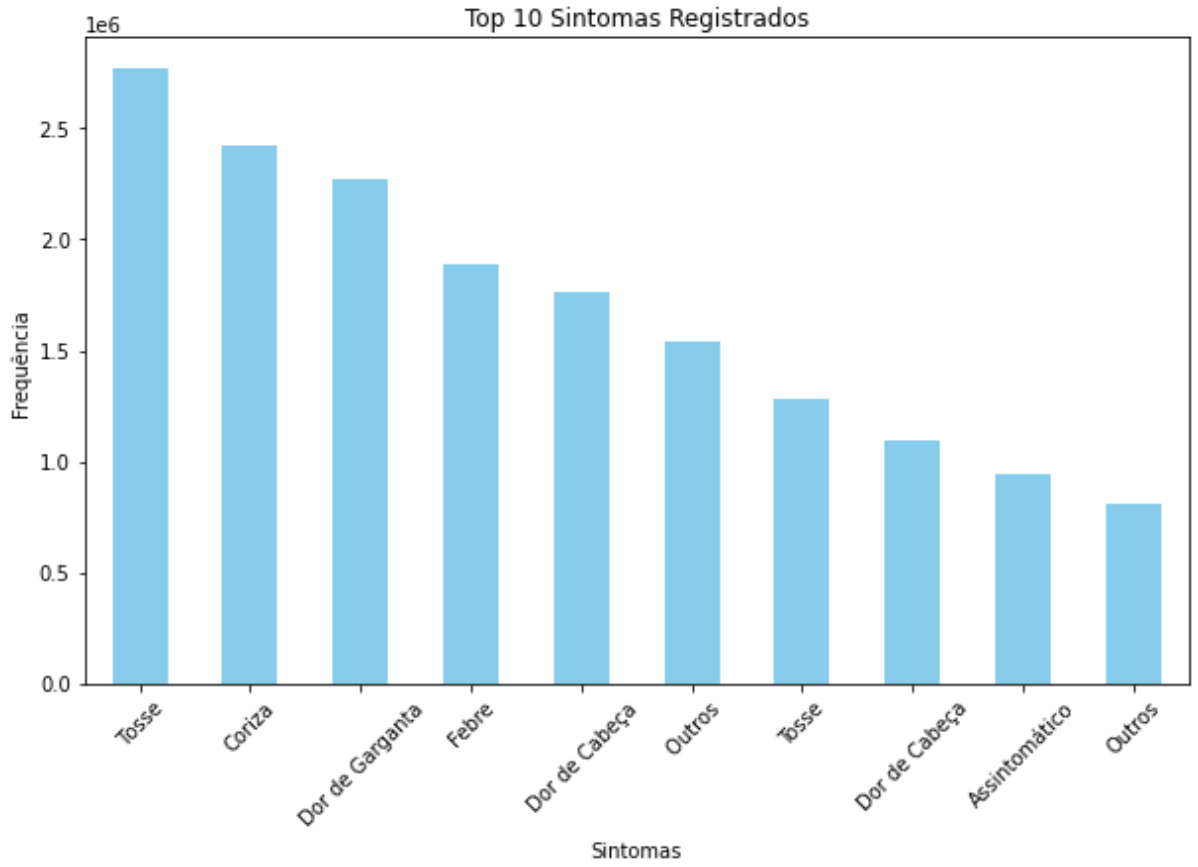

Gráfico 1 - Top 10 Sintomas Registrados na base completa

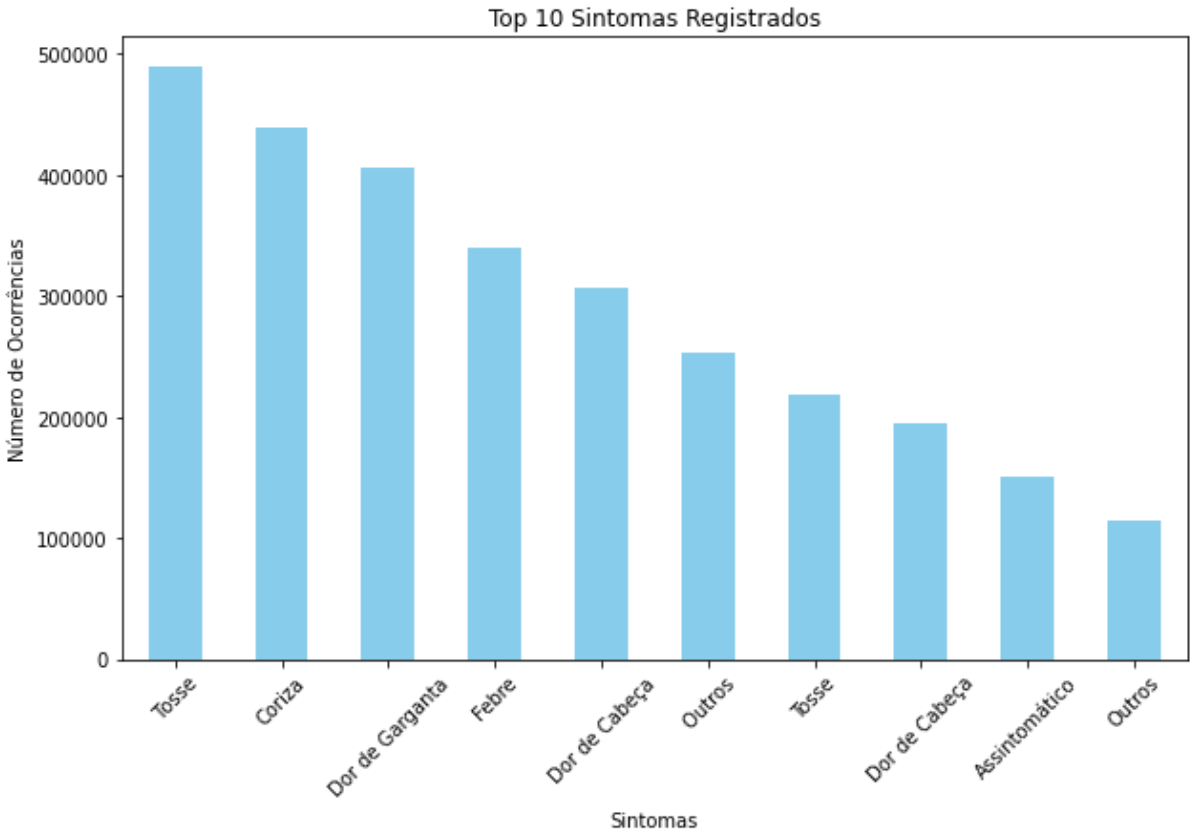

Gráfico 2 - Top 10 Sintomas Registrados na base completa

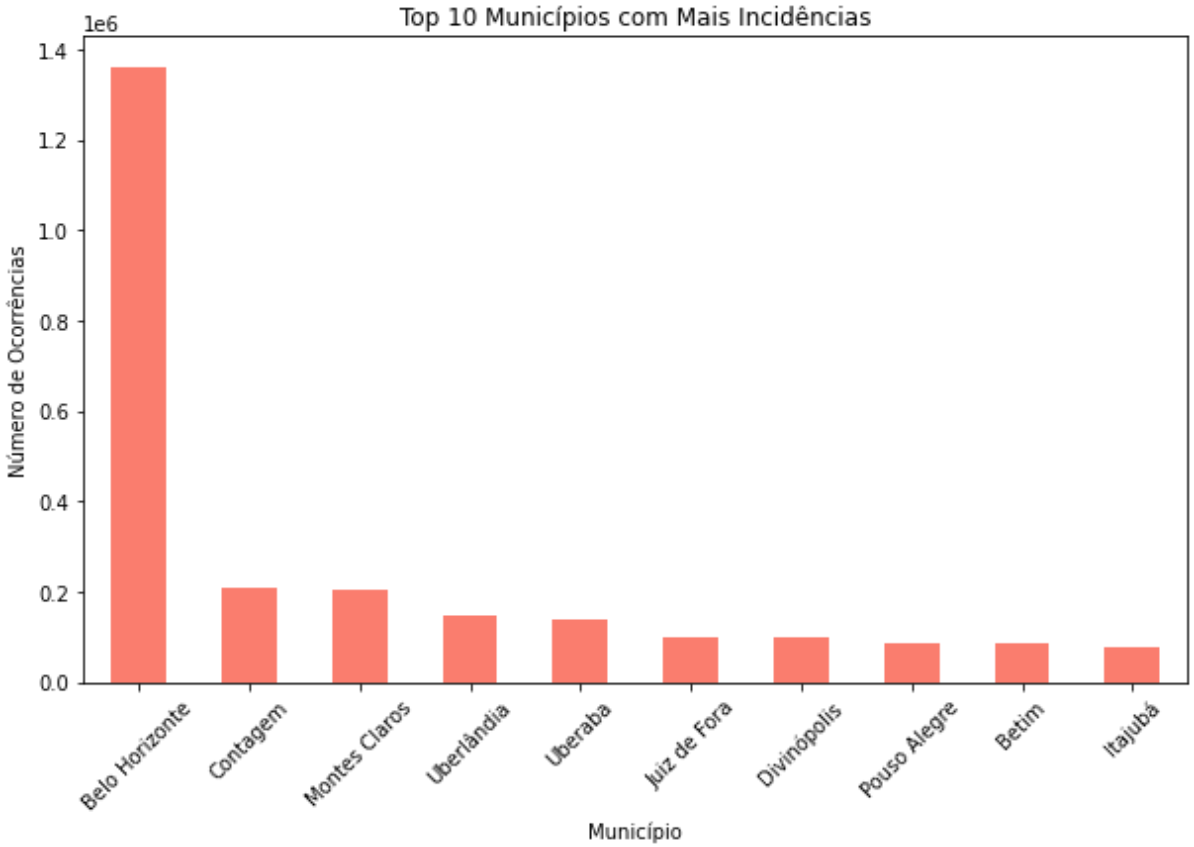

Gráfico 3 - Top 10 Municípios com mais registros na base completa

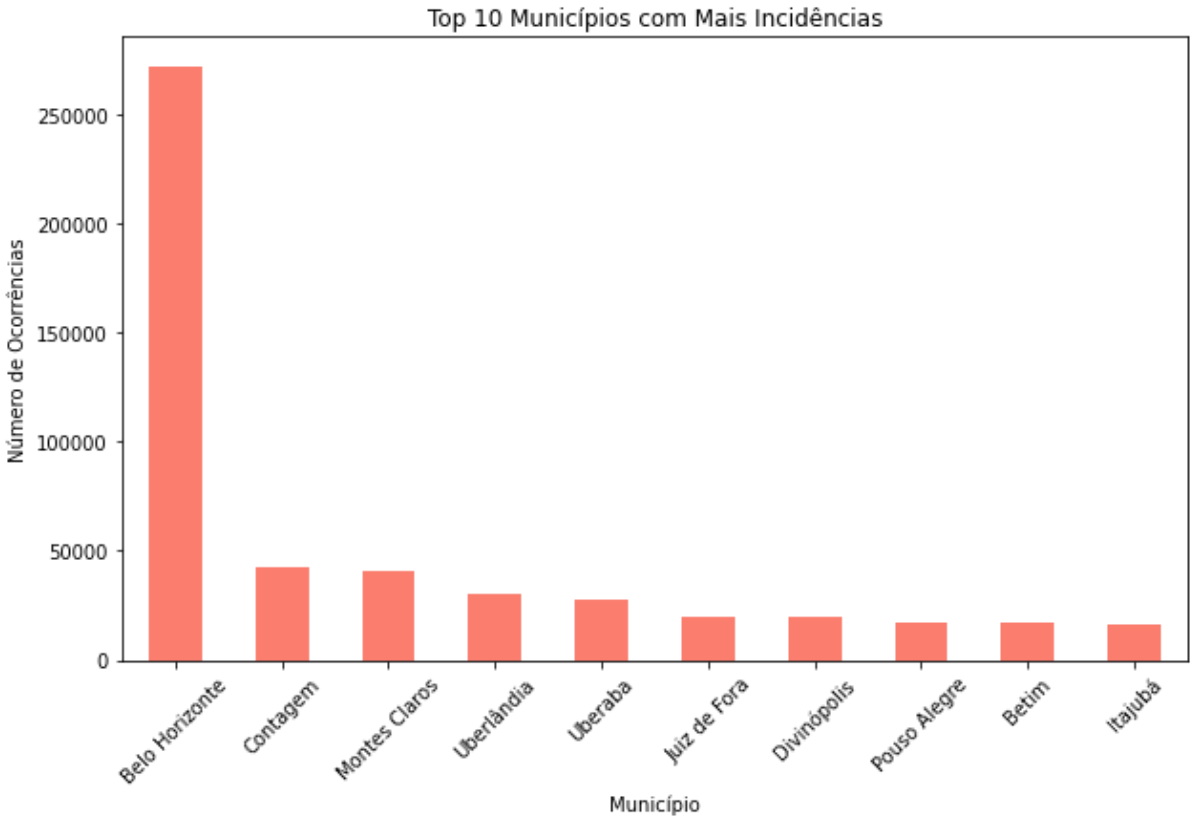

Gráfico 4 - Top 10 Municípios com mais registros na base reduzida

Como podemos observar, a base foi reduzida para cerca de 20% do tamanho inicial, respeitando a relação entre os registros que havia na base inteira.

Além disso, para os modelos de predição, fizemos o balanceamento das bases para que o número de afetados seja proporcional à população daquele município. Por exemplo, se nós entrarmos uma cidade com cerca de 50 mil notificações e mas a população daquela cidade for de 100 mil habitantes, 50% da cidade estaria doente. No entanto, se estivermos falando de uma cidade como Belo Horizonte, onde a população já chega a mais de dois milhões, 50 mil não estaria próximo de 50% da população. Então para isso, pegamos a relação de habitantes por município de Minas Gerais, e fazemos o cálculo de (Notificações / População).
Censo MG.

3. Previsão

A tarefa de previsão de notificações de SG se trata de uma tarefa de previsão de séries temporais, para esse fim existem alguns modelos que se destacam. Para o seguinte trabalho

foram utilizados os modelos de regressão polinomial, random forest e support vector regression.

3.1. Regressão Polinomial

Se baseia na regressão linear, porém utiliza de termos polinomiais para melhor representar curvas e tendências não lineares. É um modelo interessante para a tarefa de previsão de notificações visto que não possui um padrão linear, podendo ter picos ou vales em determinados meses.

3.2. Random Forest

O Random Forest Regressor é um modelo baseado em árvores de decisão que combina múltiplas árvores para criar uma previsão robusta. Ele funciona bem em séries temporais quando há relações complexas entre diferentes fatores que influenciam os valores futuros, o que pode gerar um resultado poderoso na tarefa de previsão de notificações já que muitas variáveis impactam na quantidade de casos.

3.3. Support vector regression

O Support Vector Regression (SVR) é uma versão do Support Vector Machine (SVM) adaptada para problemas de regressão. Ele tenta encontrar uma função que minimize o erro dentro de uma margem de tolerância, sendo útil para séries temporais que apresentam comportamento irregular.

Os resultados mostrados na imagem abaixo mostra uma grande divergência entre os modelos utilizados o que pode indicar um erro nos ajustes do modelos:

- O modelo de **Regressão Polinomial** pode estar com overfitting aos dados e gerando uma previsão irrealista.
- O **Random Forest** parece estar sendo muito conservador. Vale a pena testar novamente com hiperparâmetros diferentes, como aumentar a profundidade da árvore ou o número de estimadores.
- O **SVR** apresenta um resultado mais equilibrado, podendo ser um candidato melhor para previsão.