

# Model Card

## Detalhes do Modelo

- Esse modelo (versão 1.0) foi desenvolvido por Luís Eduardo Limas Brito, 27/01/2025.
- Ele foi criado a fim de estabelecer uma base para aplicação de metodologias de IA responsável. Por isso, ele não contém nenhuma implementação que siga as dimensões estabelecidas pela IAR.
- Modelo de Predição, com o objetivo de estimar o número de internações por doenças respiratórias (CID = J...) em um hospital para um dado mês.
- Há implementação de dois modelos, a fim de comparação da qualidade dos resultados. Regressão linear (implementado usando "scikit-learn") e LightGBM (modelo criado Microsoft), usando "early-stopping".

## Uso pretendido

- O modelo pode ser usado por hospitais para analisar estimativas de quantas internações podem-se esperar para o próximo mês, a fim de assistir no processo de escolha dos gastos e aquisição de recursos.

## Fatores

- Devido à quantidade de atributos, é provável que o usuário não terá acesso a todos os dados necessários. Por isso, a ausência de alguns (valores nulos) podem afetar o resultado obtido.

## Dados de Treinamento

- O modelo foi treinado nos dados do DataSUS. Especificamente, os do tipo SIH (Internações Hospitalares), com arquivos reduzidos (começam com RD), de todos os estados, dos anos de 2022 a 2024/06.
- Pré Processamento: os dados foram agrupados por hospital e mês/ano, contabilizando o total e calculando dados como média e razão de algumas colunas. Cada atributo usado para treinamento é uma defasagem temporal dos dados calculados no passo anterior (dados dos meses passados).
- Hospitais com uma média mensal abaixo de 5 internações foram reitradas.

## Dados de Avaliação

- Para o LightGBM, foi usado dados para validação (dados de 2024/7 até 2024/12 do DataSUS), seguindo o mesmo pré processamento dos dados de treinamento.
- Para realização de testes, foram usados os dados de 2025/01 até 2025/11 (com exceção dos estados do Acre e Roraima - "AC" e "RR" - que não estavam disponíveis ainda). Também passaram pelo mesmo processo citado acima.

## Métricas

- Erro Médio Absoluto (MAE): métrica para análise dos erros como valor total, sem valorizar algum tipo especial de erro. Valores altos indicam uma mistura entre alta quantidade de erros ou erros graves.
- Raiz do Erro Quadrático Médio (RMSE): métrica para análise de erros, valorizando erros graves. Útil para verificação da consistência do modelo, penalizando erros graves.
- Erro Médio Absoluto Simétrico Percentual (SMAPE): métrica para análise dos erros como valor em porcentagem, sem valorizar algum tipo especial de erro. Valores altos indicam uma alta taxa de erro.

## Avisos e Recomendações

- O modelo usa como atributo o CNES do hospital. Isso significa que ele terá uma melhor chance de acerto se o CNES providenciado for igual a um já visto antes. Por isso, esse modelo é recomendado para hospitais já cadastrados no SUS (de tamanho maior, pois hospitais com poucos casos de internações foram filtrados para fora do modelo).
- Decisões tomadas pelo modelo devem sempre ser acompanhadas por um profissional, uma vez que uma previsão nunca será sempre correta. Além disso previsões que erram "para baixo" podem ser perigosas, pois apontam que precisa-se de menos recursos que será necessário, podendo levar a complicações e mortes para pacientes.

## Análise Quantitativa

| Métrica | Base (Reg. Lin.) | LightGBM |
|---------|------------------|----------|
| MAE     | 13.902           | 12.358   |
| RMSE    | 598.748          | 434.361  |
| SMAPE   | 51.476           | 45.438   |

Table 1: Métricas por modelo