

Detalhes do Modelo

- Esse modelo (versão 2.0) foi desenvolvido por Luís Eduardo Limas Brito, 18/02/2026.
- Ele foi criado a fim de apresentar um exemplo prático de implementação da checklist proposta pelo NIAR.
- 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). Para LightGBM, devido ao uso de sementes aleatórias, decidiu-se executá-la 5 vezes e triou-se a média dos resultados avaliativos dele. As sementes usadas são, em ordem de execução: 778, 768, 758, 748, 738.

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.
- O modelo não deve ser usado para assistir casos de pacientes individualmente.

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.
- Durante o processamento dos dados, foram retirados hospitais com poucos casos (vide Data Card), a fim de evitar instabilidades no modelo. Assim, o modelo não é recomendado para uso com hospitais com baixa quantidade de internações.
- O estudo de justiça revelou que fatores sociais podem levemente afetar o resultado recebido, baseado na quantidade relativa de pessoas de diferentes sexos, idades e etnias. Pelo outro lado, diferentes regiões apresentaram alta disparidade na qualidade de resultados. Segue-se um gráfico ilustrando as diferenças na seção *Análise Quantitativa*

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 retiradas.
- Vide *data_card.md* para informações mais detalhadas sobre o processamento e propriedades dos dados.

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

- Essa versão apresenta métricas por grupo em vez de métrica geral. Como as métricas de MAE e RMSE trabalham com valores absolutos, diferentes números de amostras afetam o resultado obtido.
- Por esse motivo, decidiu-se usar apenas a métrica sMAPE (Erro Médio Absoluto Simétrico Percentual), uma vez que o valor é dado em percentual.
- A avaliação do desempenho foi separado pelos seguintes grupos: Sexo, Idade, Raça e Região. Para cada dado, categoriza-se baseado na quantidade maior de presença de algum grupo (por exemplo, um dado com maior quantidade de mulheres é categorizado como "Feminino", na categoria Sexo), exceto região, no qual cada dado naturalmente já contém essa informação por padrão.

Avisos e Recomendações

- Esse modelo foi desenvolvido somente com intuito educacional, não devendo ser aplicado em nenhum contexto real, devido à alta taxa de erro apresentada pelo modelo.
- Para informações extras sobre o processo de aplicação das metodologias propostas pelo NIAR, vide os documentos presentes dentro das pastas *docs* e *audit*.

Análise Quantitativa

