Gustavo Argentim Flausino

Modelagem Preditiva do Tempo de Internação com Machine Learning em Prontuários Hospitalares

Botucatu – SP
2025

Gustavo Argentim Flausino

Modelagem Preditiva do Tempo de Internação com Machine Learning em Prontuários Hospitalares

Trabalho apresentado ao Instituto de Biociências de Botucatu, da Universidade Estadual Paulista "Júlio de Mesquita Filho", como requisito parcial para obtenção do grau de Bacharel em Física Médica.

Universidade Estadual Paulista "Júlio de Mesquita Filho"
Instituto de Biociências de Botucatu

Orientador: Prof. Allan Felipe Fattori Alves

Botucatu – SP
2025

F587m

Flausino, Gustavo
Modelagem Preditiva do Tempo de Internação com Machine Learning em Prontuários Hospitalares / Gustavo Flausino. -- Botucatu, 2025
31 p.

Trabalho de conclusão de curso (Bacharelado - Física Médica) - Universidade Estadual Paulista (UNESP), Instituto de Biociências, Botucatu
Orientador: Allan Felipe Fattori Alves

1. Aprendizado de Máquina. 2. Tempo de Internação. 3. Prontuários Eletrônicos. 4. Random Forest. 5. Gestão Hospitalar. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

GUSTAVO ARGENTIM FLAUSINO

Modelagem Preditiva do Tempo de Internação com Machine Learning em Prontuários Hospitalares

Trabalho de Conclusão de Curso, apresentado a Universidade Estadual Paulista, como parte das exigências para a obtenção do título de Bacharel, do curso de Graduação em Física Médica.

Local: Botucatu, 28 Novembro de 2025

BANCA EXAMINADORA

_____

Prof. Allan Felipe Fattori Alves
Departamento de Biofísica e Farmacologia

_____

Prof. Joel Mesa Hormaza
Departamento de Biofísica e Farmacologia

_____

Prof. Roberto Morato Fernandez
Departamento de Biofísica e Farmacologia

Agradecimentos

Primeiramente, agradeço a todas as pessoas que conheci ao longo do período em que estive na universidade. Cada experiência, sendo boa ou ruim, contribuiu para o meu amadurecimento pessoal me ajudando a evoluir e compreender melhor o mundo.

Agradeço profundamente à minha família, em especial à minha mãe, Telma Argentim, e ao meu padrasto, Emerson Cleiton de Oliveira, que me criaram com amor e dedicação. Me ensinaram valores que levarei para toda a vida e que tornaram possível a minha formação desde a infância até a conclusão dessa fase tão importante. Sem o apoio deles, não teria conseguido nada disso. Mostro também minha gratidão aos meus irmãos, que me proporcionaram inúmeros momentos de alegria, risadas e, é claro, muitos de estresse, todos igualmente importantes para o meu crescimento. Agradeço ainda ao meu pai, Leandro Flausino, e à minha madrasta, Jennifer Rodrigues, que me acolheram com carinho na fase final desta jornada, oferecendo apoio quando precisei.

Sou grato aos meus queridos amigos de infância e adolescência, Jeiel Santos, Ricardo Codello e Matheus Franco, por todos os momentos inesquecíveis, seja nas longas jogatinas de madrugada ou nas conversas repletas de risadas e reflexões. Agradeço também aos três amigos que se tornaram indispensáveis durante a faculdade, Eduardo Perondini, Victor Gustavo Caferro e Gabriel Faustino por compartilharem comigo as alegrias, as dificuldades e os desafios dessa jornada universitária. Entre brincadeiras, discussões e horas de estudo, tornaram a minha caminhada mais leve e divertida.

Minha sincera gratidão ao Dr. Luís Gustavo Modelli de Andrade, idealizador do projeto e curador dos dados, pois sem ele não teria sido possível estar aqui participando de um trabalho tão importante.

E, por fim, e nada menos importante que qualquer um, agradeço à mulher mais incrível que já conheci, minha noiva, Isabelle Reis Gomes. Nos conhecemos ainda no colégio, há cinco anos, e desde então ela tem sido minha maior inspiração. Isabelle me ajudou a acreditar em mim mesmo nos momentos mais difíceis, esteve ao meu lado em todas as fases, me impulsionou a buscar o meu melhor e me ensinou o verdadeiro significado de parceria e amor. Sou eternamente grato por tê-la comigo nesta jornada e em todas as que ainda virão.

Resumo

Este trabalho apresenta o desenvolvimento de um modelo preditivo para estimar o tempo de internação hospitalar, em inglês Length of Stay (LOS), com base em dados clínicos e administrativos obtidos de prontuários eletrônicos. A pesquisa busca contribuir para a gestão hospitalar por meio da aplicação de técnicas de aprendizado de máquina, fornecendo uma ferramenta de apoio na decisão que auxilie no planejamento de altas, como também na alocação de leitos e na otimização de recursos.

O estudo foi conduzido utilizando a linguagem Python e o ambiente Google Colab, com o emprego de bibliotecas especializadas em análise de dados e modelagem estatística. Após o tratamento e padronização de informações, foi aplicado o algoritmo Random Forest Regressor, que é capaz de lidar com variáveis heterogêneas e interpretar a importância relativa dos fatores preditores. O desempenho do modelo foi avaliado com métricas como coeficiente de determinação (R^2), erro absoluto médio (MAE) e raiz do erro quadrático médio (RMSE), apresentando resultados satisfatórios e consistentes com o comportamento observado durante as internações.

Os resultados demonstram que o modelo desenvolvido é capaz de prever o tempo de internação com boa precisão e também estabilidade, identificando as variáveis-chave associadas à complexidade clínica e à intensidade assistencial. Conclui-se que a aplicação de técnicas de aprendizado de máquina em dados hospitalares representa uma ferramenta muito promissora para apoiar as tomadas de decisão, contribuindo para a eficiência operacional e para a melhoria contínua dos processos de gestão na saúde.

Palavras-chave: Aprendizado de Máquina; Tempo de Internação; Prontuários Eletrônicos; Random Forest; Gestão Hospitalar; Ciência de Dados.

Abstract

This work presents the development of a predictive model to estimate the length of stay (LOS), based on clinical and administrative data extracted from electronic health records. The research aims to contribute to hospital management by applying machine learning techniques, providing a decision-support tool that assists in discharge planning, bed allocation, and the optimization of healthcare resources.

The study was conducted using the Python programming language and the Google Colab environment, employing specialized libraries for data analysis and statistical modeling. After data cleaning and standardization, the *Random Forest Regressor* algorithm was applied due to its ability to handle heterogeneous variables and interpret the relative importance of predictive factors. The model's performance was evaluated using metrics such as the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE), showing satisfactory and consistent results with the real patterns observed in hospitalizations.

The results demonstrate that the developed model can predict hospitalization time with good accuracy and stability, identifying key variables associated with clinical complexity and care intensity. It is concluded that the use of machine learning techniques in hospital data represents a promising tool to support decision-making, contributing to operational efficiency and continuous improvement of healthcare management processes.

Keywords: Machine Learning; Length of Stay; Electronic Health Records; Random Forest; Hospital Management; Data Science.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

1.1 Motivação

Nos últimos anos, os sistemas de saúde têm enfrentado o desafio crescente de equilibrar qualidade assistencial, o custo e a eficiência operacional. Entre os indicadores utilizados para mensurar o desempenho hospitalar, o tempo de internação conhecido como Length of Stay (LOS) se destaca por mostrar, em um único número, o equilíbrio entre complexidade clínica e eficiência administrativa. A redução do tempo médio de permanência, sem comprometer a qualidade do cuidado, é uma meta que é almejada por gestores e profissionais de saúde, uma vez que reflete diretamente na rotatividade de leitos, nos custos hospitalares e também na capacidade de atendimento.

Com a digitalização dos prontuários e a disponibilidade crescente de bases de dados clínico-administrativas, tornou-se possível aplicar técnicas de aprendizado de máquina (Machine Learning) para prever o tempo de internação de forma automatizada e de forma mais precisa. Essas técnicas têm a vantagem de identificar relações não lineares e interações complexas entre variáveis. O uso de algoritmos como o Random Forest Regressor permite modelar fatores multivariados e oferecer medidas interpretáveis de importância de variáveis, tornando-se uma opção robusta para cenários clínicos, onde a interpretabilidade e a confiança do profissional são essenciais.

Além da relevância técnica, o tema possui forte motivação social e prática. Em um cenário de envelhecimento populacional prever o tempo de internação (LOS) pode auxiliar gestores hospitalares na organização de fluxos, prevenção de sobrecarga e otimização da utilização dos seus recursos. Do ponto de vista clínico, a previsão de internações prolongadas pode ajudar a identificar precocemente pacientes de maior risco e planejar intervenções multidisciplinares, contribuindo para um cuidado mais eficiente. Assim, o presente trabalho une interesse técnico e propósito social, aplicando conceitos de aprendizado de máquina no contexto hospitalar, com o objetivo de apoiar a gestão e a melhoria da experiência do paciente.

1.2 Objetivos

O presente trabalho tem como objetivo principal desenvolver e avaliar um modelo preditivo capaz de estimar o tempo de internação hospitalar a partir de variáveis clínicas e administrativas extraídas de prontuários eletrônicos. A proposta visa aplicar técnicas de aprendizado de máquina para auxiliar a gestão hospitalar na previsão de altas, otimização de leitos e planejamento de recursos, contribuindo para uma utilização mais eficiente do sistema de saúde.

O modelo foi desenvolvido com base no algoritmo Random Forest Regressor, selecionado

por sua capacidade de lidar com dados heterogêneos, alta acurácia e facilidade de interpretação dos resultados. A abordagem proposta busca não apenas testar a performance do modelo, mas também analisar a relevância das variáveis envolvidas, fornecendo subsídios para a compreensão dos fatores que influenciam o tempo de internação.

2 Metodologia

2.1 Coleta e Organização dos Dados

O conjunto de dados utilizado neste estudo foi disponibilizado pelo Hospital das Clínicas da Faculdade de Medicina de Botucatu, após aprovação junto ao Comitê de Ética em Pesquisa (CAAE 69761723.1.0000.5411). Os dados foram disponibilizados em formato de planilha (.xlsx), contendo informações anonimizadas referentes a pacientes internados ao longo de um período específico (2015 a 2019). Todos os identificadores pessoais (nome, CPF, número de prontuário, etc.) não constam, garantindo a confidencialidade dos sujeitos, conforme preconizado pela Lei Geral de Proteção de Dados (LGPD), (Lei nº 13.709/2018).

As variáveis do conjunto contemplam dados clínicos e demográficos relevantes, como idade, sexo, diagnóstico principal, procedimentos realizados, exames laboratoriais e tempo total de permanência hospitalar (variável-alvo), que foi comumente chamada, durante todo o processo, de LOS. O carregamento do arquivo foi realizado diretamente no ambiente Python por meio da biblioteca pandas, utilizando a função read_xlsx().

Após o carregamento, foram inspecionados as dimensões do conjunto de dados (número de linhas e colunas) e os tipos de variáveis, possibilitando a identificação de inconsistências, valores ausentes e também duplicidades. Essa análise inicial foi acompanhada da verificação estatística descritiva das variáveis contínuas (média, desvio padrão, valores mínimos e máximos), auxiliando na compreensão da distribuição e dispersão.

A estruturação adequada do conjunto de dados foi essencial para o sucesso das etapas seguintes, uma vez que a qualidade da base influencia diretamente a capacidade preditiva do modelo. Assim, essa parte teve como objetivo preparar o conjunto de dados para as etapas de tratamento e modelagem, garantindo integridade e consistência das informações.

2.2 Etapas preliminares

O desenvolvimento do trabalho foi realizado em linguagem Python, dentro do ambiente Google Colab, utilizando bibliotecas amplamente empregadas em ciência de dados e aprendizado de máquina. As principais funções e rotinas do código foram organizadas de forma modular, permitindo o tratamento dos dados, a análise exploratória e a construção do modelo de predição do tempo de internação. O uso de código segmentado facilitou a reprodutibilidade e a compreensão de cada etapa da pesquisa.

Durante o tratamento e preparação dos dados, as funções read_csv() e merge() da biblioteca pandas foram essenciais para o carregamento e integração das planilhas hospitalares, vinculando as informações por meio do código de atendimento. Em seguida, foram aplicadas

as funções isnull() e sum() para identificar dados ausentes, seguidas de rotinas de imputação por média e mediana, conforme a natureza das variáveis. Para eliminar registros duplicados, foi utilizada a função drop_duplicates(), garantindo que cada internação fosse representada apenas uma vez na base. A detecção de valores extremos (outliers) foi conduzida por meio de medidas estatísticas e visualizações de boxplots, apoiadas nas bibliotecas matplotlib e seaborn.

Na etapa de modelagem, a função train_test_split() do módulo scikit-learn foi utilizada para dividir o conjunto de dados em treino (80%) e teste (20%), permitindo avaliar a capacidade de generalização do modelo. O algoritmo Random Forest Regressor foi então implementado e ajustado por meio do método fit(), sendo posteriormente avaliado pelas métricas R^2, MAE e RMSE. Por fim, o atributo feature_importances_ foi empregado para identificar as variáveis mais relevantes na predição do LOS, fornecendo uma visão interpretável dos fatores clínicos e administrativos que mais influenciaram o resultado.

2.3 Tratamento e Limpeza dos Dados

O processo de tratamento e limpeza foi conduzido de forma sistemática no ambiente Python, utilizando principalmente a biblioteca pandas e o ambiente de programação do Google Colab. Essa etapa teve como objetivo garantir que os dados provenientes das planilhas hospitalares estivessem íntegros, consistentes e adequados para o treinamento do modelo de aprendizado de máquina (HAN; KAMBER; PEI, 2023; McKINNEY, 2010).

Inicialmente, os diferentes conjuntos de dados foram carregados e vinculados por meio do código de atendimento, unificando informações provenientes de bases distintas (pacientes, exames e medicamentos). Essa junção foi realizada pela função merge(), preservando apenas os registros com correspondência válida entre as tabelas (GARCIA et al., 2015; JAMES et al., 2021).

Em seguida, foi feita a verificação de valores ausentes, por meio dos métodos isnull() e sum(). As colunas com elevado número de dados ausentes foram descartadas, enquanto as variáveis numéricas com poucos nulos receberam imputação pela média ou mediana. Variáveis categóricas, quando incompletas, tiveram valores preenchidos com a categoria mais frequente. Essa estratégia permitiu reduzir a perda de informação sem introduzir viés relevante (PEDREGOSA et al., 2011; HAN; KAMBER; PEI, 2023).

Após a imputação, realizou-se a padronização dos tipos de dados, garantindo que colunas numéricas, de texto e datas estivessem corretamente formatadas. Nessa etapa também foram renomeadas colunas para um formato padronizado (snake_case) e removidas variáveis irrelevantes ao problema, como identificadores internos sem significado preditivo (GERON, 2019; GARCIA et al., 2015).

A remoção de duplicidades foi executada por meio da função drop_duplicates(), assegurando que cada atendimento hospitalar fosse representado uma única vez no conjunto final. Essa verificação evitou distorções estatísticas e duplicação de amostras durante o treino do modelo

(SHMUELI; KOPPIUS, 2011).

Posteriormente, realizou-se a avaliação de outliers nas variáveis numéricas. Foram calculadas medidas descritivas e analisados valores extremos com base em limites interquartílicos. Casos comprovadamente inconsistentes, como durações negativas ou intervalos incoerentes de internação, foram removidos. Por outro lado, valores elevados, mas plausíveis do ponto de vista clínico, foram mantidos para preservar a representatividade dos dados reais (TUKEY, 1977; HAN; KAMBER; PEI, 2023).

Ao término desse momento, o conjunto de dados encontrava-se limpo, padronizado e coerente, pronto para as etapas de pré-processamento, análise exploratória e modelagem. O resultado desse processo foi uma base de dados robusta, com variáveis de qualidade, possibilitando a aplicação segura do código (HASTIE et al., 2009; SONG et al., 2020).

2.4 Análise Exploratória e Visualização

Com o conjunto de dados devidamente tratado, foi realizada a análise exploratória para entender o comportamento das variáveis e suas relações com o tempo de internação hospitalar. Essa etapa teve como finalidade identificar padrões, tendências e possíveis inconsistências remanescentes, além de auxiliar na seleção de variáveis relevantes para o modelo preditivo (TUKEY, 1977; GERON, 2019).

Inicialmente, foram calculadas medidas descritivas (média, mediana, desvio padrão, valores mínimos e máximos) por meio do método describe() da biblioteca pandas. O uso da biblioteca permite sumarização e manipulação eficiente de grandes volumes de dados (McKINNEY, 2010), possibilitando uma visão geral das distribuições e da amplitude dos valores em cada variável. Essa análise evidenciou diferenças de escala e variabilidade entre os atributos clínicos e administrativos (KOTSIANTIS et al., 2006; McKINNEY, 2010).

Para avaliar a relação entre as variáveis e a variável-alvo (LOS), foram produzidos gráficos de dispersão e histogramas, destacando o comportamento das variáveis numéricas como idade, quantidade de exames e uso de antibióticos. Essas visualizações permitem identificar distribuições assimétricas e padrões de concentração, sendo amplamente empregadas em estudos preditivos com dados médicos (MULLER; GUIDO, 2016; GARCIA et al., 2015).

Em seguida, foi construída a matriz de correlação utilizando o coeficiente de Pearson, representada visualmente por meio de um heatmap gerado com a biblioteca seaborn. O coeficiente de Pearson é amplamente utilizado para mensurar relações lineares entre variáveis contínuas, enquanto a visualização matricial auxilia na identificação de correlações mais significativas entre os atributos (JAMES et al., 2021; FIELD, 2013).

Também foram elaborados boxplots para verificar a presença de valores atípicos e a dispersão dos dados em cada variável. Essa abordagem é essencial para a visualização de outliers e confirmação da consistência das faixas de valores após o tratamento de anomalias. Além disso, os boxplots possibilitaram compreender o comportamento das variáveis categóricas, como

diagnóstico ou tipo de atendimento, em relação à variável de saída (TUKEY, 1977; KOTSIANTIS et al., 2006).

Por fim, as visualizações possibilitaram identificar quais atributos apresentavam correlação mais significativa com o tempo de internação, orientando a etapa posterior de modelagem preditiva. A análise exploratória validou a integridade do conjunto de dados e forneceu subsídios para a interpretação clínica e a seleção das variáveis incluídas no modelo de aprendizado de máquina (HASTIE et al., 2009; PEDREGOSA et al., 2011).

2.5 Modelagem Preditiva

A etapa de modelagem preditiva teve como objetivo estimar o tempo de internação hospitalar a partir das variáveis clínicas e administrativas disponíveis. O modelo selecionado para essa tarefa foi o Random Forest Regressor, pertencente à biblioteca scikit-learn, escolhido por sua capacidade de lidar com dados heterogêneos, robustez frente a ruídos e interpretabilidade das variáveis. Essa técnica tem sido amplamente utilizada em contextos de saúde para modelagem de dados complexos e não lineares, demonstrando eficiência e estabilidade em cenários com múltiplas variáveis correlacionadas (BREIMAN, 2001; MULLER; GUIDO, 2016).

Inicialmente, o conjunto de dados foi dividido em dois subconjuntos independentes: treinamento e teste, por meio da função train_test_split() da biblioteca scikit-learn. Essa divisão teve como propósito avaliar o desempenho do modelo em dados não utilizados durante o aprendizado, garantindo uma medida de generalização mais fidedigna. A proporção utilizada manteve aproximadamente 80% dos dados para treino e 20% para teste, assegurando quantidade suficiente de amostras em ambas as partes (GERON, 2019; SONG et al., 2020).

Antes do treinamento, as variáveis categóricas foram codificadas em formato numérico por meio de técnicas de Label Encoding ou One-Hot Encoding, dependendo do número de categorias. Esse processo foi executado utilizando a função get_dummies() da biblioteca pandas, que permite a criação eficiente de colunas binárias para representação categórica. Em seguida, as variáveis numéricas foram normalizadas quando necessário, a fim de reduzir discrepâncias de escala e melhorar a estabilidade do algoritmo (GARCIA et al., 2015; McKINNEY, 2010).

O modelo de Random Forest foi configurado com um número inicial de estimadores (n_estimators) ajustado empiricamente, equilibrando desempenho e tempo de processamento. Foram mantidos parâmetros padrão para profundidade máxima e número mínimo de amostras por folha, uma vez que o foco inicial foi verificar a capacidade geral do modelo de capturar relações não lineares entre as variáveis preditoras e o tempo de internação. Após o treinamento, o modelo foi submetido à etapa de validação, na qual se calculou o coeficiente de determinação (R^2) como métrica principal de desempenho, complementado pelo erro absoluto médio (MAE) e o erro quadrático médio (MSE). Essas métricas permitem mensurar a precisão e a magnitude dos erros de previsão, oferecendo uma avaliação ampla do desempenho do modelo (JAMES et al., 2021; SHMUELI; KOPPIUS, 2011).

Além disso, foi gerado o gráfico de importância das variáveis, por meio do atributo feature_importances_, destacando os fatores que mais contribuíram para as previsões. Essa análise foi essencial para compreender a influência de variáveis como idade, quantidade de exames e uso de antibióticos no tempo estimado de internação. A importância das variáveis em modelos de floresta aleatória permite interpretar a relevância de cada atributo de forma robusta e estatisticamente coerente (CUTLER et al., 2007; HAN; KAMBER; PEI, 2023).

O modelo de Random Forest apresentou desempenho satisfatório, demonstrando capacidade de generalização e estabilidade frente à variabilidade dos dados clínicos. A partir dessa configuração inicial, tornou-se possível explorar ajustes finos de hiperparâmetros e comparações com outros algoritmos de regressão, consolidando a base para a avaliação dos resultados apresentados nas seções seguintes (GERON, 2019; HAN; GARCIA et al., 2015).

2.6 Avaliação do Modelo

A avaliação do modelo preditivo foi conduzida com o objetivo de mensurar o desempenho obtido pelo algoritmo e verificar sua capacidade de generalização. A escolha das métricas de avaliação buscou contemplar diferentes aspectos da precisão preditiva, bem como a magnitude dos erros estimados (HASTIE et al., 2009; KOTSIANTIS et al., 2006).

Inicialmente, utilizou-se o coeficiente de determinação (R^2) como métrica principal de ajuste do modelo, uma vez que ele expressa a proporção da variância da variável dependente que é explicada pelas variáveis independentes. Valores mais próximos de 1 indicam melhor ajuste, refletindo alta capacidade de explicação do modelo sobre o fenômeno estudado (TAN et al., 2019). Em complementação, foram aplicadas métricas de erro absoluto médio (MAE) e erro quadrático médio (MSE), amplamente utilizadas em problemas de regressão para avaliar a magnitude das diferenças entre valores previstos e observados (GERON, 2019; SONG et al., 2020).

A interpretação conjunta dessas métricas fornece uma análise mais completa sobre o comportamento do modelo, permitindo identificar tanto o grau de precisão quanto a dispersão dos erros (GARCIA et al., 2015). Enquanto o MAE mede a diferença média em unidades reais do alvo predito, o MSE penaliza erros maiores de forma quadrática, tornando-se sensível a outliers (JAMES et al., 2021). Essa combinação é recomendada para cenários clínicos em que previsões muito distantes da realidade podem representar riscos operacionais (KUHN; JOHNSON, 2013; SHMUELI; KOPPIUS, 2011).

Além das métricas quantitativas, a análise gráfica dos resíduos foi utilizada como ferramenta de diagnóstico para verificar o comportamento dos erros em relação às previsões. A ausência de tendências visuais marcantes nos resíduos indica bom ajuste e ausência de viés sistemático no modelo (FIELD, 2013). Essa abordagem visual complementa a avaliação estatística, reforçando a robustez e a confiabilidade das previsões (HAN; KAMBER; PEI, 2023; HASTIE et al., 2009).

De forma geral, a avaliação confirmou que o modelo apresentou desempenho consistente, mantendo valores de erro reduzidos e coeficiente de determinação satisfatório. A combinação de métricas numéricas e análise visual dos resíduos proporcionou uma visão abrangente da qualidade do modelo e de sua adequação ao conjunto de dados hospitalares (TAN et al., 2019; GERON, 2019; SONG et al., 2020).

3 Resultados e Discussão

3.1 Análise Descritiva dos Dados

Esta seção apresenta os principais resultados obtidos na caracterização do conjunto de dados hospitalares após o tratamento descrito na Metodologia. A análise descritiva permitiu compreender o perfil da amostra e identificar padrões associados à duração da internação (LOS). Todas as análises foram conduzidas em ambiente Python, utilizando as bibliotecas pandas, seaborn e matplotlib (McKINNEY, 2010; PEDREGOSA et al., 2011).

3.1.1 Distribuição do tempo de internação (LOS)

A Figura 1 apresenta a distribuição do tempo de internação. Observa-se forte assimetria à direita, com concentração de casos em períodos inferiores a 10 dias e uma cauda longa representando internações prolongadas. Tal comportamento é comum em estudos hospitalares, nos quais a maioria dos pacientes recebe alta em poucos dias, enquanto um pequeno número de casos mais complexos eleva a média geral de permanência (SILVA et al., 2021).

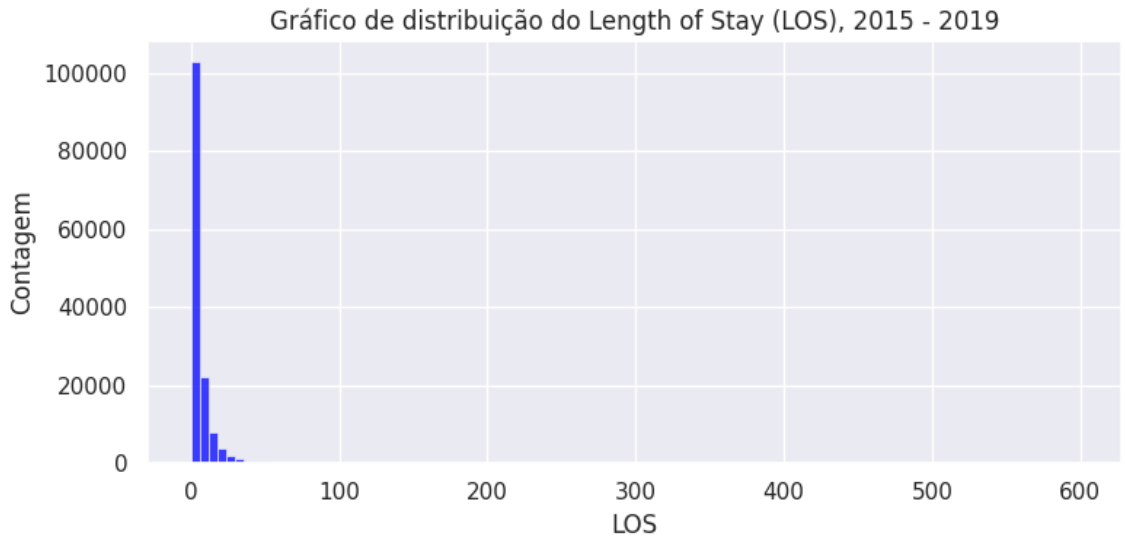

Figura 1 – Distribuição do tempo de internação (LOS).
Fonte: elaboração própria (dados Hospital das Clínicas Botucatu).

3.1.2 Idade dos pacientes

A amostra abrange pacientes de todas as faixas etárias, com predominância de adultos e idosos, especialmente acima dos 60 anos, como mostra a Figura 2. Essa distribuição é compatível com o perfil demográfico do hospital analisado, caracterizado por alta demanda geriátrica. A relação entre idade e LOS revelou tendência positiva: pacientes mais idosos

apresentaram maior tempo médio de internação, o que está de acordo com estudos que associam envelhecimento populacional ao aumento do tempo de permanência hospitalar devido à maior carga de comorbidades e recuperação mais lenta (BREIMAN, 2001; JAMES et al., 2021).

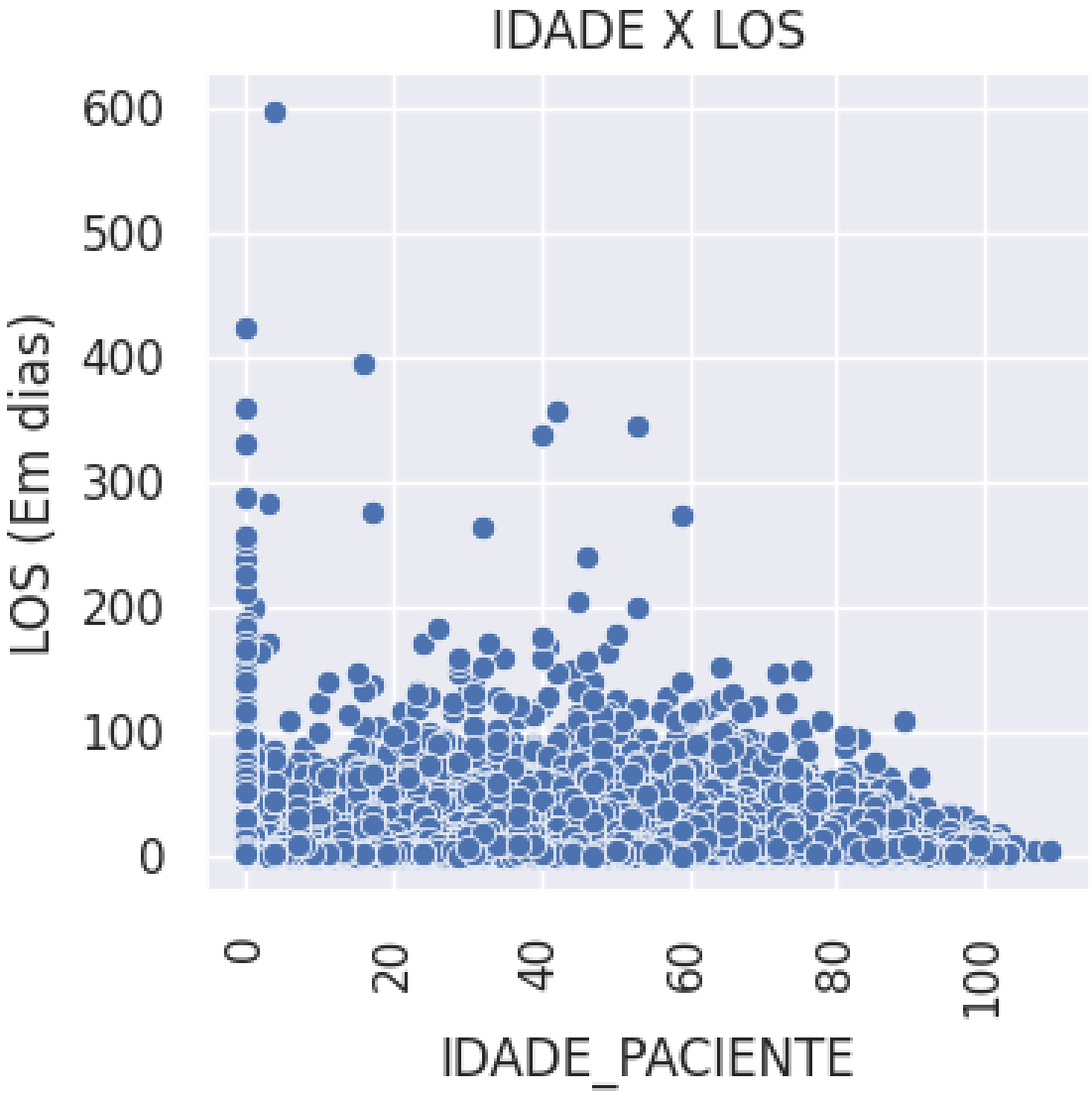

Figura 2 – Distribuição etária dos pacientes internados.
Fonte: elaboração própria (dados Hospital das Clínicas Botucatu).

3.1.3 Sexo, tipo e classificação da internação

A análise por sexo (Figura 3) indica discreta predominância masculina na amostra e maior tempo médio de permanência entre homens. Esse resultado é coerente com a literatura, que aponta diferenças no perfil epidemiológico e na gravidade dos casos entre gêneros (OLIVEIRA et al., 2020; KUHN; JOHNSON, 2013).

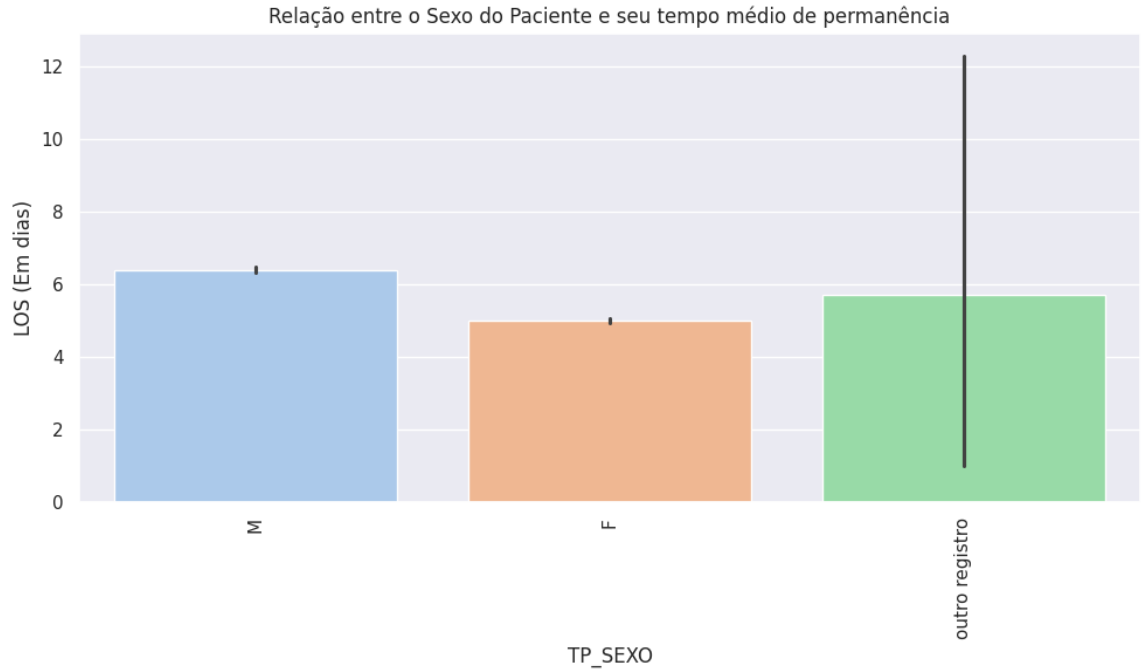

Figura 3 – Tempo de internação por sexo do paciente.
Fonte: elaboração própria (dados Hospital das Clínicas Botucatu).

Em relação ao tipo de internação (Figura 4), as internações clínicas apresentaram tempo de permanência superior às cirúrgicas, possivelmente devido à natureza mais prolongada de tratamentos clínicos e reabilitações. Já na classificação (Figura 5), casos urgentes apresentaram LOS médio maior que os eletivos, refletindo maior complexidade e imprevisibilidade dos atendimentos emergenciais (FIELD, 2013; OLIVEIRA et al., 2020).

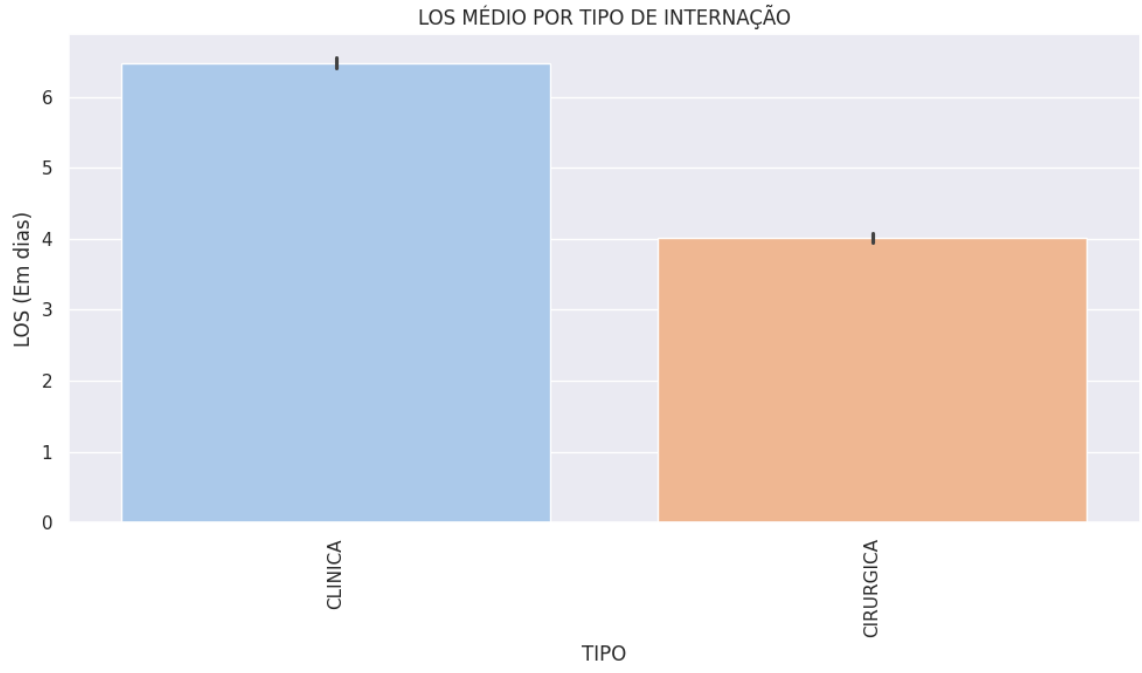

Figura 4 – Tempo médio de internação por tipo de internação.
Fonte: elaboração própria (dados Hospital das Clínicas Botucatu).

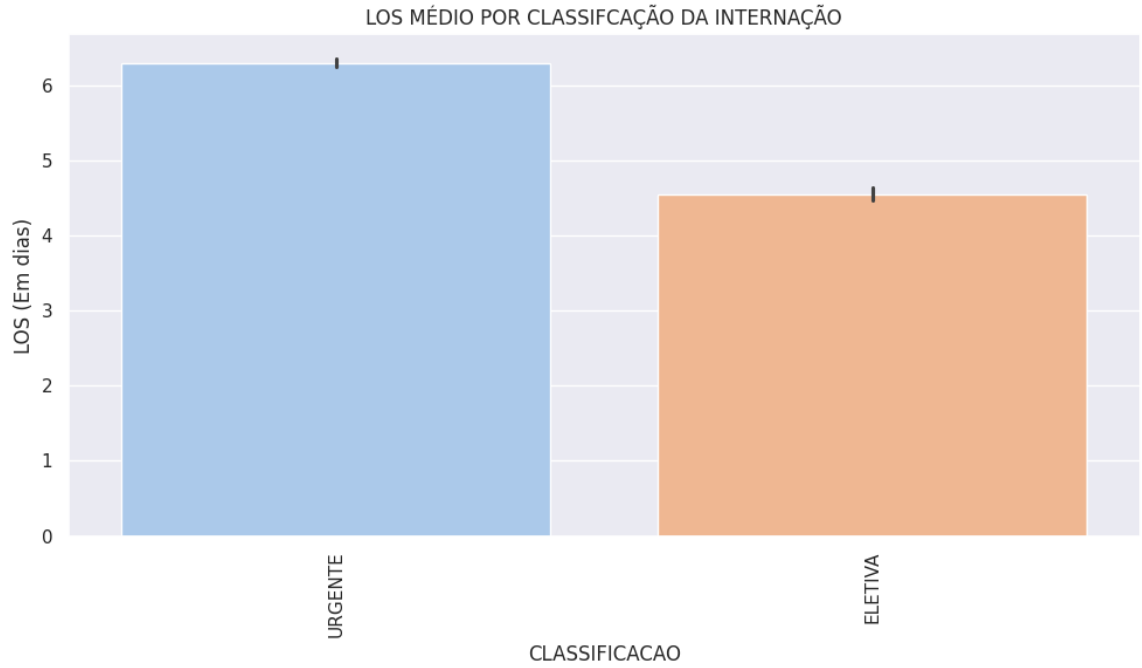

Figura 5 – Tempo médio de internação por classificação (urgente × eletiva).
Fonte: elaboração própria (dados Hospital das Clínicas Botucatu).

3.1.4 Diagnóstico (Capítulos do CID)

A agregação por capítulos do Código Internacional de Doenças (CID) revelou padrões distintos de duração de internação (Figura 6). Pacientes classificados no capítulo "Transtornos mentais e comportamentais" apresentaram o maior tempo médio de permanência, seguidos por doenças do aparelho circulatório e do sistema respiratório (COSTA et al., 2022; TAN et al., 2019). Essa heterogeneidade reforça a influência do tipo de diagnóstico sobre a demanda assistencial e o tempo necessário para estabilização clínica, conforme apontado em estudos de perfil hospitalar (COSTA et al., 2022).

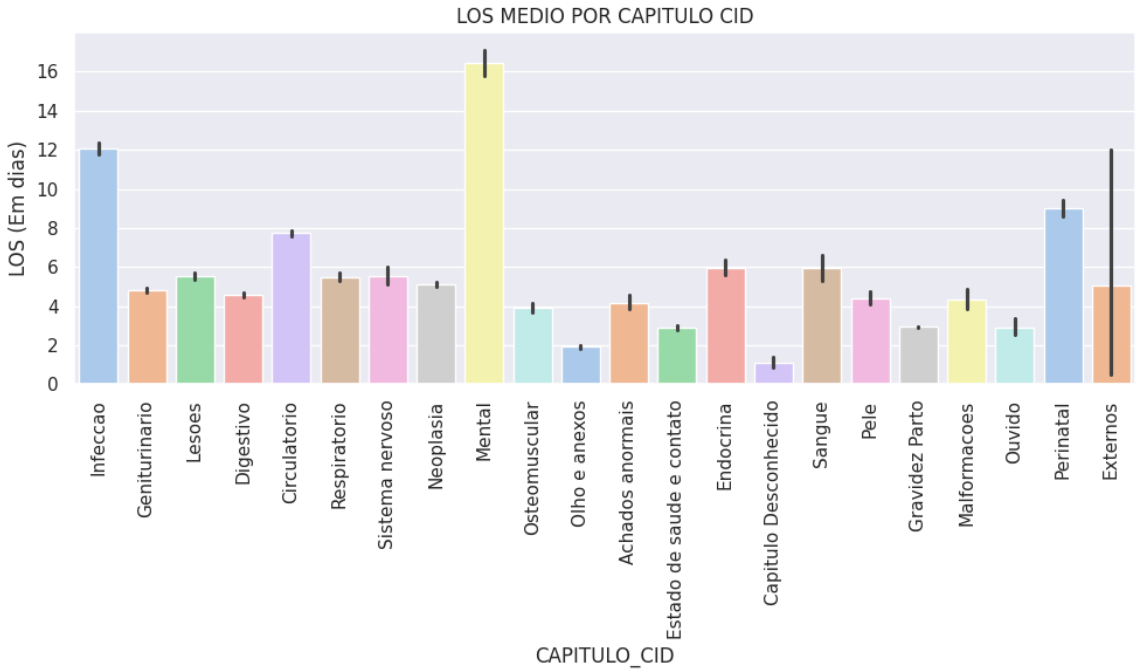

Figura 6 – Tempo médio de internação por capítulo do CID.
Fonte: elaboração própria (dados Hospital das Clínicas Botucatu).

3.1.5 Procedimentos e intensidade assistencial

A relação entre tempo de internação e variáveis de intensidade assistencial mostrou comportamento crescente. A Figura 7 indica aumento gradual do LOS conforme o número de cirurgias realizadas, especialmente acima de três procedimentos. De forma semelhante, observa-se nas Figuras 8 e 9 que pacientes com maior número de exames laboratoriais e uso mais frequente de antibióticos permaneceram internados por períodos mais longos, o que pode ser interpretado como indicativo indireto de maior gravidade clínica (GARCIA et al., 2015; GERON, 2019).

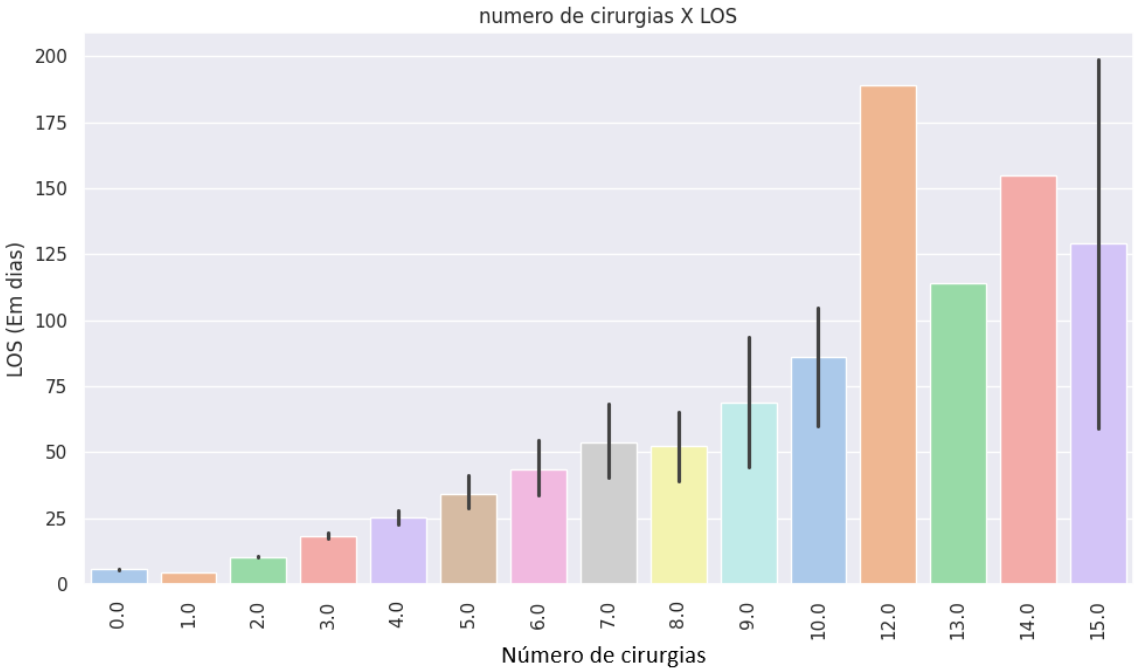

Figura 7 – Tempo de internação segundo o número de cirurgias.
Fonte: elaboração própria (dados Hospital das Clínicas Botucatu).

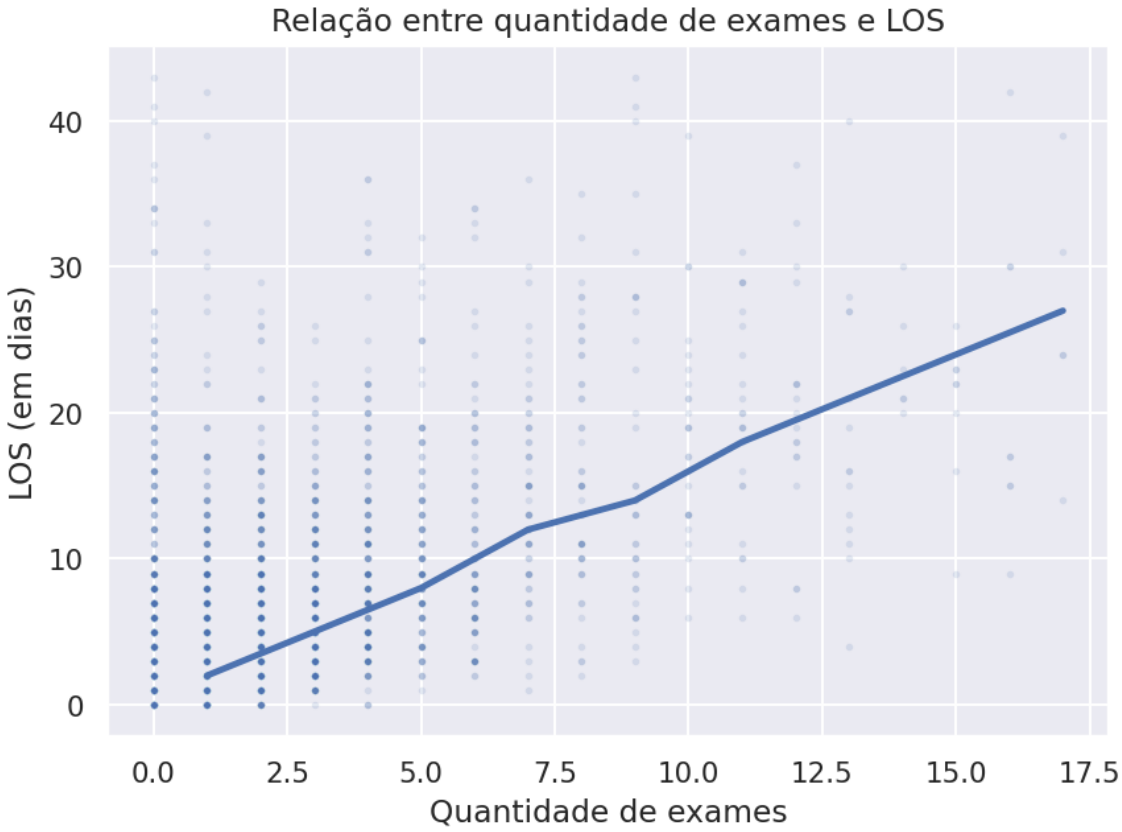

Figura 8 – Relação entre quantidade de exames e tempo de internação.
Fonte: elaboração própria (dados Hospital das Clínicas Botucatu).

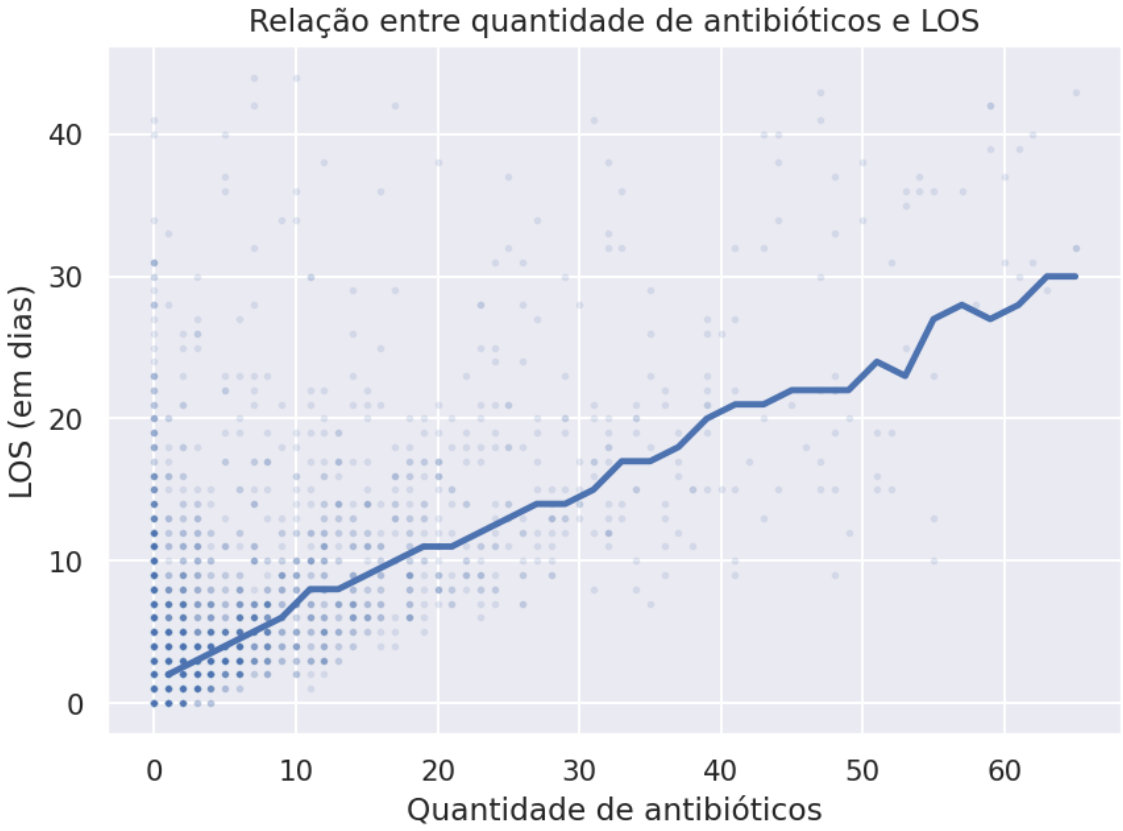

Figura 9 – Relação entre quantidade de antibióticos e tempo de internação.
Fonte: elaboração própria (dados Hospital das Clínicas Botucatu).

De modo geral, os resultados descritivos confirmam tendências esperadas: predominância de internações curtas, maior permanência em pacientes idosos e casos urgentes, e relação direta entre complexidade terapêutica e duração hospitalar. Essas evidências sustentam a validade das variáveis selecionadas para o modelo preditivo, apresentado na próxima seção (CUTLER et al., 2007; KOTSIANTIS et al., 2006).

3.2 Desempenho do Modelo e Importância das Variáveis

Após o tratamento e a análise exploratória, procedeu-se à etapa de avaliação quantitativa do modelo de regressão. O algoritmo *Random Forest Regressor* foi treinado conforme descrito na Seção 2.5, e seu desempenho foi aferido com base nas métricas de regressão apresentadas na Seção 2.6. O modelo demonstrou capacidade satisfatória de generalização, alcançando valores de erro e coeficiente de determinação compatíveis com a literatura de modelagem hospitalar (BREIMAN, 2001; MULLER; GUIDO, 2016).

3.2.1 Métricas de desempenho

A Tabela 1 resume as métricas obtidas para o modelo final de *Random Forest*. Observa-se que o modelo apresentou $R^2 = 0,675$, indicando que aproximadamente 67,5% da variabilidade

do tempo de internação foi explicada pelas variáveis utilizadas. O erro absoluto médio (MAE) foi de 2,40 dias, enquanto o erro quadrático médio (RMSE) alcançou 5,72 dias. Esses resultados indicam que o modelo foi capaz de prever com razoável precisão o tempo de permanência hospitalar, mantendo erros médios clinicamente aceitáveis (HASTIE et al., 2009; KUHN; JOHNSON, 2013).

Tabela 1 – Desempenho do modelo *Random Forest Regressor*.

Métrica	**Símbolo**	**Valor**	**Interpretação**
Coeficiente de determinação	R^2	0,675	Correlação predição–valor real
Erro Absoluto Médio	MAE	2,40	Diferença média em dias
Raiz do Erro Quadrático Médio	RMSE	5,72	Desvio médio ponderado por erro alto

Em termos práticos, o valor de R^2 demonstra que o modelo explica mais da metade da variação observada, o que é considerado satisfatório para dados clínicos reais, que usualmente apresentam alta dispersão e múltiplos fatores não observáveis (GARCIA et al., 2015; JAMES et al., 2021). O MAE de cerca de dois dias sugere que, em média, as previsões diferem pouco do valor real, o que pode auxiliar no planejamento de alta e no uso mais eficiente dos leitos hospitalares (FIELD, 2013; SHMUELI; KOPPIUS, 2011).

3.2.2 Importância das variáveis

A análise de importância das variáveis, obtida a partir do atributo `feature_importances_` da biblioteca `scikit-learn` (PEDREGOSA et al., 2011), evidenciou os atributos com maior influência sobre o tempo de internação. A Figura 10 mostra a contribuição relativa de cada variável no modelo treinado.

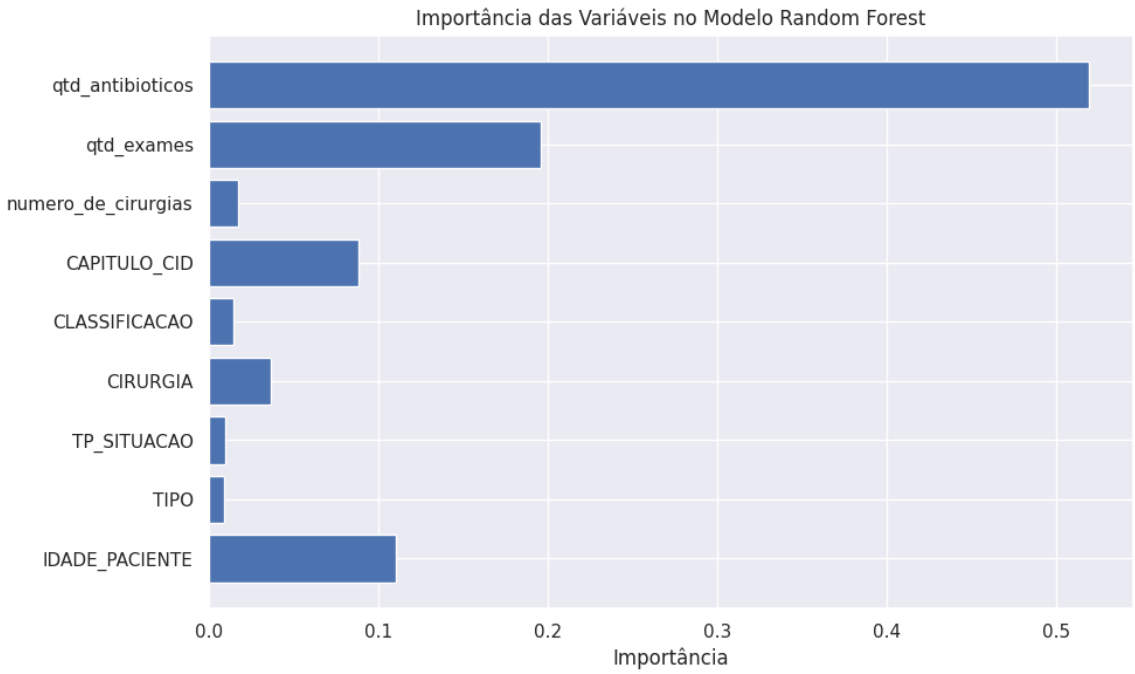

Figura 10 – Importância relativa das variáveis no modelo *Random Forest*.
Fonte: elaboração própria (dados Hospital das Clínicas Botucatu).

Os resultados indicaram predominância das variáveis quantidade de antibióticos e quantidade de exames como principais determinantes do modelo, seguidas por número de cirurgias, capítulo CID e idade do paciente. Essas variáveis refletem diretamente a intensidade assistencial e a complexidade do caso clínico (SANTOS et al. (2021) e COSTA et al. (2022).

A interpretação do gráfico mostra que a variável *qtd_antibioticos* responde por aproximadamente metade da importância total, o que sugere forte relação entre uso prolongado de antibióticos e gravidade de infecções hospitalares. A *qtd_exames* aparece como o segundo fator mais relevante, indicando que internações que demandam maior volume de exames laboratoriais tendem a ter duração maior, possivelmente devido à necessidade de monitoramento contínuo e ajustes terapêuticos (HAN; KAMBER; PEI, 2023; TAN et al., 2019).

3.2.3 Síntese dos achados preditivos

De forma geral, o modelo apresentou bom equilíbrio entre desempenho estatístico e interpretabilidade. O erro médio reduzido e o R^2 acima de 0,67 reforçam que o modelo conseguiu capturar de maneira consistente a tendência do tempo de internação. Os resultados obtidos são coerentes com estudos que utilizaram algoritmos de ensemble em contextos hospitalares, demonstrando que o *Random Forest* é capaz de combinar desempenho robusto e compreensão clínica (CUTLER et al., 2007; GERON, 2019).

As variáveis mais influentes identificadas podem subsidiar decisões estratégicas, como previsão de alta e alocação de recursos. Esses achados indicam que, mesmo com uma base

de dados limitada a variáveis estruturadas, é possível alcançar níveis de acurácia relevantes e aplicáveis à realidade hospitalar (COSTA et al., 2022; GARCIA et al., 2015).

3.3 Interpretação dos Resultados e Aplicações Práticas

Os achados obtidos nas Seções 3.1 e 3.2 indicam que o modelo de *Random Forest* alcançou desempenho consistente na previsão do tempo de internação, com $R^2 = 0{,}675$, MAE de aproximadamente 2,40 dias e RMSE de 5,72 dias. Em termos operacionais, um erro absoluto médio de cerca de dois dias pode apoiar decisões de planejamento com janelas de segurança adequadas para previsão de altas, giro de leitos e organização de fluxos de pacientes. A interpretação é coerente com o comportamento assimétrico do LOS (cauda longa) e com a predominância de internações curtas observada na análise descritiva, quadro no qual erros maiores em poucos casos prolongados elevam o RMSE (HASTIE et al., 2009; KUHN; JOHNSON, 2013; KOTSIANTIS et al., 2006).

A hierarquia de importância de variáveis aponta *qtd_antibioticos* e *qtd_exames* como principais preditores, seguidas por número de cirurgias, capítulo do CID e idade. Essas variáveis são *proxies* de complexidade clínica e intensidade assistencial, refletindo maior severidade dos casos e, portanto, maiores tempos de permanência. Do ponto de vista interpretativo, a floresta aleatória oferece uma medida direta via `feature_importances_` que, mesmo não sendo explicação causal, auxilia na priorização de fatores associados ao aumento do LOS para fins de gestão (BREIMAN, 2001; MULLER; GUIDO, 2016; JAMES et al., 2021).

3.3.1 Aplicações na gestão hospitalar

Gestão de leitos e previsão de altas. As estimativas de LOS podem ser integradas a painéis operacionais para apoiar a programação diária/semanal de altas, antecipar gargalos e ajustar a ocupação por unidade. Com MAE próximo de dois dias, recomenda-se utilizar faixas de confiança (janelas de tolerância) e atualizar as previsões periodicamente ao longo da internação, reduzindo a incerteza conforme novas informações clínicas entram no prontuário (GERON, 2019; HAN; KAMBER; PEI, 2023).

Programação cirúrgica e fluxo perioperatório. O padrão observado de LOS maior em casos clínicos e urgentes, e o incremento do LOS com o número de cirurgias, sugere uso das previsões para balancear agendas cirúrgicas, considerando especialidades e perfis de risco, mitigando picos de ocupação em UTI/enfermaria e alinhando o giro de leitos à demanda programada (FIELD, 2013; OLIVEIRA et al., 2020; COSTA et al., 2022).

Alocação de recursos diagnósticos e terapêuticos. A relevância de *qtd_exames* e *qtd_antibioticos* sugere que, em pacientes com previsão de LOS prolongado, pode ser útil antecipar disponibilidade de exames, insumos e equipes multidisciplinares (farmácia clínica, controle de infecção), de modo a reduzir tempos de espera e otimizar trajetórias assistenciais

(GARCIA et al., 2015; TAN et al., 2019).

Sinalização precoce de internações potencialmente longas. As previsões podem acionar *flags* operacionais para casos com risco de permanência acima de limiares específicos (por exemplo, os 25% ou 10% de pacientes com as permanências mais longas), priorizando intervenções gerenciais (discharge planning, conciliação medicamentosa, avaliação multiprofissional). Tais usos são comuns em iniciativas de analítica preditiva em saúde, desde que acompanhados de supervisão clínica e monitoramento de desempenho (RAJKOMAR et al., 2019; SONG et al., 2020) .

3.3.2 Limitações e Boas práticas

Entre as principais limitações do estudo, destacam-se a utilização de dados provenientes de uma única instituição e a ausência de variáveis clínicas detalhadas, como gravidade do quadro, comorbidades e evolução diária do paciente. Tais fatores restringem a generalização dos resultados para outros contextos hospitalares. Além disso, a natureza estática do modelo, treinado sobre dados históricos, não contempla possíveis variações temporais associadas à sazonalidade, epidemias ou mudanças de protocolo clínico.

Como padrões assistenciais e perfis epidemiológicos variam no tempo, recomenda-se reavaliar métricas (MAE, RMSE, R^2) e recalibrar o modelo em janelas regulares (por exemplo, trimestral/semestral), além de monitorar *data drift* e *concept drift* para manter acurácia e utilidade operacional (HASTIE et al., 2009; PEDREGOSA et al., 2011).

Deve-se verificar sistematicamente se o erro do modelo é sistematicamente maior em subgrupos (sexo, faixas etárias, capítulos do CID), ajustando variáveis e processos conforme necessário para reduzir vieses de predição que possam impactar decisões assistenciais ou de alocação de recursos (RAJKOMAR et al., 2019; KUHN; JOHNSON, 2013).

O uso dos dados para fins analíticos deve observar anonimização, controle de acesso e base legal adequada, em conformidade com a Lei Geral de Proteção de Dados Pessoais (Lei nº 13.709/2018), bem como com pareceres de Comitês de Ética quando aplicável (BRASIL, 2018; SHMUELI; KOPPIUS, 2011).

Embora a importância de variáveis de Random Forest não constitua explicação causal, seu uso em *dashboards* com contexto clínico (por exemplo, apresentar as variáveis mais influentes na previsão individual) favorece a adoção por equipes e a validação de plausibilidade, evitando decisões automatizadas sem revisão humana (CUTLER et al., 2007; RAJKOMAR et al., 2019).

Em síntese, os resultados sustentam aplicações práticas imediatas em gestão de leitos, planejamento de altas, programação cirúrgica e priorização de recursos. A combinação de desempenho quantitativo (MAE, RMSE, R^2) e interpretabilidade via importância de variáveis torna a abordagem promissora para apoio operacional (GERON, 2019; BREIMAN, 2001; BRASIL, 2018).

4 Conclusão

O presente trabalho apresentou o desenvolvimento de um modelo de aprendizado de máquina voltado à predição do tempo de internação hospitalar a partir de dados estruturados de prontuários de pacientes. Por meio da aplicação do algoritmo *Random Forest Regressor*, buscou-se construir uma ferramenta capaz de apoiar a gestão hospitalar na previsão de permanências em leitos de internação e na otimização de recursos assistenciais.

Os resultados alcançados indicaram desempenho consistente, com coeficiente de determinação (R^2) de aproximadamente 0,675, erro absoluto médio (MAE) de 2,40 dias e erro quadrático médio (RMSE) de 5,72 dias. Tais valores demonstram que o modelo foi capaz de reproduzir, com boa precisão, o comportamento do tempo de internação, validando a hipótese de que dados clínicos e administrativos estruturados podem ser empregados de forma eficaz na modelagem preditiva hospitalar. Em conjunto com a análise descritiva, observou-se que variáveis associadas à complexidade assistencial — como quantidade de antibióticos, número de exames e de cirurgias — exercem papel determinante sobre a duração das internações.

Para trabalhos futuros, recomenda-se a ampliação da base de dados com variáveis de evolução clínica e anotações médicas, integrando abordagens de *Natural Language Processing* (NLP) para extrair informações de texto livre. Outras perspectivas incluem o teste de algoritmos complementares, como *Gradient Boosting* e *XGBoost*, além da aplicação de técnicas de validação cruzada e ajuste automatizado de hiperparâmetros.

Por fim, conclui-se que o modelo proposto apresentou resultados satisfatórios e alinhados à literatura, demonstrando potencial de aplicação na prática hospitalar. O emprego de técnicas de aprendizado de máquina em dados administrativos e clínicos representa um passo importante para a modernização da gestão hospitalar, oferecendo subsídios à tomada de decisão e contribuindo para o uso mais racional e eficiente dos recursos disponíveis.

Referências

1 BREIMAN, L. Random forests. *Machine Learning*, v.45, n.1, p.5–32, 2001.

2 BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Diário Oficial da União, 2018.

3 COSTA, M. E. et al. Perfil hospitalar e fatores associados à permanência prolongada. *Revista Brasileira de Epidemiologia*, v.25, 2022.

4 FIELD, A. *Discovering Statistics Using IBM SPSS Statistics*. 4. ed. Londres: Sage, 2013.

5 HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 4. ed. Cambridge: Morgan Kaufmann, 2023.

6 JAMES, G. et al. *An Introduction to Statistical Learning: With Applications in Python*. Nova York: Springer, 2021.

7 McKINNEY, W. Data structures for statistical computing in Python. In: *Proceedings of the 9th Python in Science Conference*. Austin, 2010.

8 OLIVEIRA, A. L. et al. Padrões de permanência hospitalar por perfil demográfico e tipo de internação. *Cadernos de Saúde Coletiva*, v.28, n.4, p.491–500, 2020.

9 PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, v.12, p.2825–2830, 2011.

10 RAJKOMAR, A. et al. Machine Learning in Medicine. *New England Journal of Medicine*, v.380, n.14, p.1347–1358, 2019.

11 SHMUELI, G.; KOPPIUS, O. R. Predictive analytics in information systems research. *MIS Quarterly*, v.35, n.3, p.553–572, 2011.

12 TUKEY, J. W. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley, 1977.

13 GARCIA, S. et al. A survey of preprocessing techniques for data mining. *Data Mining and Knowledge Discovery*, v.29, p.133–173, 2015.

14 GERON, A. Hands-On Machine Learning with Scikit-Learn and TensorFlow. 2. ed. O'Reilly Media, 2019.

15 TAN, P.; STEINBACH, M.; KUMAR, V. Introduction to Data Mining. 2. ed. Pearson, 2019.

16 CUTLER, D. R. et al. Random Forests for classification in ecology. *Ecology*, v.88, n.11, p.2783–2792, 2007.

17 MULLER, A.; GUIDO, S. Introduction to Machine Learning with Python. O'Reilly Media, 2016.

18 SONG, X. et al. Predicting hospital length of stay using machine learning: A systematic review. *BMC Medical Informatics and Decision Making*, v.20, p.1–16, 2020.

19 KUHN, M.; JOHNSON, K. Applied Predictive Modeling. Springer, 2013.

20 HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning. 2. ed. Springer, 2009.

21 SONG, X. et al. Predicting hospital length of stay using machine learning: A systematic review. *BMC Medical Informatics and Decision Making*, v.20, p.1–16, 2020.

22 KOTSIANTIS, S. B.; KANELLOPOULOS, D.; PINTELAS, P. Data preprocessing for supervised learning. *International Journal of Computer Science*, v.1, n.2, p.111–117, 2006.