

Modelo Dinâmico de Análise de Sobrevivência Aplicado à Previsão do Tempo de Internação em Hospitais Públicos Municipais

Gizeuda Cariry Lacet de Barros Crispim



CENTRO DE INFORMÁTICA UNIVERSIDADE FEDERAL DA PARAÍBA



Gizeuda Cariry Lacet de Barros Crispim

Modelo Dinâmico de Análise de Sobrevida Aplicado à Previsão do Tempo de Internação em Hospitais Públicos Municipais

Dissertação apresentada ao Programa de Pós-Graduação em Informática da
Universidade Federal da Paraíba.

Orientador: Prof. Dr. Guido Lemos de Souza
Filho

Coorientador: Prof. Dr. Claurton de
Albuquerque Siebra

Catálogo na publicação
Seção de Catalogação e Classificação

C932m Crispim, Gizeuda Cariry Lacet de Barros.
Modelo Dinâmico de Análise de Sobrevivência Aplicado
à Previsão do Tempo de Internação em Hospitais Públicos
Municipais / Gizeuda Cariry Lacet de Barros Crispim. -
João Pessoa, 2026.
63 f. : il.

Orientação: Guido Lemos Souza Filho.
Coorientação: Claurirton de Albuquerque Siebra.
Dissertação (Mestrado) - UFPB/CI.

1. Análise de sobrevivência dinâmica. 2. Tempo de
internação hospitalar. 3. Modelagem preditiva. 4.
Aprendizado de máquina. 5. Gestão de leitos
hospitalares. I. Souza Filho, Guido Lemos. II. Siebra,
Claurirton de Albuquerque. III. Título.

UFPB/BC

CDU 004.8:614.2(043)



UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA



Ata da Sessão Pública de Defesa de Dissertação de Mestrado de **GIZEUDA CARIRY LACET DE BARROS CRISPIM**, candidato ao título de Mestre em Informática na área de Ciências da Computação, realizada em 27 de fevereiro de 2026.

Aos vinte e sete dias do mês de fevereiro do ano de dois mil e vinte e seis, às 14h, via videoconferência, reuniram-se os membros da Banca Examinadora constituída para julgar o Trabalho Final da discente **GIZEUDA CARIRY LACET DE BARROS CRISPIM**, vinculado a Universidade Federal da Paraíba sob a matrícula nº 20241008159, candidata ao grau de Mestre em Informática, na área de “Ciência da Computação”, na linha de pesquisa “Metodologia e Técnicas de Computação” do Programa de Pós-Graduação em Informática. A comissão examinadora foi composta pelos professores Guido Lemos de Souza Filho, orientador e presidente da banca; Claurton de Albuquerque Siebra, examinador interno ao programa; Dênio Mariz Timoteo de Sousa e Cicero Inacio da Silva, ambos examinadores externos à Instituição. Dando início aos trabalhos, o presidente cumprimentou os presentes, comunicou a finalidade da reunião e passou a palavra ao candidato para que fizesse, oralmente, a exposição do trabalho de dissertação intitulado “**Modelo Dinâmico de Análise de Sobrevivência Aplicado à Previsão do Tempo de Internação em Hospitais Públicos Municipais**”. Concluída a exposição, a candidata foi arguida pela Banca Examinadora, que emitiu o seguinte parecer: “*aprovado*”. Do ocorrido, eu, Gean Paulo P. M. de Barros, secretário, lavrei a presente ata que vai assinada por mim e pelos membros da Banca Examinadora.

João Pessoa, 27 de fevereiro de 2026.

Gean Paulo P. M. de Barros
Secretário - SIAPE 2326476

Prof. Dr. Guido Lemos de Souza Filho
Orientador (PPGI)

Prof. Dr. Claurton de Albuquerque Siebra
Examinador interno ao Programa (PPGI)

Documento assinado digitalmente
gov.br GUIDO LEMOS DE SOUZA FILHO
Data: 06/03/2026 16:40:02-0300
Verifique em <https://validar.iti.gov.br>

Documento assinado digitalmente
gov.br CLAUIRTON DE ALBUQUERQUE SIEBRA
Data: 06/03/2026 13:56:48-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Dênio Mariz Timoteo de Sousa
Examinador externo à Instituição (IFPB)

Documento assinado digitalmente
 **DENIO MARIZ TIMOTEO DE SOUSA**
Data: 15/03/2026 13:48:32-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. Cicero Inacio da Silva
Examinador externo à Instituição (Unifesp)

Documento assinado digitalmente
 **CICERO INACIO DA SILVA**
Data: 11/03/2026 10:36:12-0300
Verifique em <https://validar.iti.gov.br>

Centro de Informática, Universidade Federal da Paraíba
Rua dos Escoteiros, Mangabeira VII, João Pessoa, Paraíba, Brasil
CEP: 58058-600 Fone: +55 (83) 3216 7093 / Fax: +55 (83) 3216 7117

AGRADECIMENTOS

Primeiramente à Deus, Por sempre iluminar meu caminho e nunca me abandonar, principalmente nos momentos mais difíceis junto com a intercessão da Virgem Maria, minha mãezinha, sempre me cobrindo com seu manto protetor.

À minha família, agradeço de coração todo o apoio, carinho e paciência em mais um momento tão especial da minha vida. Não teria chegado até aqui sem eles. Em especial, à minha filha, motivação de todo meu progresso.

Ao meu orientador, Professor Guido Lemos de Sousa Filho e meu coorientador professor Claurton de Albuquerque Siebra, minha profunda gratidão. Por cada orientação, cada troca de experiência e conselho dado, que foram fundamentais para a construção deste trabalho. Vocês são mestres e profissionais inspiradores.

Aos colegas da pós-graduação que me ajudaram na prática deste estudo, serei sempre grata e disponível para a recíproca.

À coordenação do programa de pós-graduação do Centro de Informática por toda oportunidade e estrutura fornecida para a realização desta jornada.

Aos professores do Centro de Informática, por toda contribuição na minha vida acadêmica e profissional.

E por fim, mas com a mesma importância, a todos os amigos que fiz durante minha caminhada nesta pós-graduação, que me ajudaram, compartilharam as dificuldades e as dores, tornando os momentos e aprendizagem mais leves.

O meu mais sincero obrigada a todos!

RESUMO

A superlotação dos serviços de saúde impõe desafios significativos à gestão hospitalar, sendo o tempo de internação (Length of Stay – LOS) um dos principais indicadores de desempenho e planejamento assistencial. Neste contexto, este trabalho propõe uma abordagem de análise de sobrevivência dinâmica para previsão do tempo de internação hospitalar, incorporando atualização temporal das estimativas ao longo da permanência do paciente. A proposta é avaliada a partir de dados reais do sistema hospitalar público de saúde de João Pessoa – PB, com o objetivo de oferecer um modelo preditivo mais aderente à natureza evolutiva do risco de alta hospitalar. Como etapa de fundamentação e validação, a abordagem proposta é analisada em comparação com métodos consolidados na literatura, incluindo o Modelo de Riscos Proporcionais de Cox (CoxPH), Random Survival Forest e DeepSurv. O estudo contempla conceitos de aprendizagem de máquina e aprendizagem profunda, técnicas de pré-processamento de dados e fundamentos da análise de sobrevivência, estruturando o desenvolvimento do modelo em um pipeline clássico de aprendizado de máquina. O principal produto do trabalho é um modelo preditivo capaz de estimar probabilidades condicionais e dinâmicas associadas ao tempo de internação, permitindo atualizações conforme novas informações clínicas se tornam disponíveis. O desempenho é avaliado por meio do índice de concordância (C-index), possibilitando a comparação entre a abordagem dinâmica proposta e os modelos tradicionais. Resultados indicam que a modelagem dinâmica apresenta potencial para aprimorar a acurácia preditiva e a utilidade prática das estimativas, reforçando a viabilidade da aplicação de técnicas de inteligência artificial no desenvolvimento de sistemas inteligentes de apoio à decisão clínica e à gestão eficiente de leitos hospitalares.

Palavras-chave: Análise de Sobrevivência Dinâmica; Tempo de Internação Hospitalar; Modelagem Preditiva; Aprendizado de Máquina; Gestão de Leitos Hospitalares; Sistemas de Apoio à Decisão.

ABSTRACT

Healthcare service overcrowding poses significant challenges to hospital management, with Length of Stay (LOS) being one of the main performance and planning indicators. In this context, this study proposes a dynamic survival analysis approach to predict hospital length of stay, incorporating temporal updating of risk estimates throughout the patient's hospitalization. The proposed framework is evaluated using real-world data from the public hospital system of João Pessoa, Brazil, aiming to provide predictive estimates that better reflect the evolving nature of discharge risk over time. As part of the validation process, the proposed dynamic approach is assessed in comparison with well-established survival models, including the Cox Proportional Hazards model (CoxPH), Random Survival Forest, and DeepSurv. The study encompasses fundamental concepts of machine learning and deep learning, data preprocessing techniques, and survival analysis theory, structuring model development within a classical machine learning pipeline. The main outcome of this research is a predictive model capable of estimating conditional and dynamically updated probabilities associated with hospital length of stay, allowing recalibration as new clinical information becomes available. Model performance is evaluated using the concordance index (C-index), enabling comparison between the proposed dynamic framework and traditional approaches. Results indicate that dynamic modeling has the potential to enhance predictive accuracy and practical applicability, reinforcing the feasibility of artificial intelligence techniques in developing intelligent clinical decision-support systems and improving hospital bed management efficiency.

Keywords: Dynamic Survival Analysis; Hospital Length of Stay; Predictive Modeling; Machine Learning; Hospital Bed Management; Clinical Decision Support Systems. .

LISTA DE ABREVIATURAS

ARIMA – Auto regressive integrated moving average
C-INDEX – Índice de concordância
CNN – Rede neural convolucional
CoxPH - Cox Proportional Hazards
EDA - Análise Exploratória de Dados
EHR - Prontuário eletrônicos de saúde
GBM - Gradient Boosting Machine
IA -Inteligência artificial
JSON - JavaScript Object Notation
LOS – Length of stay
MAPE - Mean absolute percentage error
ML – Machine learning
MAE – Erro absoluto médio
MSE – Erro quadrático médio
PCA - Principal component analysis
PLN - Processamento de linguagem natural
RSF - Random Survival Forest
SARIMA – Seasonal regressive integrated moving average
SUS - Sistema único de saúde
SVD - Decomposição de valores singulares truncados
SVM - Máquinas de vetores de suporte
TF-IDF - Frequência de termos-frequência inversa do documento
VPN - Virtual private network
XGBoost - Extreme Gradient Boosting
XML - Extensible Markup Language

LISTA DE FIGURAS

Figura 1. Modelo de riscos proporcionais de Cox	22
Figura 2. Estrutura de um Random Survival Forest	24
Figura 3. Exemplo de roteiro de criação de modelos preditivos	27
Figura 4. Gráfico de dispersão de dados com outliers	30
Figura 5. Distribuição dos grupos CID na fase 1	41
Figura 6. Distribuição dos grupos CID na fase 2	42
Figura 7. Probabilidade cumulativa do evento no tempo remanescente estimada em diferentes dias de referência (landmark days).....	52

LISTA DE QUADROS

Quadro 1. Tabela comparativa de trabalhos realizados com o mesmo tema.....	36
Quadro 2. Quadro comparativo de métodos	38

LISTA DE TABELAS

Tabela 1. Exemplo de dataset usado durante a modelagem dinâmica	45
Tabela 2. Tabela comparativa de experimentos da primeira rodada.....	49
Tabela 3. Tabela comparativa de experimentos da segunda rodada.....	50
Tabela 4. Tabela comparativa clássico x dinâmico	50

SUMÁRIO

1 INTRODUÇÃO	14
1.1 PROBLEMA.....	14
1.2 OBJETIVOS	15
1.3 ESTRUTURA DA DISSERTAÇÃO	16
2 FUNDAMENTAÇÃO TEÓRICA	17
2.1 INTELIGÊNCIA ARTIFICIAL NA SAÚDE	17
2.2 TEMPO DE PERMANÊNCIA HOSPITALAR: UM PROCESSO DE TEMPO ATÉ EVENTO.....	18
2.3 ANÁLISE DE SOBREVIVÊNCIA: CONCEITUAÇÃO	19
2.3.1 Modelo de Riscos Proporcionais de COX (COXPH)	21
2.3.2 Random Survival Forest	23
2.3.3 DeepSurv.....	24
2.4 MODELAGEM DINÂMICA DO RISCO E ATUALIZAÇÃO TEMPORAL	25
2.5 PIPELINE CLÁSSICO PARA CONSTRUÇÃO DE MODELOS PREDITIVOS	26
2.5.1 Visão Geral.....	26
2.5.2 Coleta de dados	28
2.5.3 Exploração de dados.....	28
2.5.4 Limpeza de dados	29
2.5.5 Transformação de dados.....	31
2.5.6 Redução de dados	31
2.5.7 Divisão de dados	32
3 TRABALHOS RELACIONADOS	34
4 METODOLOGIA	38
4.1 CONJUNTO DE DADOS.....	38
4.2 MÉTODOS DE ANÁLISE DE SOBREVIVÊNCIA CLÁSSICOS	38
4.3 ATIVIDADES	39
4.3.1 Setup	39
4.3.2 Fase 1 - Experimentos com dados de todos os hospitais	42
4.3.3 Fase 2 - Utilizando dados dos hospitais gerais	44
4.3.4 Modelagem Dinâmica.....	44

5 RESULTADOS	48
5.1 MODELOS CLÁSSICOS	48
5.2 MODELO DINÂMICO	50
5.3 DISCUSSÃO	51
5.4 PERCEPÇÃO DO USUÁRIO FINAL	53
6 CONCLUSÃO	55
6.1 LIMITAÇÕES DO ESTUDO	56
6.2 TRABALHOS FUTUROS	57
REFERÊNCIAS	58

1 INTRODUÇÃO

Em diversos contextos ao redor do mundo, a superlotação dos serviços de saúde tem se consolidado como um dos principais desafios da gestão hospitalar, impactando diretamente o tempo de internação (*Length of Stay* – LOS), os custos assistenciais, o estresse das equipes multiprofissionais e a satisfação dos pacientes (Ataman; Sariyer, 2022). Nesse cenário, o gerenciamento eficiente do fluxo de pacientes tornou-se estratégico para a sustentabilidade dos sistemas de saúde. A capacidade de prever o tempo de permanência hospitalar desde a admissão até a alta possibilita aprimorar o planejamento de recursos, a gestão de leitos e a organização de cuidados pós-agudos, produzindo impactos relevantes na qualidade assistencial e na eficiência operacional (Medeiros *et al.*, 2021; Karbouband; Tabaa, 2023).

Nos últimos anos, observa-se um crescimento expressivo do uso de técnicas de aprendizado de máquina (*machine learning*) e aprendizado profundo (*deep learning*) na área da saúde, especialmente na modelagem preditiva de eventos clínicos (Stone *et al.*, 2022; Shickel *et al.*, 2018). Entre essas aplicações, a análise de sobrevivência destaca-se por sua adequação à modelagem de desfechos do tipo “tempo até o evento”, como a alta hospitalar. Diferentemente de abordagens tradicionais baseadas em regressão, os métodos de sobrevivência lidam de forma apropriada com dados censurados, situação comum em bases hospitalares, nas quais pacientes podem permanecer internados no momento da análise (Collett, 2023).

1.1 PROBLEMA

As técnicas de análise de sobrevivência evoluíram substancialmente, desde modelos estatísticos clássicos, como o modelo de Riscos Proporcionais de Cox (Cox, 1972), até abordagens contemporâneas baseadas em aprendizado profundo, como DeepSurv e DeepHit. Entretanto, a maioria dos estudos concentra-se em modelos estáticos, que produzem estimativas fixas com base em informações coletadas em um único ponto temporal, geralmente na admissão hospitalar. Tal abordagem desconsidera que o risco de alta é inerentemente dinâmico, variando ao longo da internação em função da evolução clínica do paciente, intervenções terapêuticas e intercorrências.

Nesse contexto, identifica-se uma lacuna científica relevante: ainda são limitadas as investigações que exploram, de forma aplicada e com dados reais de prontuários eletrônicos, modelos de sobrevivência que incorporem atualização temporal do risco e permitam estimativas condicionais ao tempo já decorrido de internação. Além disso, poucos estudos avaliam sistematicamente se a incorporação dessa dinâmica resulta em ganhos preditivos e maior utilidade prática para a gestão hospitalar quando comparada a abordagens tradicionais.

Dessa forma, este trabalho parte da hipótese de que uma abordagem de análise de sobrevivência dinâmica, capaz de atualizar as probabilidades de alta à medida que novas informações clínicas são incorporadas, apresenta desempenho preditivo superior e maior aplicabilidade gerencial em comparação com modelos estáticos convencionais. Ao propor e avaliar essa estrutura metodológica em um cenário real do sistema público hospitalar, busca-se contribuir tanto para o avanço metodológico da modelagem de tempo até evento em saúde quanto para o desenvolvimento de sistemas inteligentes de apoio à decisão voltados à otimização da gestão de leitos e do planejamento assistencial.

1.2 OBJETIVOS

Esta pesquisa tem o objetivo de propor e avaliar uma abordagem de análise de sobrevivência dinâmica para a previsão do tempo de permanência hospitalar, incorporando atualização temporal das estimativas de risco e analisando seu desempenho e aplicabilidade em dados reais do sistema público de saúde. Os seguintes objetivos específicos são derivados dessa visão geral:

- Estruturar e implementar um pipeline de modelagem preditiva baseado em análise de sobrevivência, contemplando etapas de pré-processamento, seleção de variáveis e tratamento de dados censurados;
- Desenvolver um modelo de sobrevivência dinâmica capaz de atualizar probabilidades condicionais de alta hospitalar ao longo do tempo de internação;
- Implementar modelos de sobrevivência tradicionais amplamente utilizados na literatura (e.g., CoxPH, Random Survival Forest e DeepSurv) para fins de referência comparativa;

- Avaliar o desempenho preditivo da abordagem dinâmica e dos modelos tradicionais por meio de métricas apropriadas, como o índice de concordância (C-index), analisando ganhos em discriminação e utilidade prática;
- Analisar a aplicabilidade gerencial do modelo proposto no contexto da gestão de leitos hospitalares e do apoio à decisão clínica em hospitais públicos.

1.3 ESTRUTURA DA DISSERTAÇÃO

Além desta introdução, que apresenta o contexto, o problema e os objetivos da dissertação, o restante do documento traz a seguinte estrutura. O Capítulo 2 discorre sobre os principais conceitos pertinentes ao entendimento desta dissertação, ou seja, a fundamentação teórica. O Capítulo 3 analisa trabalhos que são o estado da arte da aplicação de abordagens de análise de sobrevivência aplicados ao problema do tempo de internação/permanência hospitalar. O Capítulo 4 apresenta a metodologia utilizada como guia no desenvolvimento deste trabalho e avaliação dos seus resultados. O Capítulo 5 apresenta e discute os resultados obtidos, principalmente com a comparação dos modelos atualmente empregados com nossa abordagem dinâmica. Finalmente, o Capítulo 6 conclui este trabalho, destacando as principais contribuições, limitações e oportunidades para continuação desta pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 INTELIGÊNCIA ARTIFICIAL NA SAÚDE

A Inteligência Artificial (IA) exerce grande influência em todos os aspectos da vida humana, se destacando na saúde. Na atualidade, grandes bancos de dados e o uso da IA trazem melhorias em todos os aspectos do sistema de saúde, como prevenção, triagem e tratamento de doenças (Rymarz *et al.*, 2023). Segundo Silva *et al.* (2024), na medicina a IA destaca-se pela sua capacidade analítica em lidar com grandes volumes de dados, seguindo algoritmos delineados por especialistas, e oferece soluções rápidas para diversos desafios médicos com esforços direcionados para o desenvolvimento de sistemas computadorizados para diagnósticos médicos.

Já Pee *et al.* (2019), reforçam que a IA pode melhorar significativamente o atendimento ao paciente e, ao mesmo tempo, reduzir as despesas com saúde. Portanto, soluções inovadoras de IA são necessárias no setor de saúde para melhorar a eficácia e a eficiência sem aumentar os custos.

A regulamentação da inteligência artificial na saúde apresenta convergência internacional quanto à necessidade de supervisão humana, validação científica e garantia da segurança do paciente. No Brasil, a Resolução CFM nº 2.431/2025 estabelece que tecnologias digitais e sistemas baseados em inteligência artificial devem atuar apenas como instrumentos de apoio à prática médica, permanecendo o médico responsável pela decisão clínica, pela segurança do paciente e pelo cumprimento dos princípios éticos e de confidencialidade (Conselho Federal De Medicina, 2025).

De forma semelhante, a Organização Mundial da Saúde (OMS) recomenda governança ética, transparência e equidade na implementação dessas tecnologias. Nos Estados Unidos, a FDA regula sistemas de IA como dispositivos médicos baseados em software, exigindo validação clínica e monitoramento contínuo, enquanto a União Europeia classifica aplicações médicas como sistemas de alto risco no âmbito do AI Act (Lei de Inteligência Artificial da União Europeia) é a primeira legislação abrangente do mundo dedicada à regulamentação da inteligência artificial, regulamentada em 2024,

impondo requisitos rigorosos de segurança e supervisão humana (OMS, 2021; FDA, 2021; União Europeia, 2024).

2.2 TEMPO DE PERMANÊNCIA HOSPITALAR: UM PROCESSO DE TEMPO ATÉ EVENTO

O tempo de permanência hospitalar (*Length of Stay* – LOS) configura-se como uma variável temporal contínua cuja realização depende de múltiplos fatores clínicos, demográficos e organizacionais. Sob a perspectiva estatística, o LOS pode ser formalmente modelado como uma variável aleatória não negativa T , representando o tempo transcorrido entre a admissão hospitalar e a ocorrência do evento de interesse, neste caso, a alta médica (Collett, 2023).

Essa formulação insere o problema no arcabouço da análise de sobrevivência, cujo objetivo central consiste na modelagem da distribuição temporal de eventos. Diferentemente de abordagens tradicionais baseadas em regressão linear ou logística, a modelagem de tempo até evento permite incorporar explicitamente a estrutura temporal do fenômeno e tratar adequadamente observações censuradas. A censura à direita ocorre quando o evento de interesse não é observado até o término do período de acompanhamento, situação frequente em bases hospitalares (Antonou *et al.*, 2018), nas quais pacientes podem permanecer internados no momento da extração dos dados.

A análise de sobrevivência fundamenta-se em três funções principais: a função de sobrevivência $S(t) = P(T > t)$, que descreve a probabilidade de o evento não ter ocorrido até o tempo t ; a função densidade de probabilidade $f(t)$; e a função de risco ou hazard $h(t)$, que representa a taxa instantânea de ocorrência do evento condicionada à sobrevivência até o tempo t . No contexto hospitalar, a função de risco pode ser interpretada como a probabilidade instantânea de alta médica dado que o paciente permanece internado até determinado momento.

Essa estrutura probabilística torna a análise de sobrevivência particularmente adequada para modelar o LOS, pois permite estimar probabilidades condicionais ao tempo já decorrido de internação, característica central para aplicações clínicas e gerenciais.

2.3 ANÁLISE DE SOBREVIVÊNCIA: CONCEITUAÇÃO

A análise de sobrevivência compreende um conjunto de métodos estatísticos para estudar o tempo até a ocorrência de um evento (Colosimo; Giolo, 2006). Diferentemente das abordagens tradicionais de regressão, ela lida com dados censurados, quando o evento de interesse não é observado durante o período de acompanhamento (Kleinbaum; Klein, 2012). Essa característica torna a técnica particularmente útil em contextos nos quais é essencial compreender e prever a duração de processos, como em saúde, engenharia e ciências sociais.

Já Hosmer, Lemeshow e May (2008), afirmam que diferente de outras abordagens preditivas, esse tipo de análise lida com a particularidade de que nem sempre é possível observar o evento para todos os indivíduos, fenômeno denominado censura (*censoring*). Este ocorre, por exemplo, quando um paciente não apresentou o desfecho até o fim do período de estudo.

O ponto central da análise de sobrevivência é modelar a função de sobrevivência ($S(t)$), que representa a probabilidade de que um indivíduo não tenha experimentado o evento até um tempo X , e a função de risco (*hazard function*), que expressa a taxa instantânea de ocorrência do evento. Esta análise oferece ferramentas robustas para compreender processos temporais complexos, incorporando técnicas clássicas e modernas que possibilitam maior precisão na tomada de decisão. Seu caráter interdisciplinar a torna essencial para pesquisadores e profissionais de diversas áreas e torna suas aplicações amplas:

- Saúde: uma aplicação relevante é a previsão do tempo de internação hospitalar (*Length of Stay* – LOS), que permite otimizar recursos, reduzir custos e melhorar o planejamento do fluxo de pacientes (Wang *et al.*, 2020). Modelos clássicos como o modelo de Cox e abordagens modernas como *Random Survival Forest* e redes neurais profundas (*DeepSurv*) têm sido aplicados para esse fim.
- Engenharia: estimativa do tempo de falha de componentes e sistemas (Nelson, 2004).

- Economia e Negócios: análise de Churn, ou tempo até inadimplência e comportamento de clientes (Singh; Pandey, 2021). Entre os conceitos-chave da análise, destacam-se:
 - Evento de interesse: desfecho analisado, como morte, alta hospitalar, falha de equipamento, entre outros.
 - Tempo até o evento: variável de interesse que mede a duração entre o início da observação e a ocorrência do evento.
 - Censura: situações em que o evento não é observado, seja por término do estudo ou perda de acompanhamento.

Os métodos de análise de sobrevivência são aplicados conforme a natureza dos dados e os objetivos do estudo.

- Métodos Não Paramétricos: A *estimativa de Kaplan-Meier* é amplamente utilizada para estimar a função de sobrevivência sem pressupor distribuição paramétrica, permitindo a construção de curvas de sobrevivência empíricas (Kaplan; Meier, 1958). Para comparação de curvas entre grupos, é usual aplicar o *teste de log-rank*;
- Modelos Semiparamétricos: O modelo de riscos proporcionais de Cox (COXPH) (Cox, 1972) é um dos mais difundidos, por relacionar covariáveis ao risco de ocorrência do evento sem assumir uma forma específica para a função de risco basal, oferecendo grande flexibilidade analítica;
- Modelos Paramétricos: Quando a distribuição dos tempos é conhecida ou pode ser assumida, utilizam-se modelos paramétricos, como os baseados em distribuições Exponencial, *Weibull* ou *Log-Normal* (Colosimo; Giolo, 2006);
- Abordagens de Aprendizado de Máquina: O avanço da aprendizagem de máquina ampliou as possibilidades da análise de sobrevivência. Métodos como *Random Survival Forest* e redes neurais profundas (e.g., *DeepSurv* e *DeepHit*) têm sido aplicados para modelar relações complexas entre variáveis e eventos, com destaque para grandes bases de dados clínicos (Lu; Chen; Hao, 2020).

Três métodos amplamente utilizados em análise de sobrevivência, utilizados durante nosso estudo, serão abordados brevemente.

2.3.1 Modelo de Riscos Proporcionais de COX (COXPH)

O Modelo de Riscos Proporcionais de Cox, proposto por David Cox em 1972, é um dos métodos clássicos mais utilizados na análise de sobrevivência. Trata-se de um modelo semiparamétrico que relaciona o tempo até a ocorrência de um evento (e.g., óbito, alta hospitalar, falha de um equipamento) a um conjunto de variáveis explicativas, permitindo estimar o efeito dessas covariáveis sobre o risco (hazard) de ocorrência do evento ao longo do tempo (Cox, 1972). A função de risco no modelo de Cox é expressa como:

$$h(t|X) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)$$

Em que:

- $h(t|X)$ é a função de risco condicional no tempo t , dado o vetor de covariáveis X ;
- $h_0(t)$ representa a função de risco basal (*baseline hazard*), comum a todos os indivíduos;
- β_i são os coeficientes associados às covariáveis X_i .

A principal característica do modelo é o pressuposto de proporcionalidade dos riscos, isto é, a razão de riscos entre dois indivíduos com diferentes valores de covariáveis permanece constante ao longo do tempo. Esse aspecto torna o modelo particularmente útil para estudos em que o objetivo é identificar fatores de risco associados ao tempo de sobrevida ou de ocorrência de eventos, sem a necessidade de especificar a forma funcional da função de risco basal. Entre as principais vantagens do modelo estão: a flexibilidade (não especifica a função de risco basal), a interpretação direta das razões de risco e a capacidade de lidar com censuras. No entanto, violações do pressuposto de proporcionalidade podem comprometer a validade do modelo, sendo recomendada a realização de testes e diagnósticos (Kleinbaum; Klein, 2012).

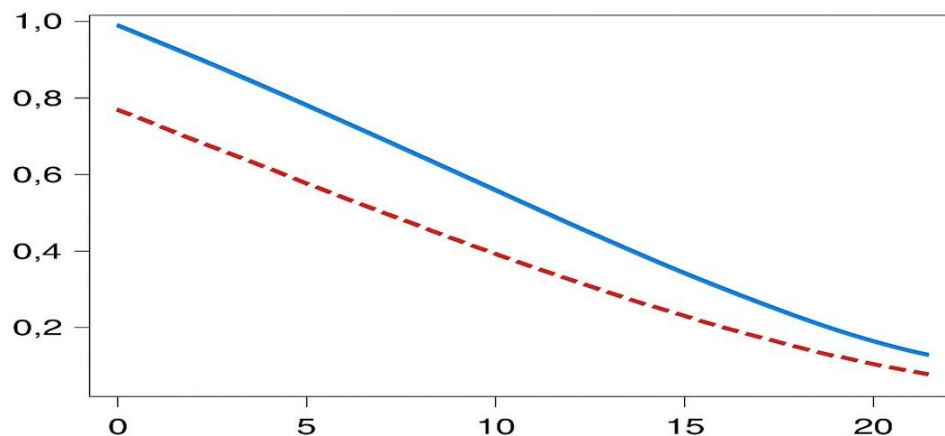
O modelo de Cox é amplamente utilizado na saúde, como na previsão do tempo de internação hospitalar (*Length of Stay – LOS*), fator crucial para o planejamento de recursos hospitalares. Pesquisas recentes têm empregado o modelo para estimar a probabilidade de alta em função de variáveis como idade, comorbidades e tipo de

procedimento (Ataman; Sariyer, 2022). Além disso, sua robustez diante de dados censurados o torna adequado em estudos clínicos, permitindo estimar efeitos mesmo quando o evento de interesse não é observado em todos os indivíduos. A Figura 1 ilustra o modelo de riscos Proporcionais de Cox, mostrando duas curvas de sobrevivência (X_1 e X_2) ao longo do tempo. Onde:

- **Eixo horizontal (x):** representa o tempo até o evento (e.g., óbito, alta hospitalar, falha de um sistema).
- **Eixo vertical (y):** mostra a probabilidade de sobrevivência, que decresce à medida que o tempo avança.
- **Curvas:** as duas linhas (azul e vermelha) representam grupos com diferentes valores de uma covariável. A curva azul (X_1) apresenta maior probabilidade de sobrevivência ao longo do tempo, enquanto a vermelha (X_2) indica maior risco de ocorrência do evento.

O paralelismo entre as curvas reflete o pressuposto de proporcionalidade dos riscos: a razão de riscos entre os grupos é constante ao longo do tempo, como mostra a Figura 1.

Figura 1. Modelo de riscos proporcionais de Cox



Fonte: Cox (1972).

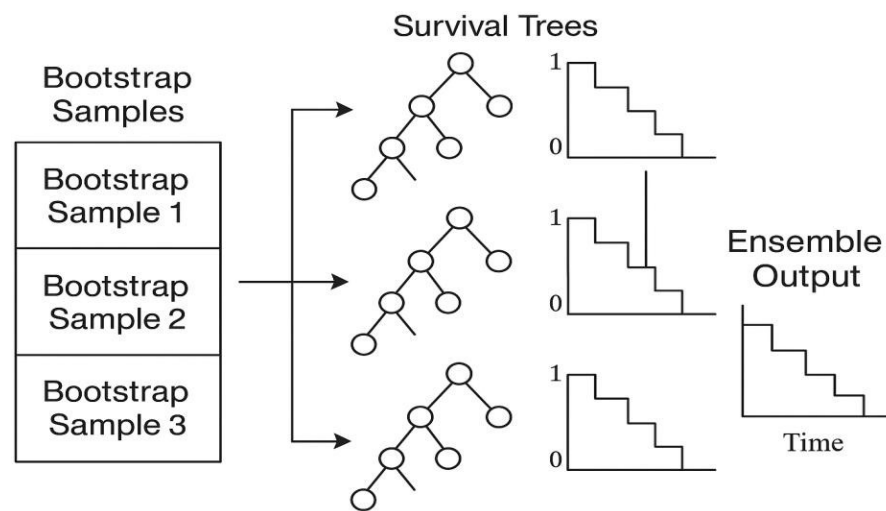
2.3.2 Random Survival Forest

O *Random Survival Forest* (RSF) é uma extensão do algoritmo *Random Forest* desenvolvido especificamente para lidar com dados de análise de sobrevivência, que frequentemente incluem desfechos do tipo tempo até o evento e presença de censura (e.g., tempo de internação hospitalar ou tempo até óbito). Proposto inicialmente por Ishwaran *et al.* (2008), o RSF adapta a estrutura tradicional de árvores de decisão para estimar funções de sobrevivência de forma não paramétrica, utilizando o método de Kaplan-Meier em cada nó terminal e o índice de log-rank como medida de pureza dos nós.

O funcionamento do RSF baseia-se na criação de múltiplas árvores de decisão de sobrevivência, onde cada árvore é construída a partir de uma amostra bootstrap dos dados e com divisão aleatória das variáveis em cada nó, promovendo diversidade entre as árvores. A predição final é obtida agregando os resultados das árvores individuais, geralmente por meio da média das funções de sobrevivência estimadas. Essa abordagem apresenta diversas vantagens em relação aos modelos tradicionais, como o *Cox Proportional Hazards*, pois não depende de suposições paramétricas (e.g., a proporcionalidade dos riscos), é robusta a dados de alta dimensionalidade e lida bem com interações complexas entre variáveis. Além disso, o RSF fornece medidas de importância das variáveis (*Variable Importance*), auxiliando na interpretação clínica e na seleção de características relevantes para o modelo.

Na área da saúde, o RSF tem sido amplamente aplicado para previsão do tempo de internação, sobrevida de pacientes com doenças crônicas e estimativas prognósticas em cenários complexos, mostrando-se particularmente útil em bases de dados grandes e heterogêneas. A Figura 2 abaixo mostra um exemplo de uma estrutura de RSF.

Figura 2. Estrutura de um Random Survival Forest



Fonte: Ishwaran *et al.* (2008).

2.3.3 DeepSurv

O DeepSurv é um modelo de aprendizado profundo para análise de sobrevivência, desenvolvido por Katzman *et al.* (2018), que combina a estrutura do modelo de Cox *Proportional Hazards* (CoxPH) com a flexibilidade de redes neurais artificiais. Seu principal objetivo é modelar relações não lineares e complexas entre covariáveis e o tempo até o evento, superando as limitações dos modelos tradicionais, que assumem proporcionalidade dos riscos e linearidade dos efeitos das covariáveis.

Já o modelo de Cox, estima o risco proporcional com base em uma função linear dos preditores. O *DeepSurv* substitui essa função linear por uma rede neural profunda, capaz de aprender representações não lineares a partir dos dados. Essa rede é treinada para maximizar a verossimilhança parcial de Cox, garantindo que a saída da rede mantenha a interpretação probabilística associada ao risco relativo do evento.

Este método tem sido amplamente utilizado em pesquisas médicas para modelagem de tempo-até-evento, incluindo aplicações na previsão do tempo de internação hospitalar, visando apoiar a gestão de recursos e a tomada de decisão clínica e outras aplicações que envolvem dados censurados (Wen *et al.*, 2022). Sua principal vantagem é lidar com grandes bases de dados heterogêneas e capturar interações

complexas entre variáveis clínicas, o que o torna promissor para sistemas de suporte à decisão clínica e personalização de tratamentos. Além disso, por ser um modelo de aprendizado supervisionado, pode ser continuamente aprimorado com novos dados, tornando-se mais robusto ao longo do tempo.

2.4 MODELAGEM DINÂMICA DO RISCO E ATUALIZAÇÃO TEMPORAL

Na prática clínica, o estado de um paciente hospitalizado evolui continuamente. Resultados laboratoriais, procedimentos realizados, resposta terapêutica e intercorrências modificam progressivamente o risco de alta ou de desfechos alternativos. Assim, a hipótese de risco fixo ao longo do tempo pode não refletir adequadamente a realidade assistencial.

A modelagem dinâmica em análise de sobrevivência busca incorporar essa característica ao permitir a atualização das estimativas de risco condicionadas tanto ao tempo já transcorrido quanto às novas informações disponíveis. Formalmente, isso implica estimar probabilidades do tipo:

$$P(T > t + \Delta t \mid T > t, \mathcal{X}(t)),$$

onde $\mathcal{X}(t)$ representa o conjunto de covariáveis observadas até o tempo t .

Essa perspectiva desloca o foco da previsão pontual para uma estrutura de previsão condicional e sequencial, mais alinhada ao processo decisório clínico. Do ponto de vista gerencial, tal abordagem pode produzir estimativas progressivamente refinadas do tempo remanescente de internação, contribuindo para planejamento de altas, organização de leitos e alocação de recursos.

2.5 PIPELINE CLÁSSICO PARA CONSTRUÇÃO DE MODELOS PREDITIVOS

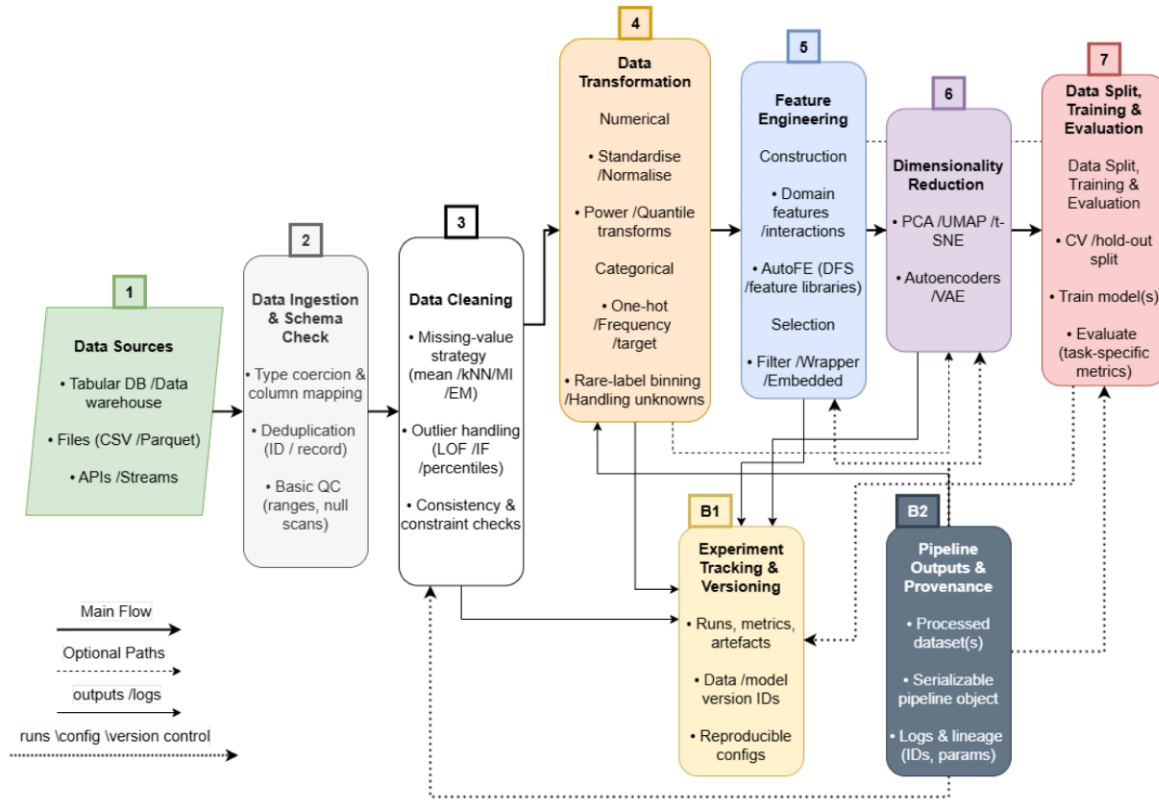
2.5.1 Visão Geral

A análise de sobrevivência envolve o uso de modelos com o objetivo de compreender a estrutura dos dados existentes e estabelecer regras que permitam realizar previsões probabilísticas. Na área da saúde, por exemplo, modelos preditivos podem ser utilizados para estimar o risco de determinado desfecho ocorrer (Santos *et al.*, 2019). A construção destes modelos geralmente segue um pipeline clássico (Figura 3), em que os dados brutos são coletados, validados e refinados por meio da limpeza, transformação e redução de dados, antes da divisão e em seguida treinamento dos mesmos.

Um projeto de modelo de predição consiste em pré-processamento, processamento e, por fim, um pós-processamento dos dados. O pré-processamento de dados constitui uma etapa fundamental no desenvolvimento de modelos de inteligência artificial (IA), funcionando como elo entre os dados brutos e a geração de informações acionáveis. Esse processo abrange um conjunto de técnicas voltadas à transformação e à limpeza dos dados, com o objetivo de elevar sua qualidade e usabilidade.

Diante do crescente volume e complexidade dos dados oriundos de diferentes domínios, um pré-processamento eficaz torna-se indispensável para a construção de sistemas de IA robustos, capazes de fornecer previsões confiáveis e insights precisos (Peace; Chris; Victor, 2024; Gorgens *et al.*, 2009).

Figura 3. Exemplo de roteiro de criação de modelos preditivos



Fonte: Koukaras e Tjortjis (2025).

O pipeline de pré-processamento geralmente contempla etapas como coleta, exploração, limpeza, transformação, redução e divisão dos dados. De fato, de acordo com Li (2019), comumente os dados brutos obtidos do mundo natural estão mal estruturados, com problemas, como duplicações, valores discrepantes ou ausentes e até contradições no conjunto de dados. Embora as técnicas atuais de pré-processamento não resolvam todos esses problemas, elas podem mitigar parte deles e contribuir para a melhoria do desempenho dos modelos. As etapas de pré-processamento, que não são necessariamente lineares, mas sim processos iterativos que podem precisar de ajustes e refinamentos, serão comentadas a seguir. As etapas usadas no nosso projeto são detalhadas nas subseções seguintes.

2.5.2 Coleta de dados

Esta etapa coleta informações de diversas fontes, como bases estruturadas, não estruturadas e formatos semiestruturados, utilizando técnicas como *scraping* (método automatizado que coleta a partir de páginas da web, sistemas online ou documentos digitais.), APIs, repositórios públicos, entre outros.

Dados estruturados (tipo de dado utilizado nesta pesquisa) são aqueles organizados em formato fixo, geralmente armazenados em tabelas com linhas e colunas, como em bancos de dados relacionais. Tem como características, uma estrutura rígida, padronizada e fácil inserção, consulta e atualização via linguagens como SQL e possuem tipos de dados definidos (números, datas, *strings* etc.). Exemplos: dados pessoais (nome, CPF, endereço, telefone) e dados financeiros.

Os dados não estruturados são dados sem formato pré-definido, que não seguem um modelo fixo, sendo de difícil armazenamento em bancos relacionais. Por isso, eles têm alta variabilidade e heterogeneidade, são de difícil processamento sem etapas de pré-processamento e mineração e costumam exigir técnicas de processamento de linguagem natural (PLN), visão computacional ou análise multimodal. Temos como exemplos os textos livres (e-mails, posts em redes sociais), imagens, áudios, vídeos e documentos digitais (PDFs).

Já os dados semiestruturados, estão entre os estruturados e os não estruturados, possuem alguma organização, mas não seguem um modelo fixo de tabelas relacionais. Eles apresentam metadados (informações que descrevem os dados), não exigem esquema rígido, mas possuem uma hierarquia ou marcação que facilita o processamento e são muito usados em integração de sistemas e big data. Exemplos: XML (*eXtensible Markup Language*), JSON (*JavaScript Object Notation*), logs de sistemas (arquivos que descrevem eventos) e mensagens de e-mail (texto + metadados como remetente e data).

2.5.3 Exploração de dados

A EDA (*Exploratory Data Analysis*) é uma etapa essencial no pré-processamento de dados, cujo objetivo é entender a estrutura, características e padrões presentes nos

dados antes de aplicar algoritmos de modelagem (Guedes; Silva, 2022). As principais funções são:

1. Compreensão da estrutura do conjunto de dados:
 - Identificar tipos de variáveis (categóricas, numéricas, temporais) e detectar relações entre variáveis.
2. Detecção de problemas:
 - Detectar valores faltantes (*missing values*), Outliers (valores discrepantes) e dados inconsistentes ou ruidosos.
3. Visualização dos dados:
 - Formação de histogramas, boxplots, gráficos de dispersão e heatmaps e verificação de correlações e padrões de distribuição.
4. Preparação para modelagem:
 - Apoiar decisões sobre limpeza, normalização e seleção de atributos.

Um exemplo é ao analisar dados de pacientes em um hospital, a EDA pode mostrar faixas etárias com maior ocorrência de internações, relação entre doenças pré-existentes e tempo de internação e presença de valores ausentes em exames laboratoriais.

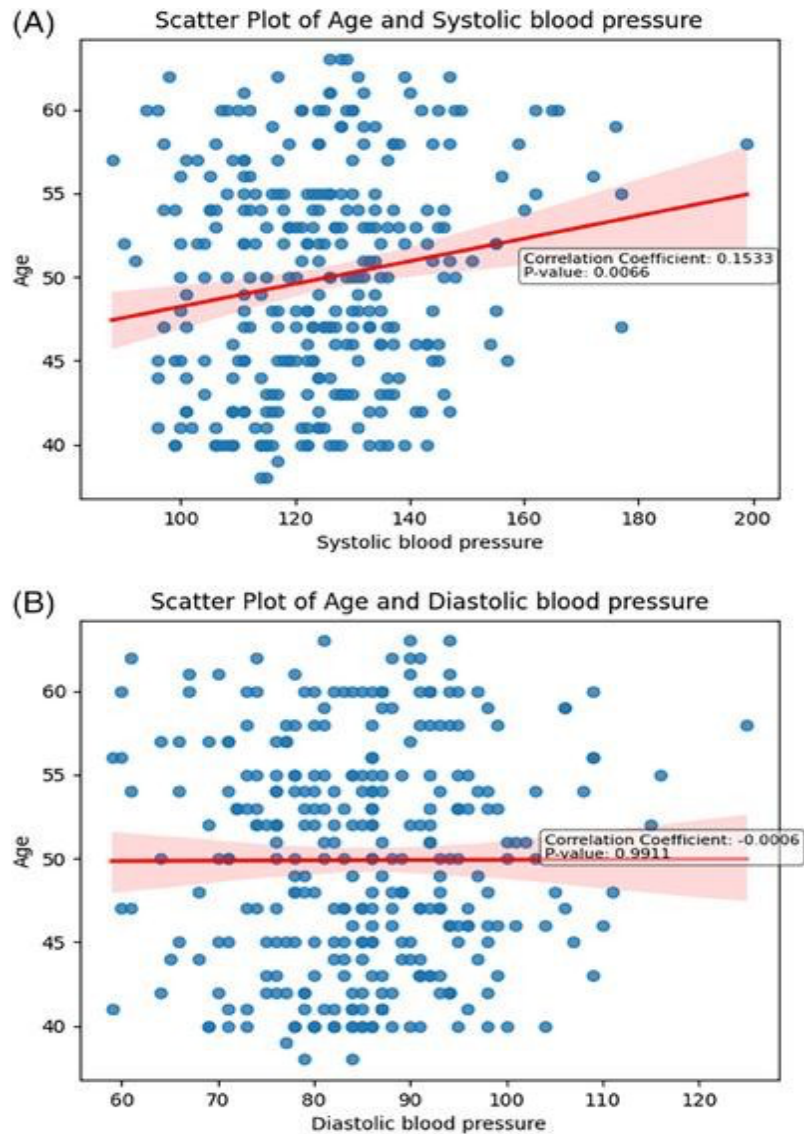
2.5.4 Limpeza de dados

A limpeza de dados consiste em corrigir, ajustar ou excluir informações incorretas, incompletas ou irrelevantes, visando garantir que o conjunto seja confiável e adequado para a modelagem. Nesta etapa, acontece o tratamento de valores ausentes, imputação (média, mediana, algoritmos) ou exclusão de registros incompletos; a remoção de duplicatas, a eliminação registros repetidos que distorcem as análises; a correção de inconsistências, padronização de formatos (e.g., datas, textos) e correção de erros de digitação; e a detecção e tratamento de *outliers* (identificar valores discrepantes e decidir se devem ser ajustados ou removidos).

Em um banco de dados hospitalar, substituir valores faltantes de exames por estimativas baseadas em pacientes com características semelhantes ou em planilhas de clientes, excluir registros duplicados ou ajustar nomes escritos de formas diferentes, são

exemplos de limpeza de dados. A Figura 4 mostra gráficos de dispersão que mostram a correlação entre a leitura da pressão arterial sistólica (A) e da pressão arterial diastólica (B) com a idade dos participantes de uma pesquisa.

Figura 4. Gráfico de dispersão de dados com outliers



Fonte: Alhawari *et al.* (2024).

2.5.5 Transformação de dados

A transformação de dados envolve alterar a forma e a representação dos dados para que fiquem adequados ao uso por algoritmos de aprendizado de máquina. Essa etapa é fundamental porque muitos modelos têm requisitos específicos quanto à escala, distribuição e formato dos dados (Li, 2019). As principais técnicas são:

1. Normalização:
 - Reduz valores para uma faixa comum (ex.: 0 a 1).
 - Útil para algoritmos sensíveis à escala, como redes neurais e K-means.
2. Padronização (*Standardization*):
 - Transforma os dados para terem média 0 e desvio-padrão 1.
 - Importante para métodos baseados em distância (SVM, regressão logística).
3. Codificação de variáveis categóricas:
 - *One-hot encoding*: cria variáveis binárias para categorias.
 - *Label encoding*: converte categorias em números inteiros.
4. Engenharia de atributos (*Feature engineering*):
 - Criação de novas variáveis derivadas dos dados originais (ex.: idade a partir da data de nascimento).
5. Redução de dimensionalidade:
 - Técnicas como PCA (*Principal Component Analysis*) para simplificar o espaço de atributos, reduzindo ruído e melhorando desempenho.

3.5.6 Redução de dados

Consiste em diminuir a quantidade de variáveis (atributos) ou registros (instâncias) em um conjunto de dados, preservando ao máximo as informações essenciais. Essa etapa é aplicada para melhorar a eficiência, reduzir o custo computacional e evitar sobreajuste (*overfitting*) em modelos de aprendizado de máquina. Por exemplo, em um dataset com milhares de variáveis genéticas, a PCA pode reduzir o número de atributos mantendo a maior parte da variabilidade. As principais técnicas são:

1. Redução de dimensionalidade:

- PCA (*Principal Component Analysis*): transforma variáveis correlacionadas em componentes independentes.
- T-SNE e UMAP: usados para visualização e redução não linear de dimensões.
- Análise de componentes independentes (ICA).

2. Seleção de atributos (*Feature selection*):

- Remoção de variáveis irrelevantes ou redundantes com métodos como seleção por importância de modelos (ex.: Random Forest) e métodos estatísticos.

3. Amostragem (*Sampling*):

- Seleção de subconjuntos representativos dos dados originais, mantendo suas características principais.

3.5.7 Divisão de dados

Consiste em separar o conjunto original em subconjuntos com finalidades específicas, geralmente:

- Treino (*Training set*): usado para ajustar os parâmetros do modelo.
- Validação (*Validation set*): utilizado para ajustar hiperparâmetros e evitar sobreajuste (*overfitting*).
- Teste (*Test set*): avalia a capacidade de generalização do modelo em dados nunca vistos (Peace; Chris; Victor, 2024). As principais técnicas são:
 - *Hold-out*: divisão simples (e.g., 70% treino, 15% validação, 15% teste).
 - *K-Fold cross-validation*: o conjunto é dividido em k partes, e o modelo é treinado/testado múltiplas vezes, alternando as partições (Joshi; Patel, 2020).
 - Amostragem estratificada: garante que as proporções das classes (em problemas de classificação) sejam mantidas em cada subconjunto.

O desempenho de modelos de aprendizado de máquina depende de variáveis que controlam seu funcionamento e aprendizado. Essas variáveis podem ser classificadas

em parâmetros e hiperparâmetros, conceitos centrais para compreender como os algoritmos são ajustados e otimizados (Russell; Norvig, 2020).

Os parâmetros são os valores internos do modelo que são aprendidos automaticamente durante o processo de treinamento. Eles ajustam o modelo aos dados e determinam como ele faz previsões. Além disso, são ajustados pelo algoritmo durante o treinamento e definem a função aprendida que mapeia entradas em saídas. Como exemplos, pesos e vieses em redes neurais, coeficientes em regressão linear ou logística e probabilidades em modelos *Naïve Bayes* (Russell; Norvig, 2020).

Os hiperparâmetros, por outro lado, não são aprendidos pelo modelo. Eles são definidos pelo desenvolvedor antes do treinamento e controlam o processo de aprendizado. Ademais, afetam a capacidade de generalização do modelo e são ajustados por meio de métodos como *grid search*, *random search* e *Bayesian optimization* (Géron, 2019). Como exemplo, tem-se a taxa de aprendizado (learning rate) em redes neurais, o número de árvores em uma Random Forest, a profundidade máxima de uma árvore de decisão e o número de vizinhos no K-NN.

Enquanto os parâmetros definem o que o modelo aprendeu a partir dos dados, os hiperparâmetros controlam como o modelo aprende. Um ajuste inadequado de hiperparâmetros pode resultar em *underfitting* (subajuste) ou *overfitting* (sobreajuste), comprometendo o desempenho do modelo (Géron, 2019). O ajuste (*tuning*) adequado de hiperparâmetros é essencial para otimizar a performance dos algoritmos. Técnicas como validação cruzada (*k-fold cross-validation*) são amplamente utilizadas para testar combinações de hiperparâmetros e escolher aquelas que maximizem o desempenho do modelo em dados não vistos (Goodfellow; Bengio; Courville, 2016).

3 TRABALHOS RELACIONADOS

A aprendizagem de máquina e a aprendizagem profunda já foram temas em diversos estudos científicos associados à predição do tempo de internação, de modo que a literatura possui algumas revisões sistemáticas neste domínio.

No estudo *“Previsão precoce de curto prazo da duração da internação no pronto-socorro usando processamento de linguagem natural para pacientes ambulatoriais de baixa gravidade”*, Chen *et al.* (2020) avaliaram o uso de processamento de linguagem natural (PLN) aplicado às primeiras anotações clínicas de 12.962 pacientes de baixa gravidade atendidos em um pronto-socorro universitário terciário em Taiwan, entre janeiro e junho de 2017.

Chen *et al.* (2020) utilizaram dados estruturados (como variáveis demográficas e sinais vitais) e diferentes seções das anotações iniciais, os autores desenvolveram seis modelos baseados em PLN (TF-IDF e decomposição de valores singulares truncados – SVD) para prever o tempo de permanência hospitalar (TP), avaliando o desempenho com o erro quadrático médio (MSE). Os resultados indicaram que pacientes de baixa gravidade representam a maior parcela do atendimento e apresentaram TP frequentemente imprevisível, e que os modelos baseados em PLN, ao permitir extração automática de recursos e capturar interações não lineares, proporcionaram previsões mais precisas do LOS. O melhor desempenho foi obtido com o modelo que combinou dados estruturados e todas as seções das notas processadas pelo TF-IDF e SVD truncado, evidenciando o potencial do PLN como ferramenta para auxiliar médicos a estimar o TP precocemente, gerenciar o fluxo de pacientes e reduzir a superlotação, embora pesquisas nesse grupo ainda sejam escassas.

Peng e Gao (2025) desenvolveram um estudo com o objetivo de aprimorar a previsão do tempo de internação hospitalar (*Length of Stay – LOS*) utilizando métodos de aprendizado de máquina baseados em *ensembles*. O trabalho resultante foi intitulado *“Prediction of hospital length of stay: leveraging ensemble tree models and intelligent feature selection”* (Previsão do tempo de internação hospitalar: aproveitando modelos de árvores em ensemble e seleção inteligente de variáveis). A pesquisa foi conduzida a partir

de um banco de dados contendo aproximadamente 100.000 registros de internações hospitalares, abrangendo 26 variáveis clínicas e demográficas.

Os dados foram coletados em múltiplas unidades de um sistema hospitalar localizado nos Estados Unidos, garantindo representatividade para diferentes perfis de pacientes. Para a modelagem, os autores empregaram algoritmos baseados em árvores de decisão, incluindo *Random Forest*, *Gradient Boosting Machine* (GBM) e *Extreme Gradient Boosting* (XGBoost), além de um modelo de rede neural convolucional (CNN).

Com o intuito de reduzir a complexidade e o tempo de processamento, foi aplicada a técnica de seleção de variáveis pela *Mutual Information*, que permitiu identificar os preditores mais relevantes sem comprometer o desempenho preditivo. O desempenho dos modelos foi avaliado por meio de métricas amplamente utilizadas na literatura, como o erro quadrático médio (RMSE) e o erro absoluto médio (MAE). Entre as técnicas testadas, o modelo *XGBoost* obteve os melhores resultados, alcançando valores de RMSE aproximados de 0,3881 e MAE em torno de 0,1553, evidenciando sua superioridade em termos de precisão. Além disso, a aplicação da seleção de variáveis por MI demonstrou eficácia na redução da complexidade computacional, mantendo níveis elevados de acurácia.

No estudo “*Modelos com reconhecimento do modo de chegada para previsão do fluxo de pacientes e tempo de permanência em departamentos de emergência*”, Atman e Sariyer (2022) investigaram a previsão do fluxo de pacientes e do tempo de permanência (LoS) em um hospital terciário em İzmir, Turquia. Foi considerado o modo de chegada (ambulância ou meios próprios) em 108.708 atendimentos entre maio de 2016 e maio de 2017.

Foram construídas quatro séries temporais representando o número diário de chegadas por cada modo e o tempo médio de permanência, analisadas com modelos ARIMA e SARIMA, conforme a sazonalidade das séries. A avaliação por MAPE mostrou boa precisão (5,43% para chegadas por ambulância, 13,09% para auto-arribos, 9,96% e 10,98% para os tempos médios de permanência). Os resultados indicaram que SARIMA captou melhor o fluxo diário, enquanto ARIMA simples foi adequado para o tempo médio de permanência. A distinção por modo de chegada proporcionou maior precisão e

forneceu subsídios estratégicos para gestão operacional e planejamento eficiente de recursos nos serviços de emergência.

O estudo *“Optimization of bed allocation based on survival analysis, treatment trajectory, and cost estimates”* de Karbouband e Tabaa (2023) propôs um modelo para otimizar a alocação de leitos em UTIs, integrando aspectos clínicos, preditivos e econômicos. A pesquisa utilizou dados de pacientes com doenças cardiovasculares internados em UTIs de hospitais em Casablanca, Marrocos, com colaboração acadêmica da França, abrangendo múltiplas classes de pacientes e variáveis clínicas relevantes.

O método combina análise de sobrevivência (estimador de Kaplan–Meier), modelagem da trajetória terapêutica (regressão logística e T-learning) e um algoritmo genético multiobjetivo que considera sobrevivência, tratamento e estimativas de custo. O desempenho foi avaliado por RMSE e MAPE, mostrando boa precisão (RMSE entre 2,055 e 2,890 leitos/dia; MAPE entre 9,4% e 13,2%) e superando métodos tradicionais, como médias móveis e modelos univariados. Os autores destacaram que a abordagem multiobjetivo ofereceu flexibilidade e suporte à decisão em contextos de alta demanda, equilibrando eficiência operacional, qualidade do cuidado e sustentabilidade econômica na gestão de leitos críticos.

O quadro 1 a seguir apresenta os principais atributos dos estudos analisados, com resumo dos datasets, técnicas, métricas e resultados.

Quadro 1. Tabela comparativa de trabalhos realizados com o mesmo tema

Trabalho	Título	Dataset	Tamanho do Dataset	Local	Técnicas	Métricas de Avaliação	Melhor Resultado
Chen et al. (2020)	Previsão precoce de curto prazo da duração da internação no pronto-socorro usando PNL	Prontuários de pacientes de baixa acuidade	12.962 pacientes	Hospital universitário terciário em Taiwan	6 modelos de predição baseados em PLN, TF-IDF, SVD truncado	MSE	Modelo com dados estruturados + todas as seções SOAP processadas por TF-IDF e SVD
Peng (2025)	Previsão do tempo de	Dataset público e	≈100.000	Múltiplas unidades hospitalar	Random Forest, GBM,	RMSE, MAE	XGBoost com MI (alta

	internação hospitalar com ensembles e seleção de variáveis	abrangente	registros	es nos EUA	XGBoost, CNN, seleção por MI		acurácia e menor tempo de processamento)
Atman e Sariyer (2022)	Modelos com reconhecimento do modo de chegada para previsão de fluxo e LOS	Registros reais de pronto-socorro urbano	108.708 registros	Hospital de grande porte em Izmir, Turquia	Modelos ARIMA e SARIMA (fluxo diário e LOS por modo de chegada)	MAPE	SARIMA para fluxo diário e ARIMA não sazonal para-LOS
Karboube Tabaa (2023)	Otimização da alocação de leitos com análise de sobrevivência e custos	Pacientes com doenças cardiovasculares em UTIs	Não informado	Casablanca (Marrocos) e Lorraine (França)	Kaplan–Meier, regressão logística, T-learning, algoritmo genético multiobjetivo	RMSE, MAPE	Algoritmo genético com integração de sobrevivência, trajetória e custos

Fonte: Elaborado pela autora (2025).

4 METODOLOGIA

4.1 CONJUNTO DE DADOS

O conjunto de dados utilizado nesta pesquisa foi o banco de dados da Secretaria Municipal de Saúde do município de João Pessoa/PB, que possui dados dos anos de 2018 a 2025, dos prontuários eletrônicos unificados dos pacientes que utilizaram algum dos hospitais públicos do município, sendo eles, a maternidade Cândida Vargas, Hospital Municipal do Valentina, Hospital Ortotrauma de Mangabeira, Hospital Municipal Santa Isabel e Hospital Municipal Prontovida, contendo mais de um milhão de registros.

Estes prontuários informam os dados pessoais dos pacientes, incluindo seus dados e histórico de saúde, como: número e tipo da autorização hospitalar; especialidade do leito; tipo de unidade de cuidado intensivo; idade, sexo e raça do paciente; identificador de gestação de risco; identificação do procedimento de saúde solicitado; complexidade do caso; caráter e duração da internação; e indicadores de internação prolongada e de morte.

4.2 MÉTODOS DE ANÁLISE DE SOBREVIVÊNCIA CLÁSSICOS

Foram utilizados três métodos relevantes:

1. Modelo de Riscos Proporcionalis de Cox (CoxPH);
2. Random Survival Forest (RSF);
3. DeepSurv;

O quadro 2 mostra um quadro comparativo entre os métodos escolhidos:

Quadro 2. Quadro comparativo de métodos

Modelo	Tipo	Suposições	Vantagens	Limitações
CoxPH	Semi-paramétrico	Proporcionalidade dos riscos, relação linear	Interpretação simples	Não captura não linearidades complexas

Modelo	Tipo	Suposições	Vantagens	Limitações
RSF	Não paramétrico, baseado em ensemble de árvores	Constrói múltiplas árvores de sobrevivência e combina seus resultados	Captura interações e não linearidades, robusto a alta dimensionalidade	Menor interpretabilidade em relação ao CoxPH
DeepSurv	Deep Learning + Cox	Estrutura de Cox similar e adaptada	Modelar relações não lineares complexas	Necessita de muitos dados e alta capacidade computacional

Fonte: Elaborado pela autora (2025).

A seleção dos métodos utilizados neste estudo foi fundamentada em trabalho anterior da mesma autora, intitulado “*Predição do tempo de permanência hospitalar utilizando análise de sobrevivência*”, que realizou um comparativo entre quatro abordagens de análise de sobrevivência (Cox, RSF, DeepSurv e DeepHit) para previsão do tempo de permanência hospitalar. O estudo avaliou requisitos de pré-processamento, complexidade dos modelos e desempenho preditivo, apontando o DeepSurv como o de melhor desempenho e o DeepHit como o de pior desempenho. Com base nesses resultados, esta pesquisa optou por incluir apenas os métodos com maior desempenho e potencial de aplicação prática em bases reais de prontuários eletrônicos, descartando abordagens menos eficazes.

O principal produto deste projeto é o modelo de previsão, utilizado para retornar probabilidades relacionadas ao tempo de internação dos pacientes em hospitais, em que o método utilizado para o desenvolvimento deste modelo vai seguir uma versão clássica de um pipeline de aprendizagem de máquina.

4.3 ATIVIDADES

4.3.1 Setup

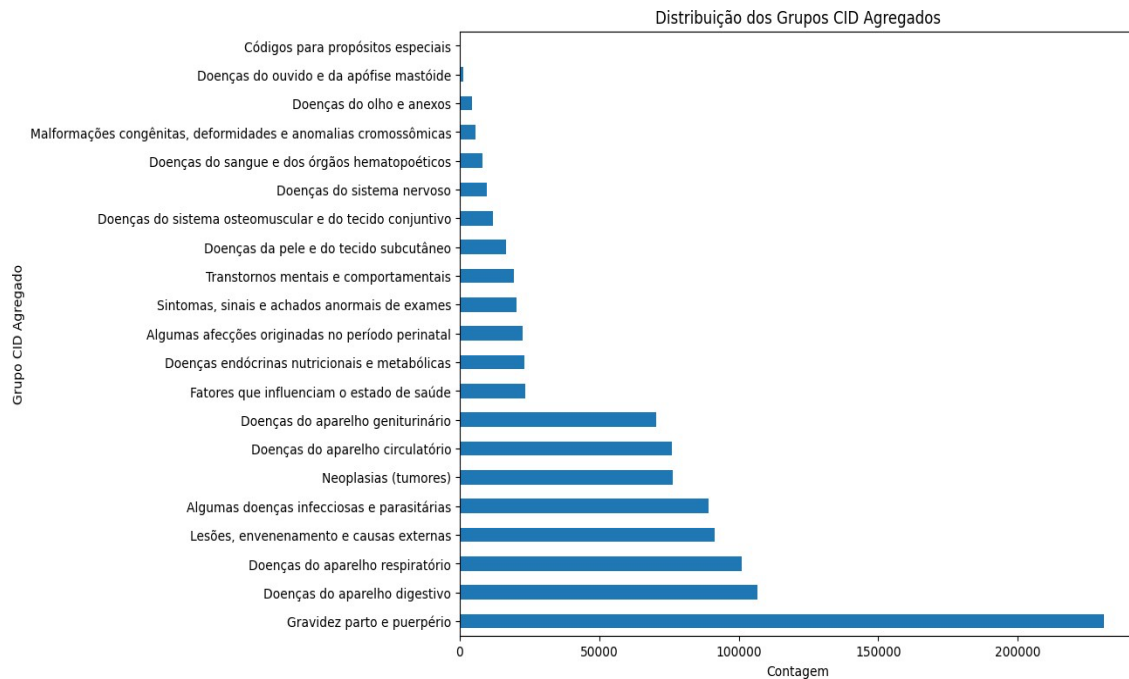
Os dados utilizados foram obtidos das bases de dados dos cinco hospitais públicos municipais de João Pessoa, que são clientes deste sistema. Foi desenvolvido um módulo de coleta que se conectou a uma cópia do banco de dados alvo através de acessos via

VPN fornecido pela Unidade Municipal de Tecnologia da Informação, onde o banco se encontra armazenado, de modo a não interferir no processamento 24 horas dos sistemas atuais. Os dados dos pacientes foram anonimizados, por se tratarem de dados pessoais, de saúde e portanto, dito sensíveis, e cada pesquisador assinou um termo de confidencialidade, cumprindo-se assim a Lei Geral de Proteção de Dados n.º 13709/2018. Cada paciente foi identificado por uma chave aleatoriamente construída no momento da coleta.

Neste banco de dados, um paciente pode permanecer internado por tempo prolongado. Nesse caso, a internação é fragmentada em registros de internação diferentes, mas referentes a mesma estadia. Tendo isso em conta, os experimentos envolveram o agrupamento das internações prolongadas de forma que duas ou mais internações foram representadas como uma só, com o tempo de saída igual ao tempo de saída do último registro.

Os testes foram divididos em 2 fases. A primeira fase envolveu a utilização dos dados de todos os hospitais presentes na base disponibilizada pela prefeitura. Entretanto, observou-se uma grande concentração de registros para o grupo CID (Classificação Internacional de Doenças) equivalente a “gravidez, parto e puerpério”, como mostrado a seguir na Figura 5:

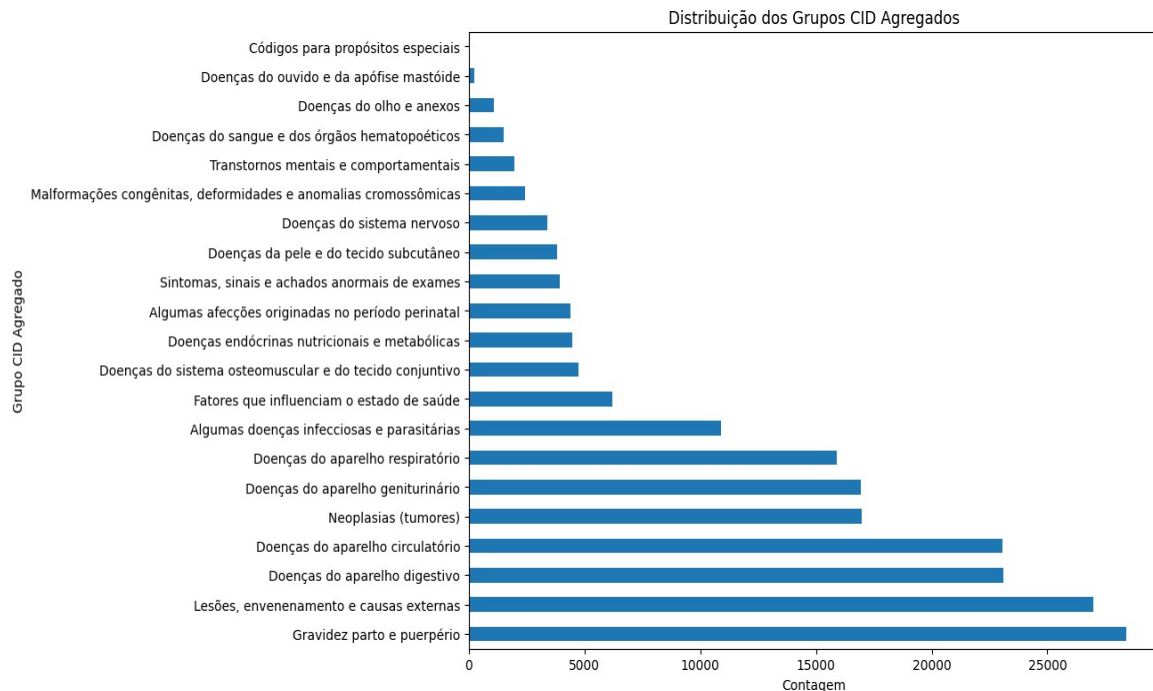
Figura 5. Distribuição dos grupos CID na fase 1



Fonte: Elaborado pela autora com base nos dados da pesquisa (2025).

A fase 2, portanto, considerou apenas os registros dos chamados “hospitais gerais”, para que as internações incluídas estivessem melhor distribuídas entre os diferentes códigos CID. Também foi percebido que o período pandêmico pode ter influenciado a obtenção dos dados, então resolveu-se retirar os dados de internações ocorridas entre os anos de 2020 e 2022. Depois dessa filtragem, a distribuição de registros ficou como segue:

Figura 6. Distribuição dos grupos CID na fase 2



Fonte: Elaborado pela autora com base nos dados da pesquisa (2025).

4.3.2 Fase 1 - Experimentos com dados de todos os hospitais

O pré-processamento envolveu diversas etapas de limpeza, transformação e codificação das variáveis. Para a modelagem de sobrevivência, a alta do paciente foi tomada como evento de interesse, e a morte foi considerada como indicativa de censura. Identificou-se que alguns registros compartilhavam o mesmo número de AIH (Autorização de Internação Hospitalar) em razão de internações prolongadas. Para corrigir essa duplicidade, os registros de uma mesma internação foram agrupados, mantendo-se a primeira data de internação e a última data de saída. O tempo de internação foi calculado como a diferença de dias entre a data de saída e a data de internação.

Aplicou-se também o *One-Hot Encoding* nas variáveis *raca_cor* (raça/cor do paciente), *marca_uci* (marcador de UCI), *marca_util* (tipo de UTI utilizada) e *espec* (especialidade do leito). Para as demais variáveis categóricas, utilizou-se mapeamento

binário: *sexo*, *complex* (complexidade) e *car_int* (caráter de internação), excluindo-se os registros referentes a acidente de trabalho.

Foram produzidas duas versões da base. Na versão 1 (*df_v1*), os 6.788 códigos CID únicos foram agrupados em 22 categorias principais, sobre as quais se aplicou One-Hot Encoding. Na versão 2 (*df_v2*), optou-se por preservar a granularidade original dos 6.788 códigos, aplicando-se *Binary Encoding* para redução dimensional, o que resultou em 13 colunas binárias. O atributo *proc_solic* (procedimento solicitado) foi simplificado pela retenção apenas do primeiro dígito do código, correspondente ao grupo do procedimento, o que resultou em 4 categorias distintas.

Após todas as etapas de pré-processamento, a base *df_v1* ficou composta por 1.009.008 registros e 69 variáveis, enquanto a base *df_v2* contém 1.009.008 registros distribuídos em 61 variáveis. Devido a alta correlação entre algumas variáveis da base *df_v1*, foi necessário retirar as colunas representativas de *espec* e outros atributos relacionados a *uci* e ao grupo de procedimentos solicitados. Já para a base *df_v2*, foram retiradas algumas das colunas representativas de *uci*, *espec*, indicador de gravidez de risco e grupo de procedimento solicitado. Para a utilização do método CoxPH, definiu-se um fator de penalização igual a 0.0001. Nenhum outro parâmetro foi definido.

Para avaliação do Random Forest, foram testados os seguintes números de estimadores: 10, 25, 50, 75, 100, 200. O melhor resultado com a base *df_v1* foi obtido com 10 estimadores. Já para a base *df_v2*, a melhor performance foi obtida com 100 estimadores. Em ambos os casos, foram definidos o número mínimo de amostras por split como 10 e o número mínimo de amostras por folha igual a 1000. A avaliação do Random Forest considerou o C-index censurado.

Para o *DeepSurv*, foram consideradas as seguintes configurações com duas camadas: [32, 32], [64, 32], [64, 64], [128, 64], [128, 128]. A arquitetura que obteve melhor desempenho foi [128, 64] para ambas as bases: *df_v1* e *df_v2*. Nesta fase, não foi utilizado batch. Outros parâmetros são *out_features* = 1, *dropout* = 0.1, *output_bias* = False, taxa de aprendizado igual a 0.01. Os resultados gerados desta fase e das outras serão mostrados na seção Resultados deste trabalho.

4.3.3 Fase 2 - Utilizando dados dos hospitais gerais

Posteriormente, foi observado que o período pandêmico pode ter influenciado a captura de dados. Além disso, percebeu-se um grande número de internações com CID referente ao grupo “gravidez, parto e puerpério”. Assim, foi decidido realizar uma nova fase de experimentos com dados de hospitais gerais e excluindo as internações ocorridas entre os anos de 2020 e 2022.

O problema da alta correlação e da baixa variabilidade de dados em alguns atributos ocorreu. Para a base *df_v1*, então, foi necessária a retirada de algumas colunas representativas de *uti*, *uci*, *espec*, indicador de gestação de risco, e grupos CID. Para a base *df_v2*, algumas colunas relacionadas a *uti*, *uci*, *espec*, indicador de gestação de risco e grupo de procedimento solicitado foram eliminadas.

Foi avaliado como os atributos influenciaram a predição feita pelos modelos. Ao estudar os coeficientes (log HR) do modelo CoxPH, percebeu-se que as variáveis relacionadas aos menores tempos de internação foram os grupos de procedimentos 2 (com finalidade diagnóstica) e 5 (transplantes), *uti* 0 (com leito sem especialidade) e *uci* (unidade de cuidado intensivo) 0 (sem leito específico). Entre os fatores que retardam a alta estão algumas UTIs (unidade de terapia intensiva) adulto, pediátrica, neonatal e cardiológica, como a *uti* 72, *uti* 79, *uti* 82, *uti* 86 e o caráter da internação. Para o modelo *Random Survival Forest* foi aplicado o método de *Permutation Importance* com 10 repetições e entre as variáveis importantes estão *uti* 0, grupo de procedimento 4 e caráter da internação.

Para o modelo DeepSurv, foi usada a técnica de explicabilidade SHAP com background de tamanho igual a 100 e 1000 amostras de teste. As variáveis com maior importância foram grupo de procedimento 4 (cirúrgicos), *uti* 0, caráter da internação e idade.

4.3.4 Modelagem Dinâmica

Em vez de considerar apenas um coorte (em inglês, *cohort*) dos dados, como foi feito até o momento, a modelagem dinâmica considera a atualização das curvas de

sobrevivência à medida que os pacientes vão tendo os seus dados atualizados. Desta forma, a pesquisa considera o “dia” como unidade temporal. Ou seja, o tempo de permanência foi modelado como um desfecho de tempo até o evento em dias, com os pacientes acompanhados desde a admissão hospitalar até a alta ou censura (15º dia). Considere o conjunto de dados a seguir como exemplo ilustrativo (Tabela 1).

Tabela 1. Exemplo de dataset usado durante a modelagem dinâmica

Id	x_1	x_2	x_3	x_n	Assessment time	Duration	Event
P1	20	200	17	33	1	4	1
P2	17	180	18	30	1	9	1
P2	nan	180	nan	nan	3	9	1
P2	20	nan	20	33	6	9	1
P3	17	nan	nan	30	2	5	1
P3	18	nan	nan	nan	4	5	1
P4	16	190	nan	28	2	20	0

Fonte: Elaborado pela autora com base nos dados da pesquisa (2025).

Esta tabela ilustra um exemplo de estrutura de dados longitudinais, na qual n características clínicas ($x_1, \dots, x_n$) são monitoradas ao longo do tempo. A variável *assessment time* indica o dia de internação em que as características presentes na mesma linha foram observadas. É importante destacar que nem todas as variáveis são mensuradas diariamente. Por exemplo, o paciente P1 possui todas as características registradas apenas no primeiro dia de hospitalização, enquanto o paciente P2 apresenta observações distribuídas em três dias distintos (1, 3 e 6).

Salienta-se que, como as medições clínicas são realizadas de acordo com rotinas assistenciais, e não segundo um cronograma fixo de amostragem, as observações das variáveis são inerentemente irregulares ao longo do tempo. Consequentemente, as entradas ausentes (representadas como valores *nan*) indicam que determinada característica não foi avaliada em um dia específico, e não que houve corrupção ou erro no registro dos dados. Essa representação permite que o modelo diferencie valores observados de não observados, preservando a ordenação temporal das medições.

O atributo *Duration* representa o desfecho de tempo até o evento, definido como o número de dias desde a admissão até a alta hospitalar ou censura. Para uma determinada internação, esse valor permanece constante em todas as linhas

longitudinais associadas ao paciente, independentemente do número de observações registradas. Neste exemplo, considera-se um acompanhamento máximo de 15 dias (valor configurável). Pacientes que permanecem internados além desse período são tratados como censurados à direita, e o respectivo indicador de evento é definido como zero.

Nesse sentido, essa organização dos dados permite que modelos de sobrevivência explorem conjuntamente covariáveis longitudinais irregulares, censura à direita e desfechos de tempo até evento, refletindo a estrutura típica de dados hospitalares do mundo real. É importante enfatizar que os valores ausentes não são imputados. Em vez disso, são codificados diretamente na função de escalonamento. Todos os valores observados são reescalados para o intervalo $[1, 6]$, enquanto as entradas ausentes são mapeadas para 0, garantindo uma separação numérica clara entre dados observados e não observados.

Por fim, foi visto até o momento que modelos tradicionais, como o de Riscos Proporcionais de Cox, assumem covariáveis estáticas e um preditor linear estimado a partir de informações observadas em um único ponto temporal, geralmente na admissão. Essa suposição limita sua capacidade de capturar dependências temporais não lineares, variações ao longo do tempo e padrões longitudinais presentes em dados clínicos reais.

Nesta nossa nova perspectiva, a previsão do tempo de permanência hospitalar constitui um problema intrinsecamente dinâmico de sobrevivência, no qual o risco de alta se modifica conforme a evolução do estado clínico do paciente ao longo da internação. Ou seja, ao adotar uma abordagem de sobrevivência dinâmica, torna-se necessário utilizar um modelo capaz de processar explicitamente dados sequenciais ou longitudinais, nos quais múltiplas medições são realizadas em diferentes momentos da internação. Redes Neurais Recorrentes (RNNs) são particularmente adequadas para modelar esse tipo de estrutura temporal, pois conseguem aprender dependências sequenciais e representar trajetórias clínicas. Contudo, essas redes não incorporam, de forma nativa, mecanismos para lidar com censura ou desfechos do tipo tempo até evento, fundamentais na análise de sobrevivência.

A solução explorou, então, uma arquitetura híbrida chamada de GRU-Cox, o qual foi adaptada (configuração estrutural e de hiperparâmetros) para servir como nossa proposta metodológica. Ao integrar uma variante de RNN, a *Gated Recurrent Unit* (GRU),

ao arcabouço do modelo de Cox, essa abordagem permitiu codificar trajetórias clínicas em representações latentes dependentes do tempo e, simultaneamente, estimar uma função de risco compatível com censura à direita e inferência em sobrevivência. Essa formulação possibilita a leitura de dados longitudinais com amostragem irregular e observações ausentes, características inerentes a dados hospitalares do mundo real, alinhando-se à proposta de modelagem dinâmica adotada neste trabalho. Detalhes da construção desta solução foram submetidos ao jornal científico “*Artificial Intelligence in Medicine*” (Qualis A1) com o título “*Rethinking Length-of-Stay Prediction: The Role of Irregular Temporal Dynamics*” e se encontra atualmente em fase de revisão.

5 RESULTADOS

5.1 MODELOS CLÁSSICOS

Como já citado anteriormente, o trabalho *Uma avaliação comparativa de métodos de análise de sobrevivência para previsão do tempo de internação hospitalar* (Crispim et. al., 2025) foi utilizado como base para o desenvolvimento deste estudo, já que realizou um estudo comparativo de quatro métodos bastante utilizados em análise de sobrevivência.

O conjunto de dados utilizado foi o dataset “*Predicting Hospital Length of Stay*” da *Microsoft Machine Learning Services*, que usa um conjunto de dados sintéticos, imitando dados realistas de hospitais, estruturados de forma semelhante aos prontuários eletrônicos de saúde. Esse conjunto inclui 100 mil pacientes e 28 variáveis. Os métodos de análise de sobrevivência utilizados foram: o modelo de riscos de COX (COXPH), o Random Survival Forest (RSF), o DeepSurv e o DeepHit, em que foram avaliados em cada um os requisitos de pré-processamento de dados, a complexidade da arquitetura do modelo e desempenho preditivo utilizando como métrica o Índice de Concordância. Este índice representa o quão bem os escores de riscos previstos por um modelo estão alinhados com a ordem real dos eventos, e podem ser entendidos da seguinte maneira:

- C-INDEX = 1 (previsão perfeita)
- C-INDEX = 0,5 (palpite aleatório)
- C-INDEX < 0,5 (pior caso)

No pré-processamento foram retirados atributos que não contribuem para o processo de aprendizado (e.g., ID do paciente) e verificado se o conjunto de dados não possuía valores ausentes. Cada método exigiu um conjunto específico de configurações. Os hiperparâmetros de todos os métodos foram otimizados por meio de busca em grade (*grid search*), combinada com validação cruzada de 10 vezes (*10-fold cross-validation*).

Como resultado, após cinco execuções de cada método, os autores relataram que o modelo CoxPH obteve uma C-index médio de 0,887; o RSF obteve uma média de 0,91; e o DeepSurv 0,945. Este último apresentando o melhor desempenho. O DeepHit obteve o pior desempenho, apresentando um valor médio de concordância de 0,6297. Desta forma, ele não foi utilizado nos estudos seguintes.

Para esta pesquisa, como já explicado na metodologia, o conjunto de dados foi separado em duas fases, cada uma com seus hiperparâmetros. Os resultados deste estudo mostraram que, na primeira fase dos experimentos, o modelo CoxPH obteve melhor C-index, o qual mede a capacidade discriminativa do modelo. Na segunda fase, o DeepSurv alcançou o melhor desempenho, como explicado a seguir.

Na primeira fase do experimento, testes com diferentes configurações foram realizados paralelamente, como indicado pela Tabela 2.

Tabela 2. Tabela comparativa de experimentos da primeira rodada

Pré-processamento de internações prolongadas	Pré-processamento de CIDs	Modelo considerado	C-index
Agregação da duração das internações prolongadas	Representação por 22 categorias principais	Cox Proportional Hazards	0.72
		Random Survival Forest	0.67
		DeepSurv	0.71
	Representação com binary encoding	Cox Proportional Hazards	0.73
		Random Survival Forest	0.67
		DeepSurv	0.72

Fonte: Elaborado pela autora com base nos dados da pesquisa (2025).

Nessa fase, no pré-processamento de CIDs com representação por 22 categorias principais e na representação com *binary encoding*, o modelo CoxPH e o DeepSurv tiveram resultados próximos, com o COX um pouco melhor. Já na segunda fase de testes do experimento, foram retirados dados dos anos relacionados à pandemia e usados dados de hospitais gerais no dataset que incluem não só os hospitais públicos municipais mas também hospitais públicos de outras esferas e hospitais privados conveniados ao município que atendem pelo SUS. Nessa fase o DeepSurv se sobressaiu nos resultados. A comparação de experimentos da segunda fase encontra-se na tabela 3.

Tabela 3. Tabela comparativa de experimentos da segunda rodada

Pré-processamento de internações prolongadas	Pré-processamento de CIDs	Modelo considerado	C-index
Agregação da duração das internações prolongadas	Representação por 22 categorias principais	Cox Proportional Hazards	0.73
		Random Survival Forest	0.72
		DeepSurv	0.75
	Representação com binary encoding	Cox Proportional Hazards	0.72
		Random Survival Forest	0.73
		DeepSurv	0.78

Fonte: Elaborado pela autora com base nos dados da pesquisa (2025).

5.2 MODELO DINÂMICO

Para assegurar uma comparação justa, usou-se o mesmo conjunto de dados considerando uma versão coorte (clássico) e uma versão longitudinal (dinâmica). Cada um desses modelos foi avaliado em 20 divisões independentes dos dados (treinamento e teste). Para cada modelo, os experimentos retornaram a média do C-index e o desvio padrão (σ) dessas execuções (Tabela 4). O teste de Wilcoxon foi também utilizado para verificar se existe uma melhoria estatisticamente significativa ($p < 0,05$) apresentada pelo modelo dinâmico.

Também se reportou a média e mediana do Δ C-index, a fim de capturar aspectos complementares da melhoria de desempenho. Enquanto a média reflete a magnitude média da variação no desempenho, a mediana fornece uma estimativa robusta do efeito típico, mitigando a influência de valores extremos.

Tabela 4. Tabela comparativa clássico x dinâmico

Modelo	C-index (média)	σ	p	ΔC-index (média)	ΔC-index (mediana)
Clássico	0.72	0.021	-	-	-
Dinâmico	0.81	0.018	0.0	0.1014	0.1050

Fonte: Elaborado pela autora com base nos dados da pesquisa (2025).

Quando o teste de Wilcoxon apresenta um valor de $p < 0,001$ (e.g., $p = 0$), isso indica que o modelo dinâmico superou consistentemente o modelo clássico de referência em todas as repetições, como observado no experimento realizado. Isso significa que, nas 20 repetições, o modelo dinâmico apresentou desempenho consistentemente e estatisticamente superior ao modelo clássico em termos de C-index.

5.3 DISCUSSÃO

O trabalho de Crispim *et al.* (2025), aplicado a dados sintéticos, motivou a aplicação do mesmo pipeline para desenvolvimento de modelos preditivos junto a um grande dataset real de prontuários eletrônicos de saúde. Desta forma, a intenção seria verificar se os valores dos modelos continuariam acima de 90% de concordância. De fato, este valor teoricamente garante um sistema confiável em sistemas reais quando aplicado à gestão de leitos e controle do tempo de internação dos pacientes.

Ademais, o estudo de Crispim *et al.* ainda relata que os padrões da forma dessas curvas fornecem informações importantes para a equipe de saúde. Por exemplo, uma curva plana ou com variações lentas configura uma hospitalização prolongada por possíveis complicações que precisa de mais monitoramento ou até mesmo mais recursos. Estas motivações deram origem a este trabalho.

Porém, os resultados iniciais mostraram que os valores de C-index ficaram bem abaixo do esperado com o uso de datasets reais. O melhor resultado foi obtido com o DeepSurv (em torno de 0.78). Em geral, ao final deste estudo, foi visto que os métodos CoxPH e DeepSurv se sobressaíram ao outro modelo testado. Porém, quando os registros de internações continham uma melhor distribuição de códigos CID, o que aconteceu com a filtragem de dados de hospitais gerais do dataset trabalhado, o DeepSurv mostrou melhor resultado.

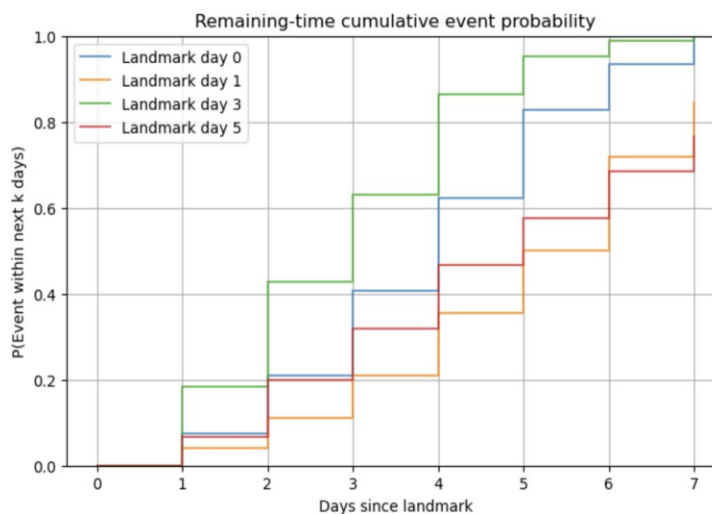
Vale mencionar que, isto já poderia ser esperado dado que o DeepSurv é um método de aprendizagem profunda que modela relações não lineares e, por conseguinte, pode detectar com mais eficiência padrões em registros com características diferentes, como as internações de diversas especialidades. Então ele pode se apresentar como um bom candidato, melhor que o CoxPH, para testes futuros com conjuntos de dados

maiores e reais, diminuindo as limitações encontradas durante a execução deste experimento.

Em relação à abordagem dinâmica, ela conseguiu um resultado melhor, chegando a uma concordância de 0.82 em sua primeira versão. Porém, a pesquisa ainda está em execução e melhoras no algoritmo já vislumbram uma concordância acima de 0.85 (ainda a ser validado), o que já seria um resultado significativo para uso real do modelo. Porém, além de um melhor C-index, o resultado atual obtido com o modelo dinâmico possui uma vantagem clara em relação ao DeepSurv.

Durante o treinamento, os parâmetros do modelo são otimizados utilizando todas as observações longitudinais disponíveis até o momento do evento ou da censura. Durante a inferência, as previsões são geradas em tempos de referência (*landmark times*) arbitrários, fixando o histórico do paciente até o marco temporal considerado e estimando o risco futuro de alta hospitalar. Por exemplo, a Figura 7 ilustra a probabilidade cumulativa do tempo remanescente até a alta estimada em diferentes dias de referência.

Figura 7. Probabilidade cumulativa do evento no tempo remanescente estimada em diferentes dias de referência (landmark days)



Fonte: Elaborado pela autora com base nos dados da pesquisa (2025).

A informação clinicamente relevante é transmitida pela forma e pela evolução das curvas após seu ponto inicial, e não pela sua altura inicial absoluta. Em particular, curvas mais íngremes indicam maior probabilidade de alta no curto prazo, refletindo uma dinâmica de recuperação mais rápida do que a esperada, enquanto curvas mais planas correspondem a processos de alta mais lentos e a internações prolongadas.

Assim, diferenças na inclinação e na progressão das curvas entre diferentes marcos temporais capturam como novos dados clínicos, observados de forma irregular ao longo do tempo, modificam a avaliação do modelo sobre a trajetória de recuperação do paciente. De forma geral, essa visualização responde diretamente à questão clinicamente relevante: “Dado o estado atual do paciente, qual é a probabilidade de o evento ocorrer nos próximos dias?”. Ao recalcular essa probabilidade em cada marco temporal (*landmark*), o modelo fornece um resumo dinâmico e temporalmente fundamentado da evolução do paciente ao longo do tempo.

5.4 PERCEPÇÃO DO USUÁRIO FINAL

Em visita a um hospital municipal, foi mostrado um protótipo da interface de um sistema-projeto construído com os dados e ferramentas utilizados neste trabalho para a diretoria do local, onde foi debatido a funcionalidade do sistema, sugestões e críticas para a melhoria, para assim quando for desenvolvido a ferramenta e colocada em uso no serviço de saúde, ela realmente possa agregar valor.

O diretor técnico relatou que atualmente a equipe de saúde faz uso da metodologia Lean, uma filosofia de melhoria de processos baseada em tempo e valor, desenhada para assegurar fluxos contínuos e eliminar desperdícios e atividades de baixo valor agregado, com o objetivo é reduzir a superlotação nas urgências e emergências de hospitais públicos e filantrópicos por meio do uso da metodologia Lean. Essa metodologia visa melhorar a gestão racionalizando recursos, otimizando espaços e insumos (BRASIL, 2020). O hospital também já faz uso da tabela do SUS que sugere a previsão de alta de um paciente interno de acordo com sua patologia.

Sobre o protótipo apresentado, os dashboards foram bem recebidos, foi sugerido a atualização da curva ser diária ou até mesmo de 12 em 12h, devido a troca de plantões da equipe médica e que o histórico seja guardado para futuras consultas. Foi sugerido que a

pesquisa de paciente seja feita pelo número de internação e que as previsões sejam separadas por setor do hospital e patologia (ex: UTI, ambulatório, UTI cardiológica...) para melhor acompanhamento. Foi relatado que o núcleo de qualidade do hospital e a alta direção que utilizaria esta ferramenta, e seria interessante níveis de acessos distintos para cada profissional. Já para equipe médica em atendimento seria interessante o uso da ferramenta para poder ordenar as altas de pacientes de acordo com a curva de previsão, liberando mais rapidamente o leito para que o NIR (núcleo interno de regulação) do serviço possa disponibilizar o mesmo para um próximo paciente.

Ao final, foi sugerido pela diretoria um estudo que comparasse as curvas de previsão de alta da ferramenta com a tabela já existente do SUS, para se necessário, em futuro próximo tivesse uma sugestão de atualização da mesma junto com o ministério da saúde.

6 CONCLUSÃO

Este trabalho teve como objetivo principal propor e avaliar uma abordagem de análise de sobrevivência dinâmica para a previsão do tempo de permanência hospitalar (*Length of Stay* – LOS), considerando a natureza evolutiva do risco de alta ao longo da internação. Diferentemente de modelos tradicionais, que assumem covariáveis estáticas e produzem estimativas fixas a partir de informações coletadas na admissão, a proposta desenvolvida buscou incorporar atualizações temporais das estimativas de risco, refletindo de forma mais fiel a dinâmica clínica dos pacientes hospitalizados.

A partir de dados reais dos prontuários eletrônicos de hospitais públicos municipais de João Pessoa – PB, foi estruturado um pipeline completo de modelagem preditiva em análise de sobrevivência, contemplando etapas de pré-processamento, tratamento de censura, seleção e codificação de variáveis, além da comparação com métodos amplamente consolidados na literatura, como o Modelo de Riscos Proporcionais de Cox (CoxPH), o *Random Survival Forest* (RSF) e o DeepSurv. Os resultados obtidos indicaram que, de modo geral, os modelos baseados em aprendizado de máquina e aprendizado profundo apresentaram desempenho superior ao modelo clássico de Cox, especialmente no que se refere à captura de relações não lineares e interações complexas entre variáveis clínicas.

No contexto da modelagem dinâmica, a abordagem proposta demonstrou potencial para aprimorar a acurácia preditiva e a utilidade prática das estimativas de tempo de internação, ao permitir a atualização das probabilidades condicionais de alta à medida que novas informações clínicas se tornam disponíveis ao longo da permanência hospitalar. Essa característica é particularmente relevante tanto do ponto de vista clínico, ao oferecer suporte mais aderente à evolução do paciente, quanto do ponto de vista gerencial, ao possibilitar melhor planejamento da ocupação de leitos, organização de fluxos assistenciais e apoio à tomada de decisão em ambientes de alta demanda.

Dessa forma, os achados desta pesquisa reforçam a viabilidade e a relevância da aplicação de técnicas de inteligência artificial e análise de sobrevivência dinâmica no contexto da saúde pública, contribuindo para o avanço metodológico da área e para o

desenvolvimento de sistemas inteligentes de apoio à decisão voltados à gestão hospitalar.

6.1 LIMITAÇÕES DO ESTUDO

Apesar dos resultados promissores, este estudo apresenta algumas limitações que devem ser consideradas na interpretação dos achados. Em primeiro lugar, os dados utilizados são provenientes de um único sistema municipal de saúde, o que pode limitar a generalização dos resultados para outros contextos geográficos, institucionais ou organizacionais. Diferenças nos perfis populacionais, nas rotinas assistenciais e nos sistemas de informação podem impactar o desempenho dos modelos quando aplicados em outros cenários.

Outra limitação relevante refere-se à qualidade e à granularidade dos dados clínicos disponíveis nos prontuários eletrônicos. Embora a base de dados seja extensa, nem todas as variáveis clínicas são registradas de forma padronizada ou em frequência regular ao longo do tempo, resultando em observações longitudinais irregulares e presença significativa de valores ausentes. Embora o modelo tenha sido concebido para lidar com essa característica, tal limitação pode restringir a capacidade de capturar de forma mais precisa a evolução clínica de determinados pacientes.

Além disso, o período pandêmico precisou ser excluído em parte dos experimentos, em função de distorções observadas no perfil das internações. Essa decisão, embora metodologicamente necessária, reduz a abrangência temporal da análise e impede conclusões específicas sobre o comportamento do LOS em cenários de crise sanitária. Por fim, a avaliação do desempenho concentrou-se principalmente no índice de concordância (C-index), não explorando de forma mais aprofundada métricas adicionais de calibração ou impacto clínico direto das previsões.

6.2 TRABALHOS FUTUROS

Como desdobramentos naturais desta pesquisa, diversas oportunidades para trabalhos futuros podem ser apontadas. Uma direção relevante consiste na validação

externa da abordagem proposta, aplicando o modelo a dados provenientes de outros municípios, estados ou sistemas hospitalares, a fim de avaliar sua robustez e capacidade de generalização em diferentes contextos assistenciais. Outra possibilidade é a incorporação de variáveis clínicas mais granulares e temporais, como resultados laboratoriais seriados, sinais vitais contínuos e informações textuais extraídas de anotações médicas por meio de técnicas de processamento de linguagem natural. A integração desses dados pode enriquecer a representação das trajetórias clínicas e potencializar o desempenho da modelagem dinâmica.

Do ponto de vista metodológico, trabalhos futuros podem explorar arquiteturas mais avançadas de aprendizado profundo para sobrevivência dinâmica, incluindo modelos baseados em atenção, *transformers* ou abordagens de múltiplos eventos, considerando desfechos concorrentes como alta, transferência ou óbito. Também é relevante investigar estratégias mais sofisticadas de avaliação, incorporando métricas de calibração temporal e estudos de impacto na tomada de decisão clínica e gerencial.

Por fim, uma linha promissora consiste no desenvolvimento de protótipos de sistemas de apoio à decisão integrados aos sistemas hospitalares existentes, permitindo a utilização prática das estimativas dinâmicas de tempo de internação por gestores e profissionais de saúde, sempre em conformidade com os princípios éticos, regulatórios e de segurança do paciente.

Em síntese, esta pesquisa contribui para o avanço da aplicação de análise de sobrevivência dinâmica na previsão do tempo de internação hospitalar e abre caminho para pesquisas futuras que aprofundem, ampliem e operacionalizem o uso dessas abordagens no contexto da saúde pública.

REFERÊNCIAS

ALHAWARI, H. *et al.* Blood pressure difference between pre and post-menopausal women and age-matched men: a cross-sectional study at a tertiary center. **The Journal of Clinical Hypertension**, v. 26, n. 12, p. 1368-1374, 2024.

ALI, O. *et al.* A systematic literature review of artificial intelligence in the healthcare sector: benefits, challenges, methodologies, and functionalities. **Journal of Innovation & Knowledge**, v. 8, n. 1, p. 100333, 2023.

ALMUQATI, M. T. *et al.* Challenges in Supervised and Unsupervised Learning: A Comprehensive Overview. **International Journal of Advanced Science and Engineering Information Technology**, v. 14, n. 4, p. 1449–1455, 2024.

ALNUAIMI, A. F. A. H. *et al.* Classification techniques in machine learning: overview and applications. *BIO Web of Conferences*, 2024. Disponível em: https://www.bio-conferences.org/articles/bioconf/pdf/2024/16/bioconf_iscku2024_00133.pdf. Acesso em: 17 fev. 2026.

ATAMAN, M. G.; SARIYER, G. Mode-of-arrival awareness models for predicting patient flow and length of stay in emergency departments. **Eurasian Journal of Emergency Medicine**, v. 21, n. 1, p. 34–44, 2022.

BRASIL. Ministério da Saúde. Lean nas emergências. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/l/lean-nas-emergencias>. Acesso em: 23 mar. 2026.

CHEN, C. H.; HSIEH, J. G. CHENG, S. L.; LIN, Y. L.; LIN, P. H. JENG, J. H. Previsão precoce de curto prazo da duração da internação no pronto-socorro usando processamento de linguagem natural para pacientes ambulatoriais de baixa acuidade. **American Journal of Emergency Medicine**, v. 38, n. 11, p. 2368–2373, 2020.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. 2. ed. São Paulo: Edgard Blücher, 2006.

CONSELHO FEDERAL DE MEDICINA. **Resolução CFM nº 2.431, de 5 de junho de 2025**. Brasília: CFM, 2025. Disponível em: https://sistemas.cfm.org.br/normas/arquivos/resolucoes/BR/2025/2431_2025.pdf. Acesso em: 20 set. 2025.

COX, D. R. Regression models and life-tables. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 34, n. 2, p. 187–220, 1972.

CRISPIM, G.; CAMELO, V.; SIEBRA, H.; SIEBRA, C.; SOUZA FILHO, G. A comparative evaluation of survival analysis methods for predicting hospital length of stay. In: Conference Of The Latin American Center For Informatics Studies – CLEI, 2025, [S. l.]. **Anais...** [S. l.: s. n.], 2025.

FOOD AND DRUG ADMINISTRATION (FDA). **Artificial Intelligence/Machine**

Learning (AI/ML)-Based Software as a Medical Device Action Plan. Silver Spring:

FDA, 2021. Disponível em: <https://www.fda.gov/media/145022/download>. Acesso em: 20 set. 2025.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. Cambridge: MIT Press, 2016.

GORGENS, E. B. *et al.* Estimaco do volume de rvores utilizando redes neurais artificiais. **Revista rvore**, v. 33, n. 6, p. 1141-1147, 2009.

GUEDES, A. F.; SILVA, M. A. Anlise exploratria de dados: fundamentos e aplicaes. **Revista Brasileira de Estatística**, v. 80, n. 3, p. 1–20, 2022.

GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras & TensorFlow**. 2. ed. Sebastopol: O'Reilly Media, 2019.

HOSMER, D. W.; LEMESHOW, S.; MAY, S. **Applied Survival Analysis: Regression Modeling of Time-to-Event Data**. 2. ed. Hoboken: Wiley, 2008.

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied logistic regression**. 3. ed. Hoboken: Wiley, 2013.

ISHWARAN, H.; KOGALUR, U. B.; BLACKSTONE, E. H.; LAUER, M. S. Random survival forests. **The Annals of Applied Statistics**, v. 2, n. 3, p. 841–860, 2008.

JOSHI, A. P.; PATEL, R. P. Data Preprocessing: The Techniques for Preparing Clean and Quality Data for Data Analytics Process. **Oriental Journal of Computer Science & Technology**, v. 13, n. 2–3, p. 78–81, 2020.

KAPLAN, E. L.; MEIER, P. Nonparametric Estimation from Incomplete Observations. **Journal of the American Statistical Association**, v. 53, n. 282, p. 457–481, 1958.

KARBUBAND, K.; TABAA, M. Optimization of bed allocation based on survival analysis, treatment trajectory, and cost estimates. **IEEE Access**, v. 11, p. 31699–31715, 2023.

KATZMAN, J. L.; SHAHAM, U.; CLONINGER, A.; BATES, J.; JIANG, T.; KLUGER, Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. **BMC Medical Research Methodology**, v. 18, n. 1, p. 1–24, 2018.

KLEINBAUM, D. G.; KLEIN, M. **Survival Analysis: A Self-Learning Text**. 3. ed. New York: Springer, 2012.

KOUKARAS, P.; TJORTJIS, C. Pré-processamento de dados e engenharia de recursos para minerao de dados: tcnicas, ferramentas e melhores prticas. **AI**, v. 6, n. 10, p. 257, 2025.

LEE, C.; ZAME, W.; YOON, J.; VAN DER SCHAAR, M. DeepHit: a deep learning approach to survival analysis with competing risks. In: **AAAI Conference On Artificial Intelligence**, 32., 2018, [S. l.]. Proceedings of the AAAI Conference on Artificial Intelligence. [S. l.: s. n.], 2018. v. 32, n. 1.

LI, C. **Preprocessing methods and pipelines of data mining: an overview**. Seminar Data Mining, Technical University of Munich, jun. 2019. Disponível em: <https://arxiv.org/abs/1906.08510>. Acesso em: 25 jul. 2025.

LU, W.; CHEN, M.; HAO, Z. Deep Learning Models for Survival Analysis. **IEEE Transactions on Neural Networks and Learning Systems**, Piscataway, v. 31, n. 10, p. 3586-3597, 2020.

MEDEIROS, N. B.; FOGLIATTO, F. S.; ROCHA, M. K. et al. Prediction of the length of stay of pediatric patients in hospitals: a scoping review. *BMC Health Services Research*, v. 21, n. 1, p. 938, 2021.

MIENYE, I. D.; SUN, Y. A comprehensive review of deep learning: architectures, applications and recent trends. **Information**, v. 15, n. 12, 2024. DOI: 10.3390/info15120755.

MOBINI SERAJI, M. H. et al. A state of the art review on machine learning techniques for driving behavior analysis: clustering and classification approaches. **Complex & Intelligent Systems**, [S. l.], artigo 386, jul. 2025. Disponível em: <https://link.springer.com/article/10.1007/s40747-025-01988-5>. Acesso em: 17 jul. 2025.

NELSON, W. **Applied Life Data Analysis**. New York: Wiley, 2004.

NURHALIZAH, R. S.; *et al.* **Analisis supervised and unsupervised learning pada machine learning**. *JIKI*, 2024. Disponível em: <https://jiki.jurnal-id.com/index.php/jiki/article/download/168/47>. Acesso em: 15 fev. 2026.

ORGANIZAÇÃO MUNDIAL DA SAÚDE (OMS). **Ethics and governance of artificial intelligence for health**. Genebra: WHO, 2021. Disponível em: <https://www.who.int/publications/i/item/9789240029200>. Acesso em: 20 set. 2025.

PEACE, P.; CHRIS, J.; VICTOR, L. **Data preprocessing for AI models**. [S. l.]: ResearchGate, nov. 2024. Disponível em: <https://www.researchgate.net/publication/385707249>. Acesso em: 25 jul. 2025.

PENG, B.; GAO, S. Prediction of hospital length of stay: leveraging ensemble tree models and intelligent feature selection. **Journal of Hospital Management and Health Policy**, v. 9, n. 1, p. 13, 2025.

PÖLSTERL, S. scikit-survival: a library for time-to-event analysis built on top of scikit-learn. **Journal of Machine Learning Research**, v. 21, n. 212, p. 1–6, 2020.

REN, Y. et al. **Deep Clustering: A Comprehensive Survey**. arXiv preprint, out. 2022. Disponível em: <https://arxiv.org/abs/2210.04142>. Acesso em: 17 jul. 2025.

RYMARZ, A. et al. Futuro do diagnóstico cardiovascular com apoio da inteligência artificial. **Revista de Educação, Saúde e Esporte On-line**, v. 44, n. 1, p. 11-22, 2023.

SANTOS, H. G. et al. Machine learning para análises preditivas em saúde: exemplo de aplicação para prever óbito em idosos de São Paulo, Brasil. **Cadernos de Saúde Pública**, v. 35, n. 7, p. e00050818, 2019.

SHICKEL, B. et al. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. **IEEE Journal of Biomedical and Health Informatics**, v. 22, n. 5, p. 1589–1604, 2018.

SILVA, G. G.; et al. Desafios do uso da inteligência artificial nos diagnósticos de saúde: uma revisão integrativa. **Cadernos Ibero-Americanos de Direito Sanitário**, [S. l.], v. 13, n. 2, p. 11-18, 2024.

SILVA, R. **Inteligência Artificial**. Enciclopédia da Conscienciologia [Internet], 2013. Disponível em: <http://www.reposicons.org/handle/123456789/3737>. Acesso em: 02 ago. 2025.

SINGH, A.; PANDEY, N. Survival Analysis in Business Applications: A Review. **Journal of Business Analytics**, v. 4, n. 1, p. 34-50, 2021.

STONE, K.; ZWIGGELAAR, R.; JONES, P.; MAC PARTHALÁIN, N. A systematic review of the prediction of hospital length of stay: Towards a unified framework. **PLOS Digital Health**, v. 1, n. 4, p. e0000017, 2022.

UNIÃO EUROPEIA. **Artificial Intelligence Act**. Bruxelas: Parlamento Europeu, 2024. Disponível em: <https://artificialintelligenceact.eu>. Acesso em: 20 set. 2025.

WAHYUDIN, et al. Machine Learning Based Clustering for Program Learning Outcomes in Higher Education: A Systematic Review. **Brilliance**, v. 5, n. 1, p. 182-189, 2025.

WEN, Y.; et al. Time-to-event modeling for hospital length of stay prediction in COVID-19 patients. **BMC Medical Informatics and Decision Making**, v. 9, n. 1, p. 100365, 2022. em: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9213016/>. Acesso em: 17 set. 2025.

TIAN, C.; ZHANG, Q.; WANG, F.-Y.; ZHANG, D. A survey on deep learning fundamentals. **Artificial Intelligence Review**, v. 58, n. 381, p. 1-108, 2025.

WANG, P. et al. Machine Learning for Survival Analysis: Methods and Applications. **ACM Computing Surveys**, v. 53, n. 4, p. 1-36, 2020.

ZHOU, S. et al. A Comprehensive Survey on Deep Clustering: Taxonomy, Challenges, and Future Directions. **ACM Computing Surveys**, [S. l.], p. 1–37, 2022.