CONSTRUÇÃO DE UM DATASET PARA ANÁLISE DE ALGORITMOS DE APRENDIZADO DE MÁQUINA EM SAÚDE PARA O HOSPITAL ESCOLA DE PELOTAS

LUCIANO LUDWIG HELING[1]; RICARDO NETTO GOULART[2]; AMANDA JHENNIFER MARQUES VIEIRA[3]; MATHEUS LOPES DE FERNANDES[4]; LEANDRO FARIAS RODRIGUES[5]; BRUNO PEREIRA NUNES[6]

[1] *Universidade Federal de Pelotas – llheling@inf.ufpel.edu.br*
[2] *Universidade Federal de Pelotas – ricardonettogoulart@gmail.com*
[3] *Universidade Federal de Pelotas – ajmvieira@inf.ufpel.edu.br*
[4] *Hospital Escola da UFPel Ebserh – matheus.fernandes@ebserh.gov.br*
[5] *Hospital Escola da UFPel Ebserh – leandro.farias@ebserh.gov.br*
[6] *Universidade Federal de Pelotas – nunesbp@gmail.com*

1. INTRODUÇÃO

A incorporação da inteligência artificial (IA) na área da saúde representa um avanço significativo, com notável potencial para transformar as práticas em saúde e aprimorar a qualidade do atendimento aos pacientes. Entre as possibilidades para esses sistemas, destacam-se a detecção precoce de doenças, a adaptação de tratamentos com base nas particularidades de cada paciente e até mesmo a predição de desfechos clínicos (MUDGAL *et al*., 2022).

No entanto, para o desenvolvimento e avaliação desses algoritmos preditivos, é fundamental a construção de conjuntos de dados (*datasets*) confiáveis e válidos para o treinamento em aprendizado de máquina (AM). Nesse sentido, um dos principais desafios reside no processamento dos dados. Os registros em saúde podem conter erros, omissões e inconsistências que precisam ser identificados e corrigidos, para garantir o bom funcionamento dos algoritmos (LIU *et al.*, 2023). Mesmo as informações de boa qualidade, podem estar em formatações que diferem entre os profissionais que os registraram. Em concomitância, é também crucial seguir os procedimentos que garantam a anonimização da amostra.

O objetivo deste trabalho é relatar o processo de processamento dos dados para a construção do *dataset,* em conjunto com o Setor de Tecnologia da Informação do Hospital Escola da Universidade Federal de Pelotas (HE/UFPel/Ebserh) — demonstrando seus passos técnicos, dificuldades e aprendizados deste processo.

2. METODOLOGIA

A metodologia empregada na construção deste *dataset* envolveu diversas etapas durante a elaboração do projeto "Inteligência Artificial como Inovação Tecnológica em Saúde para predição de desfechos entre usuários dos Leitos de Retaguarda de Urgências e Emergências do HE/UFPel". Este projeto, alvo do trabalho, está em desenvolvimento no Hospital Escola da Universidade Federal de Pelotas (HE-UFPEL/EBSERH), fazendo parte do Programa de Iniciação Tecnológica em parceria com o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). O projeto conta com um (1) bolsista de Iniciação Tecnológica, um (1) Bolsista de Iniciação Científica e voluntários nas áreas de Ciência da Computação, Engenharia da Computação, Medicina e Nutrição. Participam também profissionais que atuam no Hospital Escola nas áreas de Sistema de Informação e Inteligência de Dados, Regulação e Inovação

Tecnológica em Saúde. O objetivo final do trabalho é a criação de uma ferramenta focada na predição de desfechos — como reinternação, mortalidade e tempo de estadia — na Atenção Terciária, especificamente nos leitos da Rede de Urgência e Emergência (RUEs) do Hospital Escola da UFPel. Para isso, utilizaremos uma abordagem associada ao aprendizado de máquina e inteligência artificial.

Como primeiro passo deste processo, foram realizadas reuniões com as equipes de faturamento e Tecnologia da Informação do HE UFPel para discutir e planejar o funcionamento das ações. A primeira etapa consistiu na obtenção das planilhas abrangendo as internações dos pacientes que passaram pelos Leitos Regulados no período de janeiro a dezembro de 2022 (Figura 1). Essas planilhas forneceram a base para a pesquisa subsequente no banco de dados com informações extraídas do sistema de gestão hospitalar (AGHU), adotado como padrão em todos os Hospitais Universitários Federais da rede Ebserh.

Dessas planilhas, a principal variável utilizada foi o número de prontuários do paciente (Figura 2). Importante ressaltar que, como parte da metodologia adotada, ao fim deste processo inicial, removemos os nomes dos pacientes para garantir a anonimização das informações sensíveis, em conformidade com as diretrizes da Lei Geral de Proteção de Dados (LGPD).

Na segunda etapa, foi realizada uma análise abrangente das variáveis contidas no banco de dados — dentre elas, temos como exemplo as características demográficas dos pacientes, classificação de sua doença (CID), tempo de permanência, motivo da alta, uso de medicamentos e solicitações de exames. Essa análise visava não apenas identificar as informações disponíveis, mas também avaliar a confiabilidade e relevância dessas variáveis para nossos objetivos de pesquisa. Destacaram-se variáveis de maior significado, que se tornaram candidatas para consultas separadas devido ao volume substancial de dados. Dentre elas, tivemos acesso à variáveis relacionadas à multimorbidade, relevante preditor para desfechos em saúde (Alonso-Morán *et al.,* 2015; NUNES *et al*., 2016). O estudo foi submetido ao Comitê de Ética em Pesquisa via Plataforma Brasil e aprovado com número de parecer 5.779.581.

3. RESULTADOS E DISCUSSÃO

Na terceira etapa, unificamos as planilhas individuais em uma única fonte de dados coerente, e, com base nos parâmetros definidos durante a pesquisa, realizamos uma seleção dos Leitos Regulados de Urgência e Emergência, separando-os dos leitos que não se enquadravam no escopo do estudo.

Paciente	Prontuário	Tempo de Internação
PACIENTE	X	Y
PACIENTE	X	Y
PACIENTE	X	Y
PACIENTE	X	Y
PACIENTE	X	Y
PACIENTE	X	Y
PACIENTE	X	Y
PACIENTE	X	Y

Figura 1. Ilustração do perfil de dados brutos enviados pelo Hospital (HE-UFPel).

No entanto, surgiram dúvidas relacionadas às transferências de pacientes. Dentre elas, como avaliaríamos os movimentos dos pacientes entre os leitos da

rede de saúde. Então, como resultado de reuniões em conjunto, foi definido que, para uma maior precisão, qualquer tipo de transferência seria removida, chegando a um número fixo de 1066 internações. Na figura abaixo, é demonstrado como foi feita a conferência da primeira extração, onde temos a esquerda pacientes que estavam na lista recebida inicialmente do faturamento e à direita pacientes resultantes da primeira pesquisa no banco de dados.

Paciente LEITOS REGULADO	Prontuário	CONSULTA AGHUX	Prontuário		COMPARA NOME	COMPARA PRONTUÁRIO
PACIENTE	x	PACIENTE	x		VERDADEIRO	VERDADEIRO
PACIENTE	x	PACIENTE	x		VERDADEIRO	VERDADEIRO
PACIENTE	x	PACIENTE	x		VERDADEIRO	VERDADEIRO
PACIENTE	x	PACIENTE	x		VERDADEIRO	VERDADEIRO
PACIENTE	x	PACIENTE	x		VERDADEIRO	VERDADEIRO
PACIENTE	x	PACIENTE	x		VERDADEIRO	VERDADEIRO
PACIENTE	x	PACIENTE	x		VERDADEIRO	VERDADEIRO
PACIENTE	x				FALSO	FALSO
PACIENTE	x	PACIENTE	x		VERDADEIRO	VERDADEIRO
PACIENTE	x	PACIENTE	x		VERDADEIRO	VERDADEIRO
PACIENTE	x	PACIENTE	x		VERDADEIRO	VERDADEIRO
PACIENTE	x	PACIENTE#	x		FALSO	VERDADEIRO
PACIENTE	x	PACIENTE	x		VERDADEIRO	VERDADEIRO
		PACIENTE	x		FALSO	FALSO
PACIENTE	x	PACIENTE	x		VERDADEIRO	VERDADEIRO
PACIENTE	x	PACIENTE	x		VERDADEIRO	VERDADEIRO
PACIENTE	x	PACIENTE	x		VERDADEIRO	VERDADEIRO
PACIENTE	x	PACIENTE#	x		FALSO	VERDADEIRO

Figura 2. Conferência dos dados recebidos pela regulação do Hospital Escola (HE-UFPel), com as informações coletadas pelo Aplicativo de Gestão para Hospitais Universitários (AGHU).

Usando a linguagem SQL (Structured Query Language), linguagem para gerenciar e manipular bancos de dados relacionais (DILLING, 2020), encontramos um número de 1016 internações, sendo necessário destinar tempo para encontrar os 50 pacientes faltantes. A fase de testes de extração se tornou uma etapa complexa, especificamente devido a erros no número de identificação do prontuário, que serviram como a principal chave de pesquisa no banco de dados, assim como imprecisões nas datas de internação e alta. Consequentemente, este momento exigiu a atenção detalhada na validação dos dados, pois a precisão dessas informações é crucial para assegurar a confiabilidade e a representatividade das análises e resultados derivados das consultas. Para fins de visualização da coleta e extração de dados, trazemos na Figura 3 uma demonstração de possíveis informações a serem obtidas por meio da linguagem SQL.

n_internacao	permanencia	motivo_alta	unidade_internacao
X	Y	Z	REDE DE URGÊNCIA E EMERGÊNCIA III
X	Y	Z	REDE DE URGÊNCIA E EMERGÊNCIA II
X	Y	Z	REDE DE URGÊNCIA E EMERGÊNCIA II
X	Y	Z	REDE DE URGÊNCIA E EMERGÊNCIA II
X	Y	Z	REDE DE URGÊNCIA E EMERGÊNCIA III
X	Y	Z	REDE DE URGÊNCIA E EMERGÊNCIA II
X	Y	Z	REDE DE URGÊNCIA E EMERGÊNCIA III
X	Y	Z	REDE DE URGÊNCIA E EMERGÊNCIA II

Figura 3. Exemplo hipotético de banco de dados passível de extração via SQL.

4. CONCLUSÕES

Ao término deste processo, emergem reflexões sobre a construção deste conjunto de dados, *dataset,* para a análise de algoritmos de inteligência artificial na área da saúde. Além disso, a aquisição deste conhecimento sobre técnicas de aprendizado de máquina e a busca por noções na construção de um algoritmo de predição, utilizando linguagem de programação Python, adicionou uma dimensão valiosa ao nosso conjunto de ferramentas, que possam transformar esta ideia em realidade.

Portanto, os resultados obtidos até o momento refletem não apenas a progressão bem-sucedida do projeto, mas também a promessa de contribuir significativamente para a aplicação de inteligência artificial na área da saúde dentro de um hospital universitário do Sistema Único de Saúde. À medida que avançamos para a próxima fase da pesquisa, essas conclusões fornecerão uma base sólida para o desenvolvimento e a implementação de algoritmos de IA destinados a melhorar a assistência em saúde e aprimorar o atendimento aos pacientes.

5. REFERÊNCIAS BIBLIOGRÁFICAS

ALONSO-MORÁN, E. *et al*. Multimorbidity in risk stratification tools to predict negative outcomes in adult population. **European Journal of Internal Medicine, v. 26, n. 3, p. 182–189**, 1 abr. 2015.

CHICCO, D.; ONETO, L.; TAVAZZI, E. Eleven quick tips for data cleaning and feature engineering. **PLOS Computational Biology**, v. 18, n. 12, p. e1010718–e1010718, 15 dez. 2022.

DILLING, T. J. Artificial Intelligence Research: The Utility and Design of a Relational Database System. **Advances in radiation oncology**, v. 5, n. 6, p. 1280–1285, 1 nov. 2020.

LIU, M. *et al*. Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. **Artificial Intelligence in Medicine**, v. 142, p. 102587–102587, 1 ago. 2023.

MUDGAL, S. K. *et al*. Real-world application, challenges and implication of artificial intelligence in healthcare: an essay. **The Pan African Medical Journal**, v. 43, p. 3, 2 set. 2022.

NUNES, B. *et al*. Multimorbidity and mortality in older adults: A systematic review and meta-analysis. **Archives of Gerontology and Geriatrics**, v. 67, p. 130–138, 1 nov. 2016.