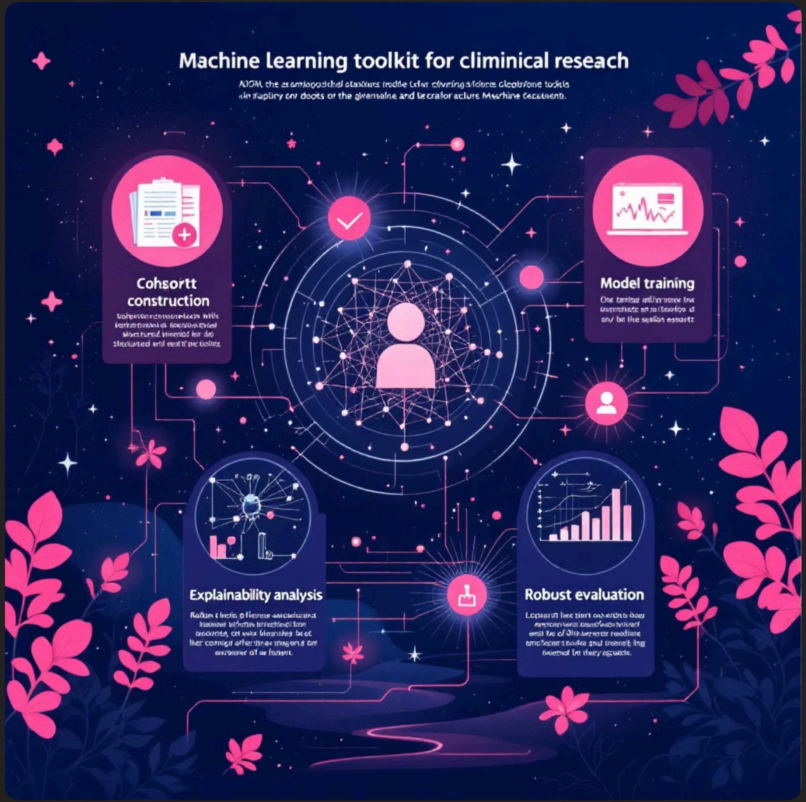
Integration Manual: ML Study in COVID-19 ICU

Welcome to the complete and reproducible pipeline for developing predictive models of mortality and length of stay in the ICU for critically ill patients with COVID-19, using Brazilian national data (INFLUD/SRAG, 2020–2024).

Project Overview

This repository offers basic tools for exploring DataSUS clinical data in ML:

- Path to raw files
- Jupyter Notebook
- ICU adult cohort construction
- Main graphics on explored data



Data Source and Initial Configuration

INFLUD dataset National database on Severe Acute Respiratory Syndrome (SARS), 2020–2024, available at opendatasus.saude.gov.br

Storage Raw

files (hundreds of MB) should be saved in data/raw/ before starting processing.

Inclusion Criteria

Only adult ICU patients with confirmed COVID-19. Pediatric cases and inconsistent records were excluded.

Important: The files are large. Make sure you have adequate RAM.

Repository Folder Structure

```
icu-rep/
 ├──── data/
 │    ├──── raw/
 │    ├──── processed/
 ├──── notebooks/
 │    ├──── evolucao_V12_git.ipynb
 ├──── src/
 │    ├──── cohort_building.py
 │    ├──── variable_curation.py
```

Modular organization facilitates navigation, maintenance, and collaboration among research team members.

Variable Curation: Final Hybrid Selection

1

Excluded Administrative Variables

DT_ENCERRA and DT_EVOLUCA were removed because they did not add predictive value (2 → 0 variables). 2 Added Comorbidities HEMATOLOGY and HEPATIC manually included due to clinical relevance (0 → 2 variables)

2

Included Clinical Variables (comorbidities)

2 Added Comorbidities HEMATOLOGY and HEPATIC manually included due to clinical relevance (0 → 2 variables)

3

Demographic and Epidemiological

10 variables including age, sex, municipal and regional codes, epidemiological weeks

4

Clinical and Support

19 variables covering symptoms (dyspnea, fever, cough), comorbidities (heart disease, lung disease, kidney disease), and ventilatory support ...

5

ICU Evolution

Three critical time variables: DT_ENTUTI, DT_SAIDUTI, and calculated LOS_ICU

> Total final variables: 34 — balance between informational richness and model complexity

Cohort Construction Pipeline

Load Raw Dataset

Import CSV from INFLUD

Filter Adult Patients

Exclude NU_IDADE_N < 18 anos and not probable ages (>110) to guarantee valid cohort

Confirm COVID-19 Cases

Use CLASSI_FIN to confirm diagnostics, removing influenza, RSV and other etiologies

Select Cohort ICU

Maintaing only registers with DT_ENTUTI, excluding patients that have not entered ICU in their digital documentation válida.

Treating Absent Values

Apply specific techniques

Final Curation

Mantain 34 selected variables for data/processed/cohort_v12.csv

Reproducibility: Quick Start

01

Repository

```
git repository
https://github.com/SEU_USUARIO/
icu-rep.git
cd icu-rep
pip install -r requirements.txt
```

02

Cohort

```
python src/cohort_building.py
```

Processes raw data and applies all inclusion filters

03

Visualizations

```
python
src/visualization/shap_plots.py
python
```

Team Work

Begin Researching

You now have the tools you need to contribute to this research project. The pipeline is designed to be intuitive, but please feel free to consult the technical documentation or contact the team: lia.graca@unifesp.br; lia.systemslab@gmail.com

Resources

- Cohort notebook
- Documentation at README.md
- Graphics

How to cite

> GRAÇA, Lia da. *Sistema de soporte translacional con aprendizaje automático para la optimización de la asignación equitativa de camas de UCI y el traslado de pacientes.* 2025. Tesis (PhD Program in Sciences) – Universidade Federal de São Paulo, São Paulo, ano. Available at: https://repositorio.unifesp.br/items/7188041c-7d05-46fe-9637-d5d7880cd339

Support: For technical questions or contributions, open an issue in the GitHub repository or contact the project maintainers lia.graca@unifesp.br