# COVID-19 ICU Admission Prediction Project

## Requirements & Evidence Matrix

### Executive Summary

This project develops a machine learning–based early-warning system to predict ICU admissions for COVID-19 patients in Brazil using the Sírio-Libanês Hospital dataset (1,925 records, 231 features, 5 temporal windows). It bridges data science, hospital operations, and clinical leadership, aiming to optimize ICU resource allocation under pandemic strain. The deliverables include a unified Jupyter notebook, executive presentation, and technical research report, all verified in GitHub.

Stakeholders:

• Product Owner: QS Academy Faculty – Clinical Data Analytics Track

• Data Scientist: Nathan Weber

• Reviewers: 2025 Cohort – Applied Machine Learning Capstone

• Use Case: Early ICU prediction and hospital resource optimization  
  
GitHub link: [https://github.com/isthatarequirement/CAPSTONE](https://na01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fgithub.com%2Fisthatarequirement%2FCAPSTONE&data=05%7C02%7C%7C6a34fb33f58541d3fa2808de06cf0ef4%7C84df9e7fe9f640afb435aaaaaaaaaaaa%7C1%7C0%7C638955683532262754%7CUnknown%7CTWFpbGZsb3d8eyJFbXB0eU1hcGkiOnRydWUsIlYiOiIwLjAuMDAwMCIsIlAiOiJXaW4zMiIsIkFOIjoiTWFpbCIsIldUIjoyfQ%3D%3D%7C0%7C%7C%7C&sdata=k0SGTjJZopOcsIOKxiRFqKxDUXnZRRgxn%2B44spXXkic%3D&reserved=0)

### Requirements–Evidence Traceability Matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Requirement Area | Deliverable / Description | Evidence / Proof of Completion | Priority |
| PSR-1 | Dataset Load & Integrity Check | Load and preview dataset; confirm shape, dtypes, column list, and head/tail. | Jupyter notebook cells 5–10: df.shape, df.info(), df.head() | High |
| PSR-2 | Unified Notebook Submission | Single `.ipynb` with all code, markdown, visuals; README summarizes workflow. | GitHub repo: CAPSTONE\_NWEBER\_notebook.ipynb, README.md | High |
| PSR-3 | Missing/Outlier Cleaning | Drop duplicates; show missingness; remove >70% missing; impute rest. | Cells 15–25: missing-value summary before/after cleaning | High |
| PSR-4 | Feature Encoding | Encode categorical features (AGE\_PERCENTIL, comorbidities). | map() and get\_dummies() operations in notebook | High |
| PSR-5 | Standardization | Scale numeric features using StandardScaler. | Pipeline definition in cells 40–45; verified on test set | High |
| PSR-6 | EDA & Visualization | ≥5 labeled charts: ICU rate, demographics, comorbidities, vitals, completeness. | Charts in cells 50–90; exports: icu\_rates\_chart.png, icu\_demographics\_chart.png | High |
| PSR-7 | Temporal Analysis | ICU admission probability across 0–2h → >12h windows. | Line chart shows 8.3% → 50.6% increase (cells 55–60) | High |
| PSR-8 | Modeling – Baselines | Train Logistic Regression, Random Forest, Gradient Boosting baselines. | Cells 100–120: pipeline outputs with F1, ROC-AUC, confusion matrices | High |
| PSR-9 | Hyperparameter Tuning | Nested 5-fold CV using GridSearchCV optimizing F1. | Cells 125–150: best params table & fold F1 results | Critical |
| PSR-10 | Ensemble & Stacking | Combine tuned models into a stacking ensemble with logistic meta-learner. | Cells 155–170: ensemble code and F1 ≥ 0.90 | High |
| PSR-11 | Recall Optimization | Tune thresholds and class weights to minimize false negatives. | Precision-recall and threshold plots (cells 175–185) | High |
| PSR-12 | Feature Importance & Interpretability | Display top-10 features with clinical relevance notes. | Exports: rf\_feature\_importance.jpg, shap\_feature\_importance.jpg | High |
| PSR-13 | Model Evaluation Metrics | Report accuracy, F1, ROC-AUC, recall, specificity, confusion matrices. | Cells 190–205: classification\_report(), confusion\_matrix() | High |
| PSR-14 | Executive Presentation | Narrative deck: Summary, Clinical Insights, Operational Recommendations. | Capstone\_Model\_Slides.pptx | High |
| PSR-15 | Repository & Documentation | GitHub repo with commits, reports, and structured folders. | Verified commit log, README.md, folder structure | High |
|  |  |  |  |  |

### Deliverable Summary

|  |  |  |
| --- | --- | --- |
| Deliverable | Description | Format |
| Exploratory Data Analysis | Comprehensive temporal, demographic, and lab analysis | Jupyter Notebook / PDF |
| Data Preparation | Cleaning, encoding, imputation, scaling | Python (.ipynb) |
| ML Model Development | Tuned LR, RF, GBM, and stacking ensemble | Python (.ipynb) |
| Executive Presentation | Leadership-focused slide deck | PowerPoint (.pptx) |
| Research Report | Technical report with methodology and CI analysis | Word / PDF |
| ~~Power BI Dashboard~~ | ~~Live ICU analytics dashboard (optional)~~ | ~~.pbix~~ |
| GitHub Repository | Full reproducible code and docs | Public link |