A Data-Driven and Explainable Machine Learning Approach to Predict Prolonged Hospitalization in Brazilian SARS Patients

Fernando Oliveira[1]
orcid.org/0000-0008-1078-8416

Cleyton Rodrigues[1]
orcid.org/0000-0003-0003-656X

[1]Escola Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil. E-mail: fhmo@ecomp.poli.br

DOI: 10.25286/repa.v11i1.3537

Como citar este artigo pela NBR 6023/2018: Fernando Oliveira; Cleyton Rodrigues. A Data-Driven and Explainable Machine Learning Approach to Predict Prolonged Hospitalization in Brazilian SARS Patients. Revista de Engenharia e Pesquisa Aplicada, v.11, n. 1, p. 41-50, 2026.

ABSTRACT

Severe Acute Respiratory Syndrome (SARS) continues to pose a substantial public health challenge in Brazil, with prolonged hospitalizations increasing pressure on healthcare resources. This study utilized Brazil's national SIVEP-Gripe surveillance system, a comprehensive repository of anonymized, individual-level records for SARS cases including influenza and other respiratory viruses, to develop and evaluate machine learning models. Using data from 2024, we constructed a preprocessed dataset consisting of 64,238 hospitalized patient records. This dataset was built using 32 independent variables, all of which are available at the time of patient admission. The focus of this dataset is to predict prolonged hospital length of stay (PLOS > 7 days). Three ensemble tree-based algorithms—Random Forest, XGBoost, and CatBoost—were trained after data preprocessing and robust imputation, using stratified 5-fold cross-validation with AUC maximization. The models exhibited moderate but consistent predictive performance, with AUC values around 0.65. XGBoost achieved the best balance between sensitivity and specificity, while Random Forest achieved higher recall for prolonged-stay cases. Explainable AI analysis using SHAP values revealed asthma, age, oxygen saturation, and geographic region as the most influential predictors. These findings underscore the potential of explainable machine learning approaches to support early hospital resource planning using routinely collected surveillance data. Future research should incorporate dynamic and clinical progression variables to further enhance predictive performance and real-world applicability.

KEY-WORDS: Severe Acute Respiratory Syndrome; Explainable Artificial Intelligence; Machine Learning; Hospital Length of Stay; Prolonged Hospitalization.

1 INTRODUCTION

Severe acute respiratory syndrome (SARS) poses a major global public health challenge, significantly increasing morbidity, mortality, and healthcare burdens, especially among vulnerable groups like young children and the elderly. In Brazil, SARS surveillance is primarily conducted through the SIVEP-Gripe (Sistema de Informação da Vigilância Epidemiológica da Gripe) system. This system has been instrumental in monitoring the circulation of various respiratory pathogens, and most notably since 2020, SARS-CoV-2, the causative agent of COVID-19.

The COVID-19 pandemic has highlighted the critical importance of early risk stratification and efficient hospital resource management, particularly in low- and middle-income countries [1]. In this context, the surveillance of SARS remains a key component of Brazil's public health monitoring strategy. Despite advancements, length of stay (LOS) prediction research remains fragmented and often specific to individual hospitals, lacking broad applicability [2]. Prolonged Length of Stay (PLOS) in hospitalized patients with SARS is a critical concern, associated with increased healthcare costs, higher risk of nosocomial infections, greater patient morbidity, and strained hospital resources [3]. Identifying patients at high risk for PLOS early in their hospital course can enable targeted interventions, optimize resource allocation, and potentially improve outcomes.

While numerous studies have characterized the clinical and epidemiological features of SARS in Brazil [9][10], a significant gap exists in developing predictive models for PLOS using only data available at hospital admission. Previous research on COVID-19 has identified various predictors for hospital stay duration, such as respiratory parameters, laboratory markers, and patient demographics [1, 4-7], but these have not been synthesized into a broadly applicable model for the Brazilian SARS population.

A key challenge was rigorously preparing the raw, heterogeneous SIVEP-Gripe data for modeling. The complex dataset used numerical or character codes, requiring extensive cleaning and decoding via a data dictionary. Missing data—coded as 'ignored' or 'unknown'—necessitated a careful imputation strategy. Importantly, all data cleaning and feature engineering steps were constrained to variables available at hospital admission, a crucial constraint for developing an early-prediction model. This meticulous process of handling data heterogeneity, cleaning inconsistencies, and strategically imputing missing values was fundamental to constructing a reliable dataset for the subsequent modeling and validation phases.

Machine learning (ML) approaches are well-suited to address the heterogeneity of data. They can analyze complex, high-dimensional datasets, such as SIVEP-Gripe, to identify subtle patterns among patient characteristics that contribute to prolonged hospitalization. However, the "black box" nature of some complex ML models can hinder clinical adoption due to a lack of transparency. This has driven interest in Explainable Artificial Intelligence (XAI), which provides the interpretability and transparency necessary for clinical applications [7]. By making model predictions understandable, XAI can foster trust and facilitate the integration of ML tools into healthcare decision-making.

Therefore, this study develops and evaluates machine learning models to predict prolonged length of stay (PLOS) in hospitalized SARS patients using Brazil's SIVEP-Gripe system. While PLOS is commonly defined as exceeding the median LOS [16] (6 days in our dataset), we used >7 days for a more conservative threshold to identify high-resource cases. By analyzing demographic, clinical, and comorbidity data available at admission, we aimed to identify key predictors of prolonged hospitalization and evaluate various modeling techniques. The insights gained could enhance clinical decisions and public health strategies, leading to better patient outcomes and more efficient healthcare resource management. This study presents a reproducible and interpretable data science workflow applied to public health data, highlighting the challenges of missing data, model selection, and explainability.

The remainder of this manuscript is structured as follows: Section 2 presents the background necessary for work understand. Section 3 details the materials and methods, including data description, preprocessing, feature engineering, model development, and evaluation strategy. Section 4 reports the results of exploratory data analysis, model performance, and explainability using SHAP values. Section 5 discusses the study's limitations, and Section 6 concludes with key findings and directions for future work.

2 BACKGROUND

Severe acute respiratory syndrome (SARS) is characterized by fever, cough or sore pain, alongside shortness of breath or oxygen saturation levels below 95% [15]. These conditions may lead to hospitalization or result in fatal outcomes, regardless of admission status.

Length of stay (LOS) is a critical healthcare metric, alongside mortality and readmission rates, for patient care and resource optimization [7]. Prolonged LOS (PLOS) is an adverse outcome that modern predictive models aim to forecast early. PLOS imposes a significant economic and operational strain on global healthcare systems. In Brazil, where public hospitals often operate at near capacity, extended LOS has a direct impact on the Sistema Único de Saúde (SUS), diminishing efficiency and escalating costs related to intensive care, comorbidities, and readmissions [3]. Consequently, precise prediction of PLOS is imperative for optimizing hospital management, predicting resource requirements, and guiding public health policy.

Machine learning (ML) has emerged as a powerful tool for modeling LOS, owing to its capacity to discern intricate, nonlinear relationships among patient demographics, comorbidities, and clinical indicators. Clinical data-driven frameworks exhibit superior predictive performance for PLOS, with tree-based ML models achieving notably high accuracy [7]. However, the inherent lack of transparency in some ML models—the so-called 'black box' issue—has hindered their clinical acceptance. To address this challenge, Explainable Artificial Intelligence (XAI) techniques are increasingly used in healthcare to bolster model transparency and foster trust [17].

Explainable Artificial Intelligence (XAI) methodologies are gaining traction in predicting PLOS risk. Among the most prevalent XAI methods are: (1) SHAP (SHapley Additive exPlanations) [8], which elucidates model outputs by attributing them to feature contributions using cooperative game theory principles; and (2) LIME (Local Interpretable Model-agnostic Explanations) [11], which approximates complex models with simpler, interpretable models on a local scale. Explanations generated through XAI, often via SHAP, pinpoint influential predictors for PLOS risk, thereby aiding clinicians in comprehending the underlying determinants of extended hospital stays.

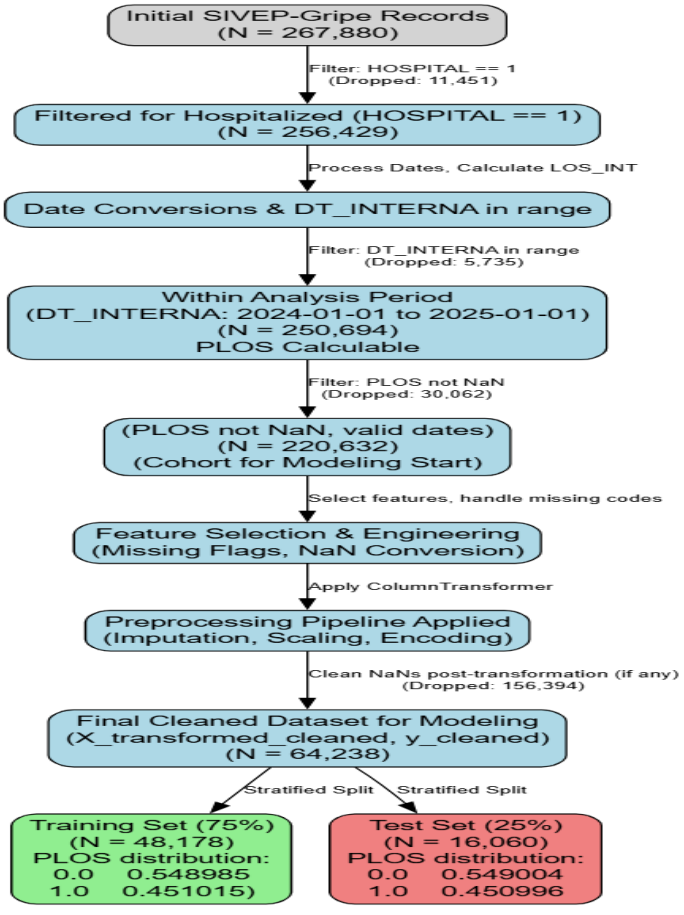
This study builds upon these foundations by developing ML models to predict PLOS (defined as higher than 7 days) in Brazilian SARS patients using admission data, and employing XAI techniques to enhance model interpretability and clinical relevance.

3 MATERIALS AND METHODS

This study's methodology for predicting PLOS followed a sequential data-processing pipeline, as illustrated in Figure 1.

The process started with the acquisition and cleaning of data from the national surveillance system, ensuring data quality and consistency. Subsequently, a feature engineering phase was conducted to select and transform the most relevant variables from the cleaned dataset. The final, engineered dataset then served as the foundation for the modeling and validation stages, where machine learning algorithms were trained and evaluated.

Figure 1 – Data Science Workflow of the PLOS Prediction Pipeline.



Source: Author.

DOI: 10.25286/repa.v11i1.3537

3.1 DATA DESCRIPTION

The study utilized data from the SIVEP-Gripe (Sistema de Informação da Vigilância Epidemiológica da Gripe) system, a Brazilian national surveillance platform for SARS. The dataset was obtained from the official Brazilian Ministry of Health's open data portal.

This comprehensive dataset contains anonymized individual-level records of SARS cases, including influenza and other respiratory viruses, notably COVID-19. Each record encompasses a wide array of variables covering patient demographics, clinical presentation, comorbidities, vaccination history, hospitalization details, diagnostic test results, and case outcomes. The raw data are semi-structured and coded (numerically or with characters), requiring a data dictionary for interpretation.

3.2 DATA PREPARATION

For this study, the initial dataset was filtered to include only patients who were hospitalized due to SARS, identified by the H*OSPITAL* variable (*Value: 1 = Yes*). The analysis focused on admissions occurring between January 1, 2024, and January 1, 2025, based on the *DT_INTERNA* (*admission date*) field. The primary outcome, Prolonged Length of Stay (PLOS), was derived from *DT_INTERNA and DT_EVOLUCA* (outcome date), defined as a hospital stay exceeding 7 days.

Table 1 presents the features selected for predictive modeling, along with their coding according to the SIVEP-Gripe data dictionary. The selection criteria were that only variables available at admission were included, to develop a model for early risk stratification.

3.3 DATA PREPROCESSING

Data were programmatically retrieved and imported into a pandas DataFrame for subsequent analysis, using *Latin-1* encoding and *low_memory=False* to ensure accurate type inference for all columns. Initial preprocessing focused on preparing the date-related fields. Columns representing key dates, including admission date (*DT_INTERNA*) and outcome date (*DT_EVOLUCA*) were converted to datetime objects, with any parsing errors coerced to *Na*T (Not a Time). The study cohort was restricted to patients recorded as hospitalized (*HOSPITAL = 1*).

The primary outcome variable, Length of Stay (LOS), denoted as *LOS_INT*, was calculated as the difference in days between outcome date and admission date. To ensure data integrity, *LOS_INT* was only computed for records where both dates were present and *DT_EVOLUCA* was on or after *DT_INTERNA*; other cases resulted in a *NaN* for *LOS_INT*. For the purpose of predictive modeling, *LOS_INT* was binarized into a target variable PLOS (Prolonged Length of Stay), where *PLOS = 1* if *LOS_INT > 7* days and *PLOS = 0* otherwise. Records with missing *LOS_INT* (and consequently PLOS) were excluded. The analysis was further confined to hospitalizations within the specified date range.

Table 1 - Description of Selected Variables.

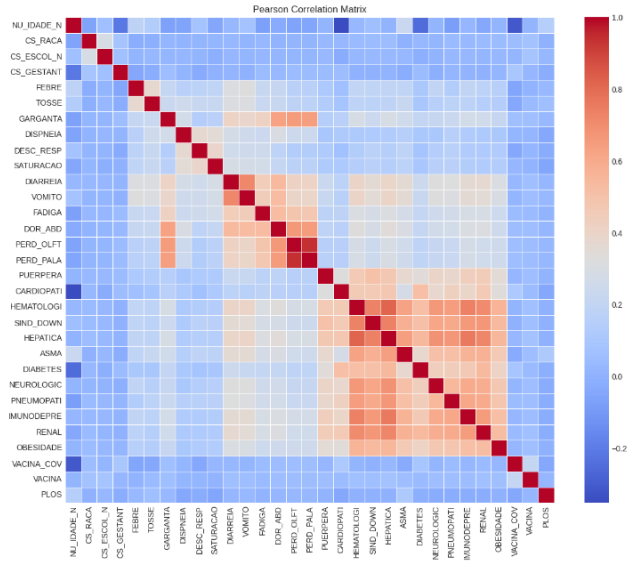
Variable	Code	Description
Demographics and General Information		
CS_SEXO	M, F, I	Sex: Male (M), Female (F), Ignored/Missing (I)
NU_IDADE_N	Numeric	Age in years
CS_RACA	1–5, 9	Race/Ethnicity: 1=White, 2=Black, 3=Asian, 4=Mixed/Brown, 5=Indigenous, 9=Ignored/Missing
CS_ESCOL_N	0–9	Education level: 0: No schooling/Illiterate 1: Incomplete Elementary (1st cycle) 2: Complete Elementary (2st cycle) 3: Complete High School 4: Complete Higher Education 5: Not applicable 9: Ignored/Missing
CS_GESTANT	1–6, 9	Gestational status: 1: 1st Trimester 2: 2nd Trimester 3: 3rd Trimester 4: Gestational age ignored 5: Not pregnant 6: Not applicable (e.g., male) 9: Ignored/Missing
SG_UF_NOT	AA	State of notification (two-letter Brazilian state code)
Clinical Symptoms (1=Yes, 2=No, 9=Ignored)		
FEBRE	1,2,9	Fever
TOSSE	1,2,9	Cough
GARGANTA	1,2,9	Sore throat
DISPNEIA	1,2,9	Dyspnea / Shortness of breath
DESC_RESP	1,2,9	Respiratory distress
SATURACAO	1,2,9	O_2 saturation $< 95\%$
DIARREIA	1,2,9	Diarrhea
VOMITO	1,2,9	Vomiting
FADIGA	1,2,9	Fatigue
DOR_ABD	1,2,9	Abdominal pain
PERD_OLFT	1,2,9	Loss of smell (dropped due to high correlation)
PERD_PALA	1,2,9	Loss of taste
Comorbidities (1=Yes, 2=No, 9=Ignored)		
PUERPERA	1,2,9	Puerperium (postpartum period)
CARDIOPATI	1,2,9	Heart disease
HEMATOLOGI	1,2,9	Hematologic disease (dropped)
SIND_DOWN	1,2,9	Down syndrome
HEPATICA	1,2,9	Liver disease
ASMA	1,2,9	Asthma
DIABETES	1,2,9	Diabetes
NEUROLOGIC	1,2,9	Neurological disease
PNEUMOPATI	1,2,9	Other chronic lung disease
IMUNODEPRE	1,2,9	Immunodeficiency / Immunosuppression
RENAL	1,2,9	Chronic kidney disease
OBESIDADE	1,2,9	Obesity
Vaccination Status (1=Yes, 2=No, 9=Ignored)		
VACINA_COV	1,2, 9	COVID-19 vaccination received
VACINA	1,2, 9	Seasonal influenza vaccination

Source: Author.

A curated set of features relevant to SARS patient demographics, symptoms, comorbidities, and vaccination status was selected for model development. This selection was guided by domain knowledge and common practices in SARS research. Specifically, features included *CS_SEXO* (sex), *NU_IDADE_N* (age in numbers), *CS_RACA* (race/ethnicity), *CS_ESCOL_N* (education level) *CS_GESTANT* (gestational status), symptom indicators (*FEBRE, TOSSE*, etc.), comorbidity flags (*CARDIOPATI, DIABETES*, etc.), and vaccination details (*VACINA_COV, VACINA*). The state of notification (*SG_UF_NOT*) was also included for potential regional analysis.

To handle missing data, we identified codes for '*ignored*' '*not applicable*' or '*unknown*' (e.g., '9' for categoricals) based on the documentation. For original comorbidity columns, a binary flag (*col_missing_flag*) was engineered for each, indicating whether the original value was one of these missing codes. This step was performed before converting the missing codes themselves to np.nan to preserve the information about original missingness. This strategy allows the model to potentially learn patterns from the act of data being missing. After flag creation, the identified missing codes in the original feature columns were replaced with np.nan. Features identified as highly correlated (> 80%) based on preliminary analysis (*PERD_OLFT* and *HEMATOLOGI*) were removed (Figure 2).

Figure 2 - Pearson Correlation Matrix of Selected Features



Source: Author.

A robust preprocessing pipeline was constructed using scikit-learn's *ColumnTransformer*. Numerical features (*NU_IDADE_N*) were imputed using *IterativeImputer* and standardized using StandardScaler. Nominal categorical features (*CS_SEXO, CS_RACA, SG_UF_NOT*) were imputed using the most frequent value and then one-hot encoded.

The ordinal categorical features (*CS_ESCOL_N and CS_GESTANT)* were also imputed with the most frequent value, followed by *OrdinalEncoder* to preserve its inherent order. Binary features (symptoms and vaccine status) were imputed with their most frequent value and then one-hot encoded. The original comorbidity columns (e.g., *CARDIOPATI, DIABETES*), after their specific missing codes were converted to *np.nan*, were imputed using *SimpleImputer (strategy='constant', fill_value=2*), where *'2'* typically represents "*No*" in SIVEP-Gripe coding for comorbidities. This ensures that if a comorbidity status was originally marked as "*Ignored*" it is treated as "*No*" after imputation. The engineered *_missing_flag* columns, being already binary (0 or 1), were passed through without further transformation, assuming they contained no *NaNs* after their creation.

The pipeline was fitted to training data and applied to both sets to avoid leakage. Rows with remaining *NaN* values were removed. The substantial reduction in dataset size after applying the imputation pipeline and subsequent removal of *NaN* values is due to specific features, such as comorbidities, that were not fully resolved by the current imputation strategy.

3.4 MODEL DEVELOPMENT AND EVALUATION

The preprocessed dataset with 64,238 records and 32 independent variables was split into training (75%) and testing (25%) sets, stratified by the PLOS target variable to maintain class proportions in both splits (*PLOS* distribution: 0 = 55%, 1= 45%). A *random_state* was used for reproducibility.

To develop and evaluate machine learning models for predicting Prolonged Length of Stay (*PLOS*), three distinct algorithms were selected for their robustness and common application in healthcare predictive modeling [7]: RandomForest [12], XGBoost [13], and CatBoost [14]. For comparative purposes, a *DummyClassifier* using the *'stratified'* strategy was employed to set a minimum performance benchmark.

DOI: 10.25286/repa.v11i1.3537

Each model underwent hyperparameter tuning using *RandomizedSearchCV* with *5-fold stratified cross-validation*, optimizing for the Area Under the Receiver Operating Characteristic Curve (*AUC*). The hyperparameter search space for each model was intentionally kept concise, with only two candidates explored for each parameter combination. The hyperparameter search space was intentionally restricted (*n_iter=2*) to facilitate a rapid demonstration, acknowledging that a more exhaustive search might yield additional performance improvements.

The performance of the best model (identified by *RandomizedSearchCV*) for each algorithm was then evaluated on the unseen test set. Key metrics, including *accuracy, precision, recall, and AUC,* were calculated. Classification reports and confusion matrices were also generated to provide a comprehensive view of model performance.

3.5 EXPLAINABLE TECHNIQUE

SHAP (SHapley Additive exPlanations) [8] was applied to interpret the top-performing model, providing both global and local feature importance. SHAP values were visualized via plot to identify the most influential predictors. This approach provides insights into the critical drivers of extended hospitalization, delivering actionable information for healthcare practitioners.

4 RESULTS

The following subsections presents the results of the data, models performance and model outcome explainability analyses. The discussion of the obtained results will also be presented.

4.1 DATA ANALYSIS

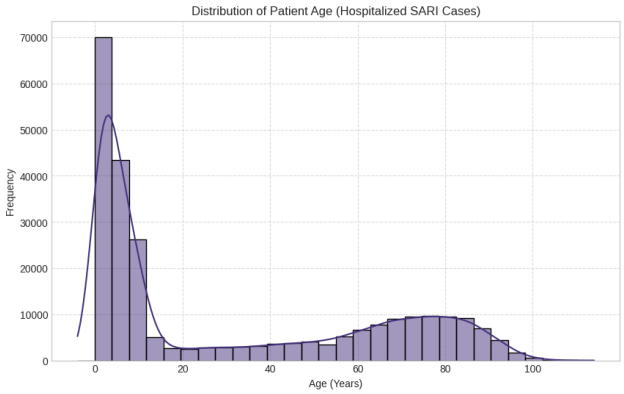
Exploratory data analysis of 2024 hospitalized SARS cases reveals key demographic and clinical characteristics.

A bimodal age distribution highlights high hospitalization rates in young children and older adults (Figure 3). The sex distribution is nearly balanced, with a slight male predominance.

Temporal trends show distinct epidemic waves, likely reflecting influenza and COVID-19 seasonality (Figure 4).

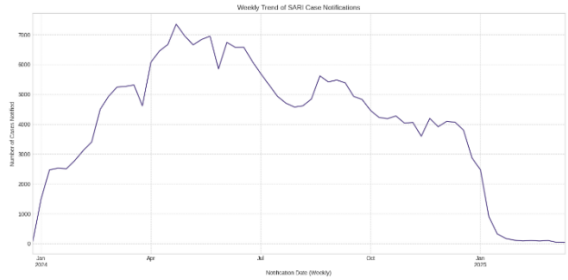
Common symptoms include cough, dyspnea, and respiratory distress, while heart disease, diabetes, and asthma are prevalent comorbidities (Figure 5).

Figure 3 – Age Distribution of Hospitalized Patients.



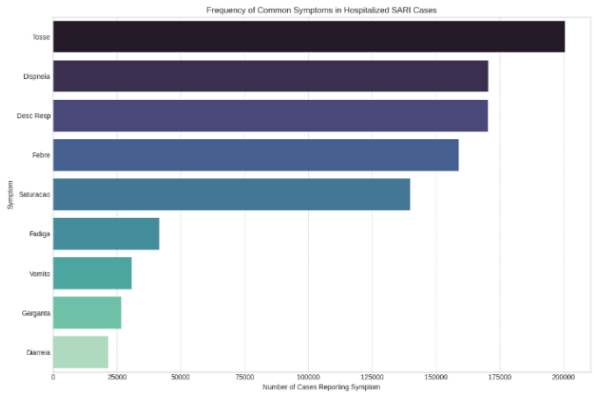
Source: Author.

Figure 4 –Weekly Temporal Trends of Hospitalized SARS.



Source: Author.

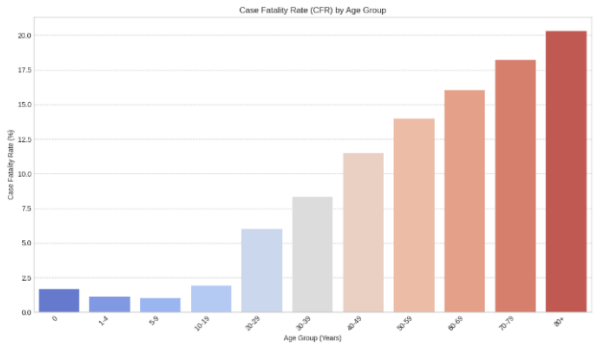
Figure 5 – Frequency of the Most Common Symptoms.



Source: Author.

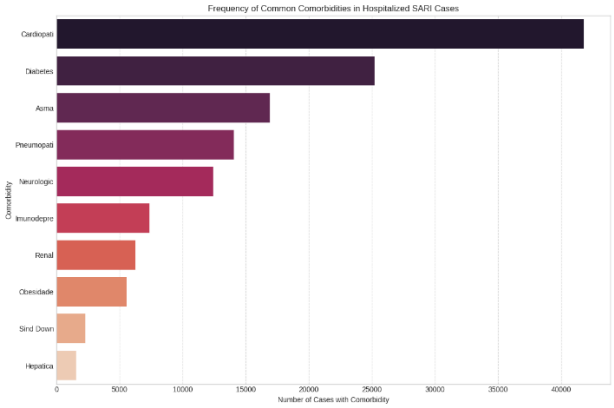
The fatality rate increases sharply with age, particularly affecting the elderly, as shown in Figure 6. Common symptoms include cough, dyspnea, and respiratory distress, and prevalent comorbidities are heart disease, diabetes, and asthma, as illustrated in Figure 7.

Figure 6 – Case Fatality Rate by Age Group.



Source: Author.

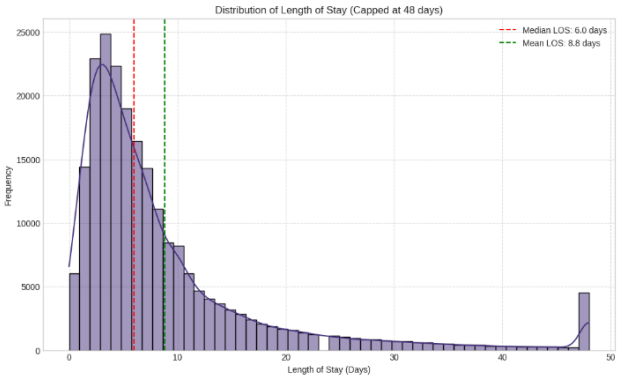
Figure 7 – Prevalence of Common Comorbidities.



Source: Author.

The length of stay (*LOS*) is typically short, with a median of 6 days and mean of 8.8 days, but has a long tail of prolonged hospitalizations, as shown in Figure 8.

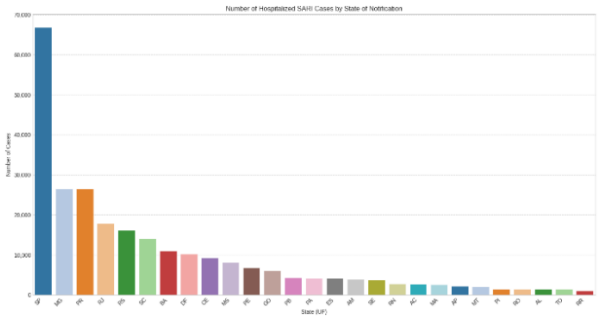
Figure 8 – Distribution of Length of Stay in Days.



Source: Author.

The geographical distribution of cases reflects population density, with the five states reporting the highest case counts being São Paulo, Minas Gerais, Paraná, Rio de Janeiro, and Rio Grande do Sul, as depicted in Figure 9.

Figure 9 – Geographic Distribution of Hospitalized Cases.



Source: Author.

4.2 MODELS EVALUATION

The performance evaluation of the machine learning classifiers was conducted to assess their ability to predict prolonged hospital length of stay (*PLOS*). Multiple complementary metrics—accuracy (*ACC*), *precision* (Prec.), *recall* (Rec.), and the area under the receiver operating characteristic curve (*AUC*)—were employed to capture both overall discrimination and class-specific behavior. These indicators provide a comprehensive view of model reliability, balancing correct predictions, sensitivity to positive (*PLOS*) cases, and robustness against false alarms. Table 2 presents the single set of parameters found for all models by the limited search.

Table 2 - Set of Parameters Models.

MODEL	PARAMETERS
Random Forest	n_estimators=300, min_samples_split=5, min_samples_leaf=4, max_depth=30.
XGBoost	n_estimators=300, max_depth=3, learning_rate=0.1
CatBoost	learning_rate=0.1, iterations=300, depth=8

Source: Author.

Table 3 summarizes the comparative performance of all machine learning models based on *accuracy* (ACC), *precision* (Prec.), *recall* (Rec.), and *AUC* metrics. The baseline *DummyClassifier* achieved random-level performance (*ACC* = 0.50, *AUC* = 0.50). Among the trained models, XGBoost and CatBoost exhibited the highest overall accuracy (*ACC* = 0.61 and 0.60, respectively) and *AUC* values (0.66 for both), slightly outperforming Random Forest (*ACC* = 0.61, *AUC* = 0.65).

DOI: 10.25286/repa.v11i1.3537

Table 3 - Comparative Performance of ML Models.

MODEL	ACC	Prec.	Rec.	AUC
DummyClassifier	0.50	0.45	0.45	0.50
Random Forest	0.61	0.55	0.66	0.65
XGBoost	0.61	0.58	0.49	0.66
Cat Boost	0.60	0.58	0.51	0.66

Source: Author.

In terms of precision, both XGBoost and CatBoost (Prec. = 0.58) surpassed Random Forest (Prec. = 0.55), indicating fewer false positives. However, Random Forest achieved the highest recall (Rec. = 0.66), suggesting superior sensitivity to prolonged length-of-stay cases. Overall, the gradient-boosted models demonstrated balanced and robust predictive performance, while RandomForest favored recall at the expense of precision.

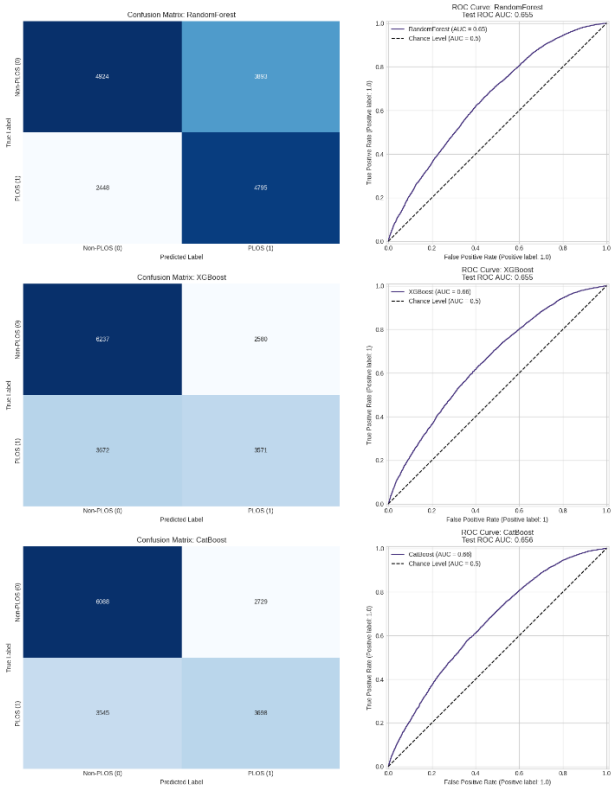
Figure 10 presents the ROC curves and corresponding confusion matrices for the three classifiers—Random Forest, XGBoost, and CatBoost. All models exhibit comparable overall discrimination, with *AUC* values clustering around 0.65. These similar *AUCs* indicate that, under the current feature set and limited hyperparameter tuning, each algorithm learns similarly effective decision boundaries.

The confusion matrices further quantify these trade-offs at the selected classification threshold. Random Forest correctly classifies more true *PLOS* cases (higher *recall*) than the boosting models, reflecting its sensitivity to positive instances. In contrast, XGBoost achieves a higher number of true negatives, indicating better specificity, while CatBoost produces a nearly identical distribution of true and false predictions compared to XGBoost.

Ultimately, model selection depends on the relative costs of misclassification. Random Forest's higher recall makes it preferable when missing a prolonged-stay case is especially costly, such as in clinical triage, whereas XGBoost and CatBoost, with their better precision and slightly higher overall accuracy, are better suited when minimizing false alarms is the priority.

For a balanced compromise between sensitivity and specificity, XGBoost emerges as the most suitable option, offering marginally superior *AUC* and accuracy under these experimental conditions.

Figure 10 – Comparative Performance of ML Models.



Source: Author.

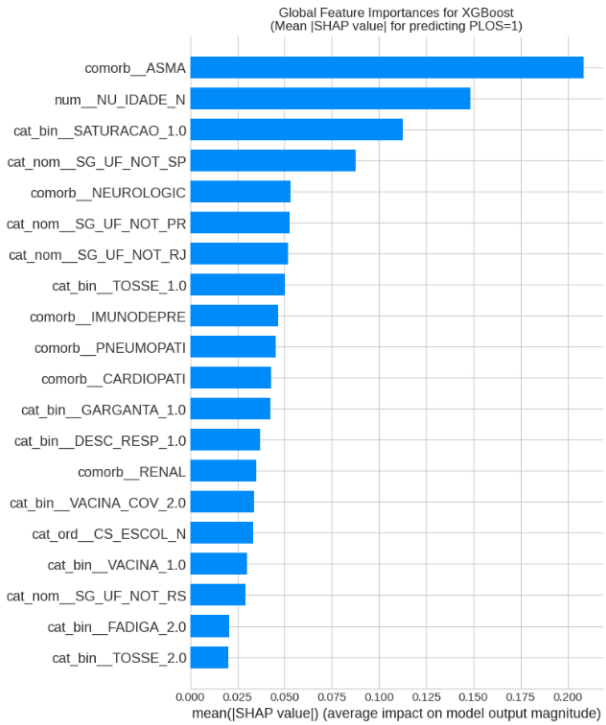
4.3 PREDICTIVE OUTCOMES OF XGBOOST EXPLAINABILITY

To interpret which factors most influenced the XGBoost model's predictions for prolonged hospital length of stay (*PLOS = 1*), global feature importances were analyzed using SHAP values, as illustrated in Figure 11.

The most influential predictor was the presence of asthma comorbidity, underscoring its critical role in the model's decision-making process. Patient age and oxygen saturation levels followed in importance, emphasizing the relevance of both demographic and physiological indicators in early *PLOS* prediction.

Geographical attributes—particularly notification states such as São Paulo and Paraná—also ranked among the top contributors, suggesting that regional healthcare or reporting disparities may influence patient outcomes. Additional relevant features included neurological comorbidities, cough symptoms, and chronic conditions such as immunodeficiency, pneumopathy, and cardiopathy, all of which align with known clinical risk factors for severe respiratory syndromes.

Figure 11 – Feature Importance for the XGBoost.



Source: Author.

In summary, the results show that machine learning models moderately predicted prolonged hospital stays (PLOS) among SARS patients using admission-time data from Brazil's SIVEP-Gripe system. Consistent *AUC* values (around 0.65) across Random Forest, XGBoost, and CatBoost suggest that, despite limited features and minimal tuning, ensemble methods captured clinically meaningful patterns. The XGBoost model offered a balanced trade-off between sensitivity and specificity, while Random Forest prioritized recall, making it particularly suitable for scenarios in which missing prolonged-stay cases could have serious operational implications.

From an interpretability perspective, SHAP analysis highlighted asthma, age, and oxygen saturation as key predictors, reflecting known links between respiratory comorbidities, hypoxemia, and adverse outcomes. The prominence of geographical variables suggests that regional disparities in healthcare delivery and data reporting may affect hospitalization dynamics.

Overall, these results align with previous studies on SARS hospitalization risks and underscore the feasibility of surveillance-based predictive modeling in low- and middle-income countries. However, the moderate predictive performance highlights the need for more comprehensive temporal, clinical, and institutional data. Crucially, explainable AI techniques such as SHAP help bridge algorithmic outputs and clinical reasoning, promoting transparency and supporting integration into healthcare decision-making.

5 LIMITATIONS

This study has limitations: First, the SIVEP-Gripe database may have incomplete data, underreporting, and regional quality variations. Although imputation methods were applied, residual bias from missing or inconsistent entries may persist. Second, the operational definition of prolonged length of stay (PLOS > 7 days) is somewhat arbitrary and may not generalize across clinical settings. Third, the models were restricted to admission-time features, excluding dynamic variables such as treatment interventions, laboratory trajectories, and hospital capacity indicators that could improve predictive accuracy. Fourth, hyperparameter optimization was deliberately constrained for computational efficiency, potentially limiting model performance. Finally, external validity remains uncertain beyond Brazil's 2024 dataset, underscoring the need for cross-temporal and multi-institutional validation.

6. CONCLUSIONS

This study successfully developed and compared three ensemble machine learning models—Random Forest, XGBoost, and CatBoost—for predicting prolonged hospital stays among SARS patients in Brazil using early admission data. All models exhibited comparable, moderate discrimination, with XGBoost achieving marginally superior overall AUC and accuracy, and Random Forest excelling in recall. Explainable AI analysis highlighted key predictors—particularly asthma, age, oxygen saturation, and geographic region—reinforcing their clinical plausibility and interpretability.

These findings highlight the feasibility of leveraging national surveillance data for predictive analytics in healthcare resource management. Despite moderate performance, the models' data-efficient design makes them a solid foundation for scalable early-warning tools in public health. It is important to mention that the proposed pipeline can be extended to other public datasets, offering a flexible and interpretable framework for predictive health analytics.

DOI: 10.25286/repa.v11i1.3537

However, as outlined in the limitations (Section 5), challenges such as data incompleteness in SIVEP-Gripe, the arbitrary PLOS threshold (>7 days), and exclusion of dynamic clinical variables may constrain generalizability. These issues underscore the need for cautious application in diverse settings.

Future research should integrate more sophisticated temporal and clinical features, alongside optimized hyperparameter tuning and prospective validation, to enhance the generalizability of models. Embedding such models into clinical decision support systems and tailoring them to regional and demographic contexts may improve hospital resource allocation and patient outcomes. This could involve integrating electronic health records for dynamic variables or testing within international datasets to broaden applicability.

REFERÊNCIAS

[1] LÓPEZ-CHEDA, A. et al. Estimating lengths-of-stay of hospitalised COVID-19 patients using a non-parametric model: a case study in Galicia (Spain). Epidemiology and Infection, v. 149, p. e102, 2021.

[2] STONE, K. et al. 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.

[3] ROSENTHAL, V. D. et al. The impact of healthcare-associated infections on mortality in ICU: a prospective study in Asia, Africa, Eastern Europe, Latin America, and the Middle East. American Journal of Infection Control, v. 51, n. 6, p. 675-682, 2023.

[4] QI, X. et al. Machine learning-based CT radiomics model for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: a multicenter study. MedRxiv, 2020-02.

[5] QUIROZ-JUÁREZ, M. A. et al. Identification of high-risk COVID-19 patients using machine learning. PLOS ONE, v. 16, n. 9, p. e0257234, 2021.

[6] ZHENG, Y. et al. Clinical characteristics and predictors of delayed discharge among children with SARS-CoV-2 Omicron variant infection. Biomedical Reports, v. 20, n. 2, p. 29, 2023.

[7] BOPCHE, R. et al. In-hospital mortality, readmission, and prolonged length of stay risk prediction leveraging historical electronic patient records. JAMIA Open, v. 7, n. 3, p. ooae074, 2024.

[8] LUNDBERG, Scott M.; LEE, Su-In. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17. Red Hook, NY, USA: Curran Associates Inc, 2017. p. 4768-4777.

[9] MARCOLINO, M. S. et al. Clinical characteristics and outcomes of patients hospitalized with COVID-19 in Brazil: results from the Brazilian COVID-19 registry. International Journal of Infectious Diseases, v. 107, p. 300-310, 2021.

[10] PASSARELLI-ARAUJO, H. et al. Machine learning and comorbidity network analysis for hospitalized patients with COVID-19 in a city in Southern Brazil. Smart Health, v. 26, p. 100323, 2022.

[11] RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos. " Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. p. 1135-1144.

[12] BREIMAN, Leo. Random forests. Machine learning, v. 45, n. 1, p. 5-32, 2001.

[13] CHEN, T; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016. p. 785-794.

[14] PROKHORENKOVA, Liudmila et al. CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, v. 31, 2018.

[15] BECK, Jéssica Luíza et al. Clinical and sociodemographic aspects of cases of severe acute respiratory syndrome in southern Brazil. Revista de Epidemiologia e Controle de Infecção, v. 13, n. 3, p. 150-157, 2023.

[16] HURISA DADI, Habtamu; HABTE, Netsanet; MULU, Yenework. Length of hospital stay and associated factors among adult surgical patients admitted to surgical wards in Amhara Regional State Comprehensive Specialized Hospitals, Ethiopia. PLoS One, v. 19, n. 8, p. e0296143, 2024.

[17] ABBAS, Qaiser; JEONG, Woonyoung; LEE, Seung Won. Explainable AI in Clinical Decision Support Systems: A Meta-Analysis of Methods, Applications, and Usability Challenges. In: Healthcare. MDPI, 2025. p. 2154.