Big Data-Driven Prediction of ICU Length of Stay

Large Scale Data Science

Francisco Tavares 202205119

Rodrigo Batista 202206259

Rodrigo Taveira 202206425

Introduction

In critical care environments such as Intensive Care Units (ICUs), the ability to estimate how long a patient will remain hospitalized—the Length of Stay (LOS)—is essential for managing hospital resources, reducing costs, and improving patient outcomes. Accurate LOS predictions can aid in bed planning, staffing, and treatment decisions, particularly in high-demand environments like the Medical ICU (MICU).

The MIMIC-III database is one of the most comprehensive publicly available datasets for ICU research. It includes detailed patient records, covering vital signs, medications, diagnoses, and administrative information. For this project, we focus on a subset of the data corresponding to MICU patients and use structured features that are commonly available early during admission, such as age, gender, ethnicity, insurance, and timestamps related to admission and discharge.

The main objective of this project is to develop and evaluate predictive models for LOS using scalable data science tools that can handle large volumes of healthcare data. This includes the use of Google BigQuery for querying and filtering data, BigFrames for transformations, and distributed processing frameworks like PySpark and Dask for modeling.

Overview of the MIMIC-III Dataset

The **Medical Information Mart for Intensive Care III (MIMIC-III)** is a large, publicly available database that contains de-identified health-related data associated with over 40,000 critical care patients admitted to the Beth Israel Deaconess Medical Center in Boston between 2001 and 2012. The dataset includes rich and detailed information such as patient demographics, vital signs measured at the bedside, laboratory test results, procedures, medications, caregiver notes, and mortality outcomes both within and outside the hospital.

MIMIC-III is widely used in academic research due to its comprehensive nature and its potential to support studies in clinical decision support, predictive modeling, and health outcomes analysis.

Dataset Description

This study leverages five core tables from the MIMIC-III database to construct the cohort and extract relevant features for LOS prediction: `PATIENTS`, `ADMISSIONS`, `ICUSTAYS`, `CHARTEVENTS`, and `DIAGNOSES`. Each table contributes specific types of information necessary for patient identification, demographic profiling, administrative event tracking, clinical measurement extraction, and diagnostic classification

- **CHARTEVENTS**: This is the largest and most granular table, recording time-stamped physiological measurements and clinical observations made during ICU stays. Each row includes a unique identifier (`ITEMID`), measurement time (`CHARTTIME`), and value (`VALUENUM`, `VALUEUOM`). This table is the primary source of vital signs used for predictive modeling, such as heart rate, blood pressure, respiratory rate, and oxygen saturation. Observations were harmonized across EHR systems (CareVue and MetaVision) using `ITEMID` mappings and aggregated into hourly intervals.

- **PATIENTS**: Provides static demographic data for each patient, including biological sex (`GENDER`), date of birth (`DOB`), and vital status (`EXPIRE_FLAG`, `DOD`). This table is essential for calculating age and filtering out pediatric cases or patients with incomplete demographic records.

- **ADMISSIONS**: Contains administrative information related to each hospital admission, such as admission and discharge timestamps (`ADMITTIME`, `DISCHTIME`), admission type (`ADMISSION_TYPE`), insurance category (`INSURANCE`), and social background variables (e.g., `RELIGION`, `MARITAL_STATUS`, `ETHNICITY`). These attributes help contextualize the clinical episode and serve as predictors of hospital resource use.

- **ICUSTAYS**: Focuses on the ICU segment of each hospital stay. It defines the ICU admission and discharge times (`INTIME`, `OUTTIME`) and the unit of care (e.g., `FIRST_CAREUNIT`). This table is used to compute ICU-specific LOS and to ensure temporal alignment between clinical events and ICU episodes.

- **DIAGNOSES**: Lists ICD-9 diagnostic codes associated with each hospital admission. This table enables grouping patients into clinically meaningful disease categories (e.g., circulatory, respiratory, endocrine), which were used for subgroup analyses and cohort segmentation.

Together, these tables support a comprehensive modeling pipeline that integrates demographic context, admission characteristics, diagnostic categories, and real-time physiological measurements to predict ICU Length of Stay.

Performance Testing with CHARTEVENTS Data

To assess the feasibility of working with large volumes of ICU time series data, we tested the ingestion and basic manipulation of one million rows from the CHARTEVENTS table using several scalable tools: BigQuery, Dask, and PySpark.

The raw file (~4 GB) was accessed from Google Cloud Storage using `gsutil`, and decompressed locally using Python's built-in `gzip` module. Initially, a sample of one million rows was read into a Pandas DataFrame and exported to a CSV for subsequent use.

To validate scalability and integration with distributed systems, we loaded the same sample using:

- **Dask**, which allowed efficient chunk-based reading and lazy computation;

- **PySpark**, where we created a Spark DataFrame and executed queries to count rows and null values;

- **BigQuery** (partially attempted), with integration via the Spark BigQuery Connector.

This process confirmed the readiness of these tools to handle MIMIC-III data at scale. The Spark DataFrame correctly recognized the structure (15 columns, 1 million rows), and no missing values were detected in the sample. Null inspection using `isNull()` and `isnan()` functions showed clean data for the selected subset.

These experiments support the robustness of cloud-based and distributed pipelines for future extensions of the project, particularly when working with the full CHARTEVENTS table, which contains over 330 million rows.

Why do we use Group Diseases?

To better understand the clinical and demographic characteristics of ICU patients, we chose to group diagnoses into broader disease categories based on the ICD-9 classification system. ICD-9 codes are structured such that the first three digits represent a general category of disease. For example, codes in the range 390–459 correspond to circulatory system diseases, while codes in the range 460–519 correspond to respiratory system diseases. Grouping diagnoses by these categories allows us to move beyond fragmented, code-level analysis and instead focus on patterns that are clinically interpretable and statistically meaningful.

Using disease groups offers several advantages. First, it reduces the dimensionality of the data by collapsing thousands of specific diagnostic codes into a smaller number of high-level categories. This simplification makes it easier to explore trends and draw conclusions without being overwhelmed by granular detail. Second, these groups align more closely with how clinicians think about patient conditions—by system or domain (e.g., cardiovascular, respiratory, endocrine)—which enhances the clinical relevance of the analysis. Third, focusing on disease groups allows for more robust aggregation and comparison of outcomes, such as length of stay (LOS) and mortality rates, across different types of conditions.

To guide our analysis, we first computed the frequency of admissions associated with each ICD-9 disease group. We then selected the three most prevalent groups in our dataset to explore in greater detail. These were:

1. Circulatory System Diseases (ICD-9: 390–459) – e.g., heart failure, arrhythmias, hypertension,etc.
2. Respiratory System Diseases (ICD-9: 460–519) – e.g., pneumonia, respiratory failure, etc.
3. Endocrine, Nutritional and Immunological Disorders (ICD-9: 240–279) – e.g., diabetes, malnutrition, metabolic disorders,ect.

By focusing our exploratory analysis on these three groups, we aim to gain deeper insights into patient characteristics, treatment patterns, and factors influencing hospital length of stay across different clinical domains.

Disease Grouping Datasets

For subsequent analysis and model development, we created four separate datasets based on ICD-9 disease groupings:

- One dataset including only patients diagnosed with **Endocrine, Nutritional, and Immunological Disorders**;
- One with patients affected by **Circulatory System Diseases**;
- One with patients presenting **Respiratory Diseases**;
- And a combined dataset aggregating patients from all three groups.

This segmentation enables us to analyze LOS patterns and model performance within clinically homogeneous subpopulations, while also supporting comparisons with a merged, more heterogeneous cohort.

Exploratory Data Analysis

Before training predictive models, we conducted an exploratory data analysis (EDA) to better understand the distribution, quality, and variability of the key features involved in ICU Length of Stay (LOS). This process helps identify trends, detect anomalies, and assess the representativeness of the dataset. We analyzed both demographic and clinical attributes, paying special attention to outliers, missing values, and group-level patterns related to diagnostic categories. The following visualizations and summaries guided decisions on data filtering, feature engineering, and model design.

Cohort Selection

To ensure consistency with clinical standards and maintain data quality, we applied several filters to define our final cohort

- We included an age restriction, which considers only patients aged over 15 years. This excludes pediatric cases, which typically follow different clinical protocols and exhibit distinct patterns of care and length of stay (LOS).

- We restricted ICU stays to a duration between 12 hours and 10 days. Very short stays (<12h) may reflect pre-admission administrative errors or early transfers, while very long stays (>10 days) often correspond to complex or exceptional clinical cases. These extreme values tend to introduce noise and distort the learning process, especially in regression models.

- For patients with multiple ICU admissions, we retained only the first ICU stay. This choice is motivated by two factors:
    - Most patients in the MIMIC-III dataset are admitted only once.
    - Considering only the first admission avoids data leakage between multiple episodes for the same patient and aligns with methodologies adopted in previous studies using MIMIC.

These criteria resulted in a cohort that is both clinically meaningful and statistically reliable, allowing our models to focus on representative and generalizable ICU episodes.

Feature Selection

We began by selecting a set of features that are clinically relevant for predicting LOS. These included demographic, administrative, and physiological variables. The following features were chosen based on prior literature and exploratory analysis:

The selected attributes include:

- **ethnicity** – the patient's ethnic background

- **religion** – the patient's religious affiliation (if recorded)

- **insurance** – type of insurance coverage (e.g., Medicare, private)

- **admission_type** – type of hospital admission (e.g., emergency)

- **admittime** – hospital admission timestamp

- **dischtime** – hospital discharge timestamp

Additionally, we included **gender**, extracted from the PATIENTS table.

These attributes were chosen based on their widespread availability across patient records, clinical and epidemiological relevance, and support in the literature demonstrating their value in predicting Length of Stay (LOS). All variables were appended to the previously filtered tables, where patients were already restricted by age, LOS range, and first ICU admission only.

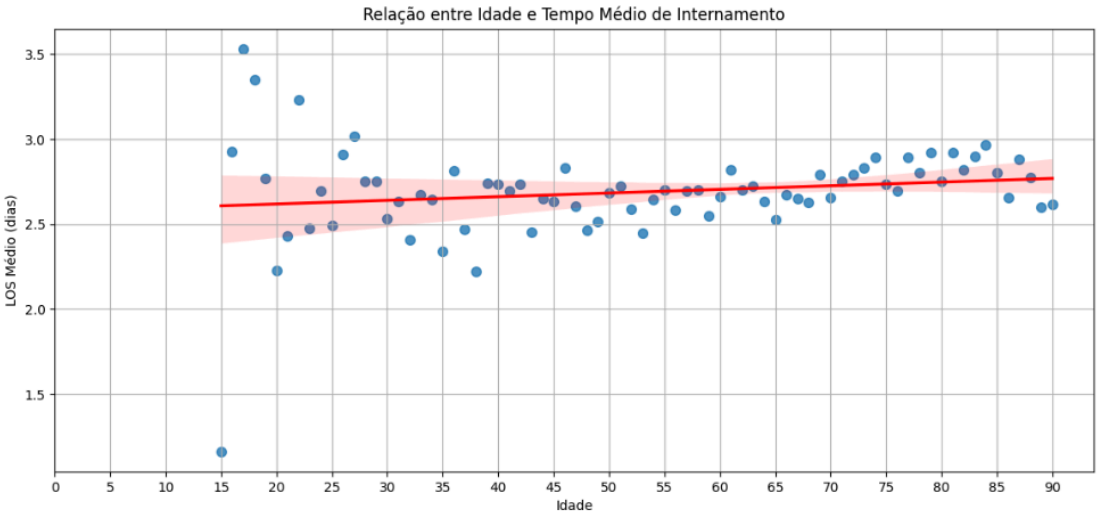
In addition to selecting relevant features, we also excluded certain columns that do not provide predictive value or could introduce data leakage.

Analysis of the selected features

Age

Age is a fundamental clinical variable that directly influences patients' health status, the complexity of the required care, and consequently, the length of hospital stay (LOS – Length of Stay). Older patients tend to have comorbidities, greater frailty, and a higher likelihood of complications, which can lead to longer hospitalizations.

To ensure privacy, MIMIC-III masks patient ages above 89 by assigning them a placeholder value such as 300. To maintain consistency and respect the dataset's anonymization protocol, we replaced all ages ≥ 299 with a fixed value of 90 years. This standardization allows us to preserve the age distribution's interpretability while complying with the data's de-identification policy.



Using the presented charts, we analyzed both the age distribution within the dataset and its relationship with the average LOS. A scatter plot with a regression line was used to observe the general trend of LOS across different ages.

This analysis helps determine whether the variable age significantly contributes to explaining variations in LOS and supports its inclusion in predictive models. Additionally, it allows for the identification of potential age groups with distinct hospitalization patterns, such as very young or very old patients.

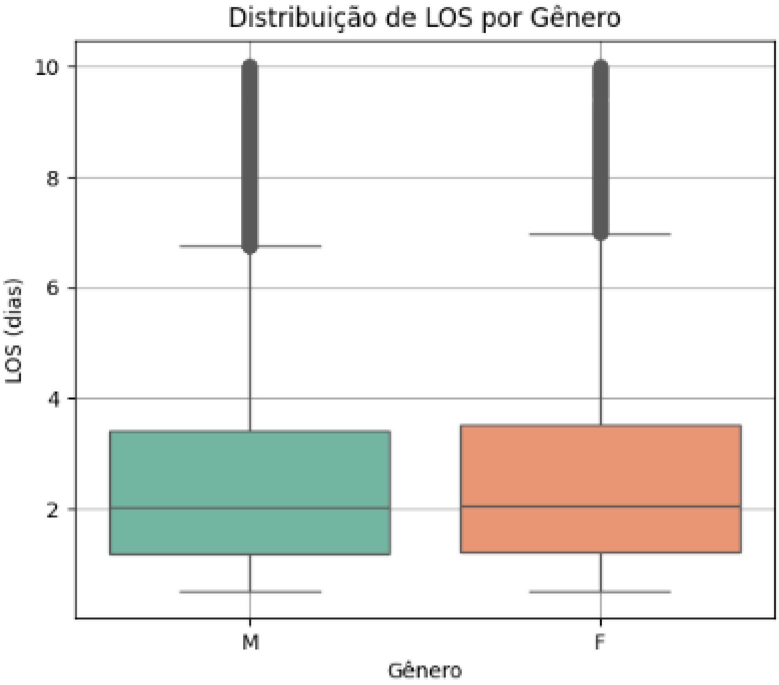
The trend line shows a slight upward slope, indicating that the average length of stay tends to gradually increase with age. Despite data dispersion, especially among younger ages, LOS remains relatively stable in the 2.5 to 3-day range.

Variability is higher among younger patients (under 25 years old), likely due to a smaller number of patients or more diverse clinical conditions in this group. From age 60 onward, the average LOS tends to stabilize or increase slightly, which aligns with the greater clinical complexity typically observed in elderly patients.

Gender

The patient's gender can influence the length of hospital stay due to various factors, including physiological differences, the prevalence of specific diseases, or even variations in clinical treatment between men and women.

Using the presented charts, I assess the distribution of patients by gender and examine whether there are significant differences in the average length of stay (LOS) between the groups. This analysis helps determine whether the gender variable can add value to the predictive model and whether it should be included as an explanatory factor.



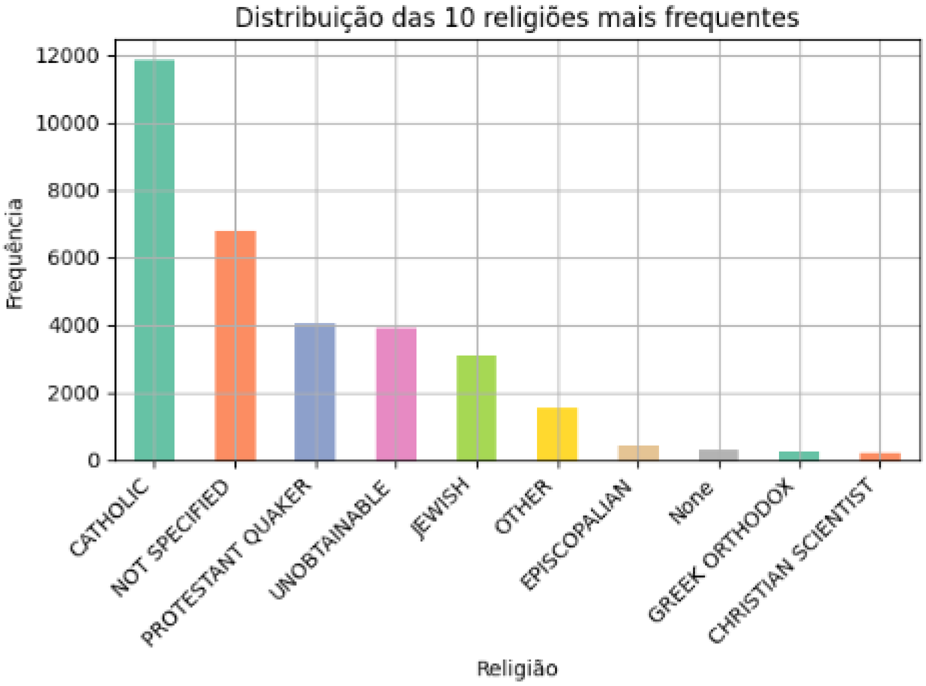
Moreover, this variable is binary and well represented in the dataset, which facilitates its interpretation and integration into machine learning models.

Religion

The patient's religious affiliation can, in some clinical contexts, influence healthcare practices, treatment decisions, and even the length of hospital stay. For example, certain beliefs may affect the acceptance of medical procedures, blood transfusions, or palliative care, which can impact the duration of hospitalization.

Using the presented charts, we analyze the distribution of patients by religion and examine whether different groups show variations in the average length of stay (LOS). This analysis helps assess whether religion could be a relevant variable in the predictive model, aiding in the identification of patterns and potential cultural or institutional biases.

Although it is a sensitive variable that must be handled with ethical care, its analysis is useful for both validating data quality and understanding possible contextual factors that affect healthcare.

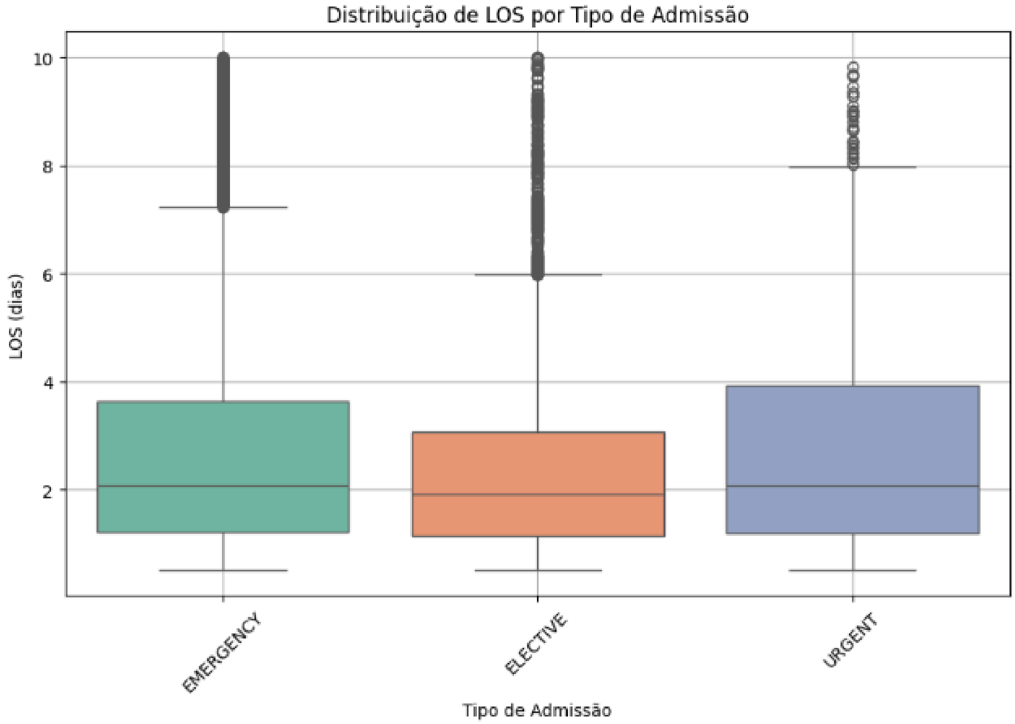


Admission Type

The admission_type variable indicates the type of hospital admission for the patient, which may include categories such as emergency, urgent, elective, among others. This information is clinically relevant, as different admission types reflect varying degrees of severity and planning of the hospitalization.

Using the presented charts, we analyze the frequency of each admission type and its relationship with the average length of stay (LOS). Emergency admissions, for example, are often associated with more severe and unpredictable clinical conditions, which can lead to longer hospital stays.

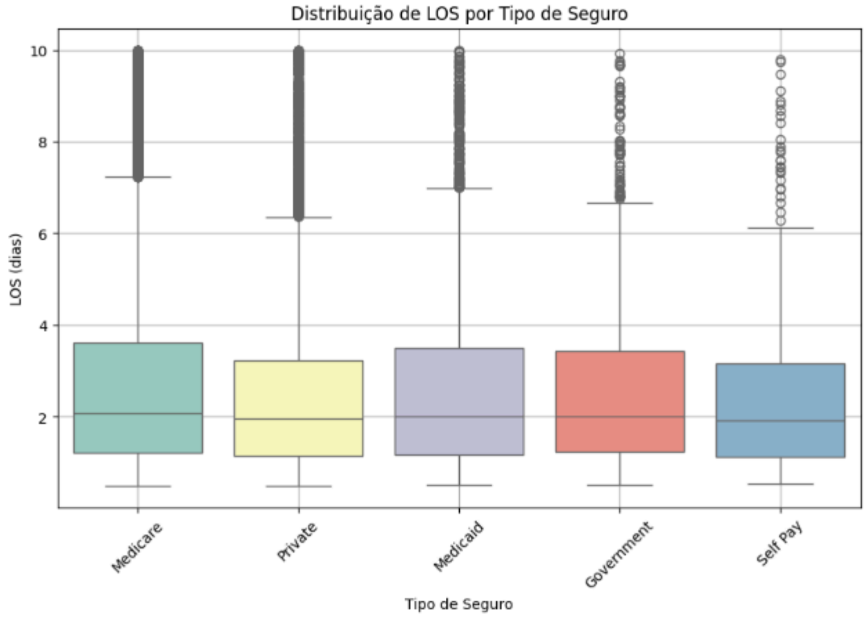
This analysis is important to understand the impact of the admission context on the duration of hospital care and to justify the inclusion of this variable in predictive models.



Insurance

This chart shows the distribution of different insurance types in the dataset. It is important to understand the representation of each class before using this variable in a predictive model, especially if it is imbalanced — as this may affect the model's robustness.

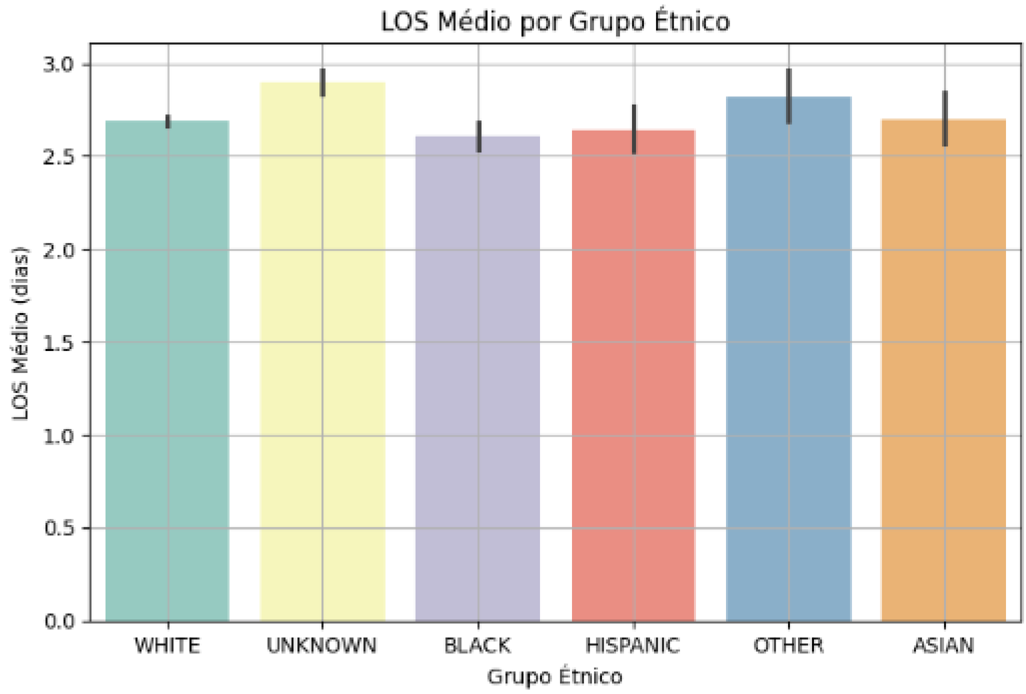
The second chart relates the type of insurance to the average length of stay (LOS), helping to identify whether there is a relevant association between these two variables. This relationship may suggest that the type of insurance influences the duration of hospitalization, making it a potentially useful feature for the model.



Ethnicity

The ethnicity variable represents the declared or assigned ethnic background of patients at the time of hospital admission. Although not directly related to physiological factors, ethnicity may reflect social, cultural, and healthcare access differences that can potentially influence the length of hospital stay (LOS).

Using the presented charts, we analyze the frequency of different ethnic groups in the dataset and examine whether there are differences in average LOS among them. This analysis is also important for identifying potential biases in the data and ensuring a fair and representative approach in predictive models.



When used responsibly, the ethnicity variable can help reveal relevant patterns and contribute to building more comprehensive models that are aware of the patients' social context.

Given the high number of specific categories in the ethnicity variable, we decided to group ethnicities into broader categories to simplify analysis and improve data visualization. This approach helps highlight clearer trends and avoids distortions caused by underrepresented classes.

Data Analysis Using Dask and PySpark

Dask

Dask was primarily used for exploratory data analysis and visualization, enabling the handling of datasets larger than available memory and facilitating parallel operations. With Dask, it was possible to generate density plots and histograms for clinical variables such as heart rate, temperature, and oxygen saturation. These visualizations provided better insight into data distribution and helped identify patterns, outliers, and potential anomalies, contributing to more informed decisions in the preprocessing and analysis of clinical data.

Integration of Vital Signs

For this study, we selected key clinical variables from the *CHARTEVENTS* table, namely heart rate, respiratory rate, temperature, oxygen saturation (SpO_2), and systolic and diastolic blood pressure.

These vital signs are widely recognized as early indicators of physiological deterioration in intensive care settings and show high coverage across patients in the MIMIC-III dataset, making them reliable features for analysis.

To ensure compatibility between data recorded in the two hospital systems used during the MIMIC-III data collection — CareVue (prior to 2008) and MetaVision (post-2008) — we included multiple `itemid` versions per variable. For example:

- `220045` → Heart Rate (bpm)
- `220277` → Oxygen Saturation (SpO_2, %)
- `220050` → Systolic Blood Pressure
- `223761` → Temperature (°C)

We extracted the corresponding measurements using SQL queries that applied `MAX(IF(...))` logic to retrieve the appropriate value regardless of the system version.

Furthermore, in alignment with the methodology described in MIMIC-Extract, we aggregated the measurements into 1-hour intervals to create a regular time series representation.

Heart Rate

This chart compares the distribution of heart rate measurements recorded by two sets of sensors: the more recent values (heart_rate) and the older ones (heart_rate_old). We used a density plot to better visualize the shape of the distribution, regardless of the total number of observations.

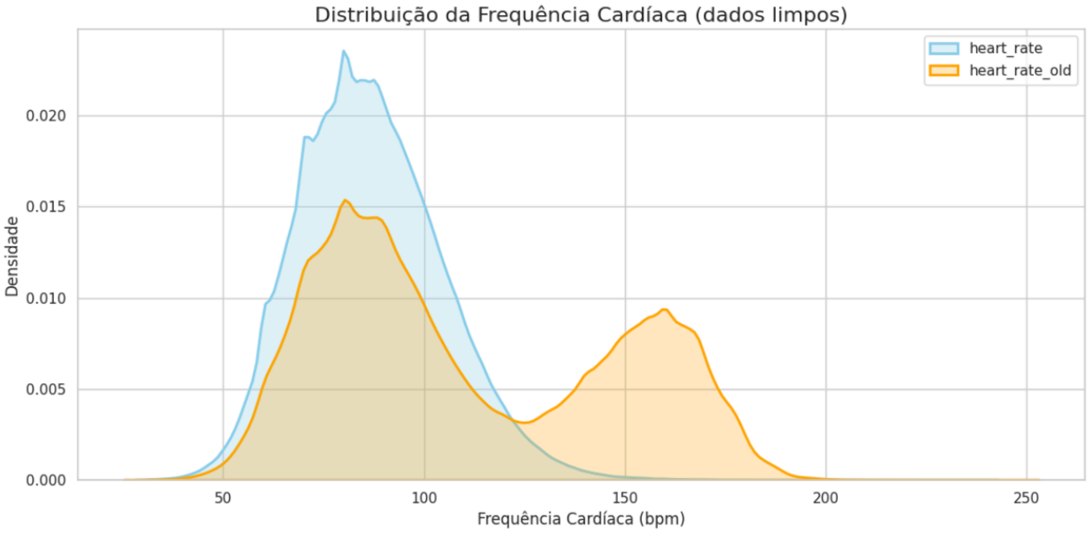
The density represents the relative proportion of values along the heart rate axis (beats per minute), allowing for distribution comparison even with different sample sizes.

The following observations can be made:

- The heart_rate distribution shows a single peak centered between 70–90 bpm, which aligns with normal resting physiological values.

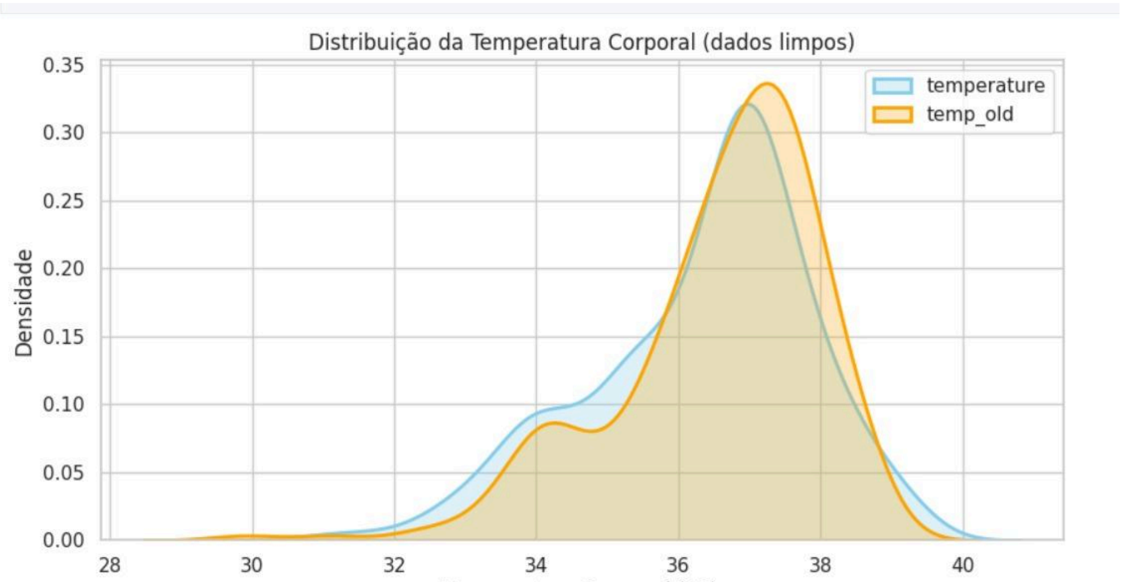
- In contrast, heart_rate_old displays a bimodal distribution, with a second peak above 140 bpm, suggesting possible inconsistencies in the older data, differences in measurement methods, or a higher number of cases with tachycardia.

This type of analysis is useful for validating data quality and understanding how monitoring systems have evolved over time.



Temperature

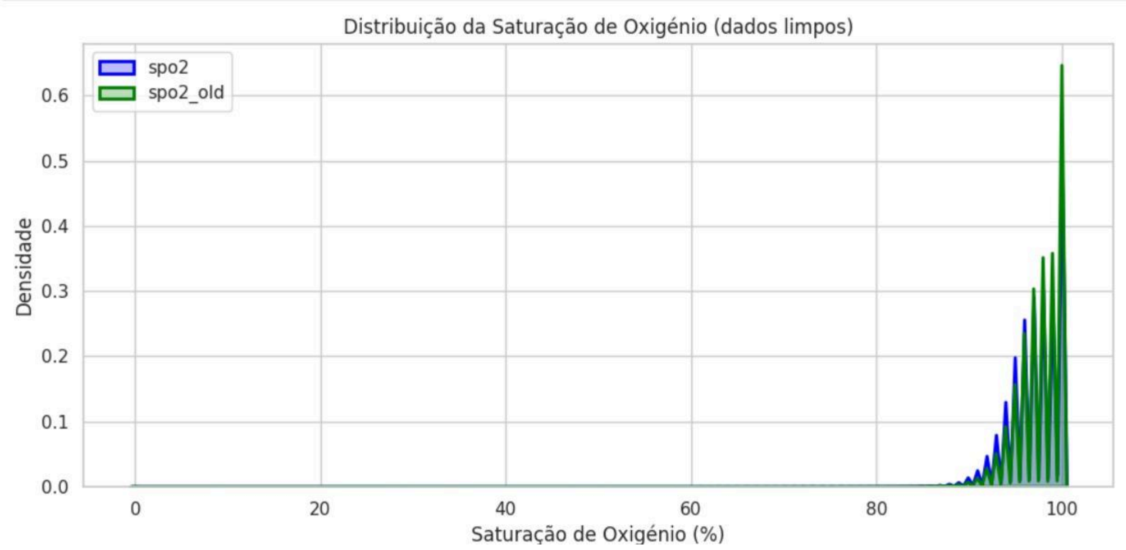
This chart compares the distribution of body temperature recorded by two types of itemid: a more recent one (temperature) and an older one (temp_old). Both curves show a unimodal distribution centered within the normal human temperature range (approximately 36.5 °C to 37.5 °C), which aligns with expected physiological values.



The data were previously filtered to exclude anomalous values (<25 °C or >45 °C), ensuring a more reliable analysis. The similarity between the curves indicates consistency between the two types of records, although the temperature curve appears slightly more dispersed, possibly reflecting greater sensitivity or a larger volume of recent data.

This chart is useful for validating data quality and identifying potential discrepancies between older and more recent data sources.

Oxygen Saturation



This chart shows the distribution of oxygen saturation measured using the variables `spo2` (more recent version) and `spo2_old` (older version). The density analysis reveals a high concentration of values between 95% and 100%, with a sharp peak at 100%, which is expected in stable patients.

This behavior is typical in clinical data:

- Most hospitalized patients maintain near-normal oxygen saturation levels due to ventilatory support.

- The "stepped" shape of the curves suggests that the values are recorded as discrete integers rather than continuous values.

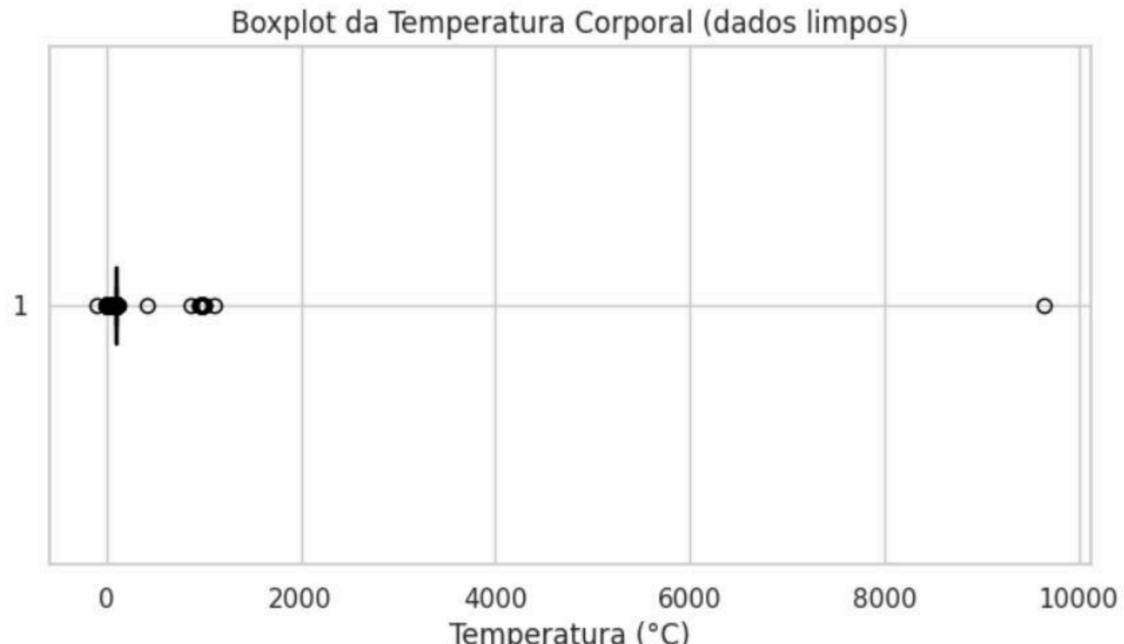
- The strong density on the far right reinforces the idea that values are highly concentrated around physiologically normal or ideal limits.

This visualization confirms that both fields (`spo2` and `spo2_old`) capture measurements consistent with clinical practice, although with slight differences in distribution.

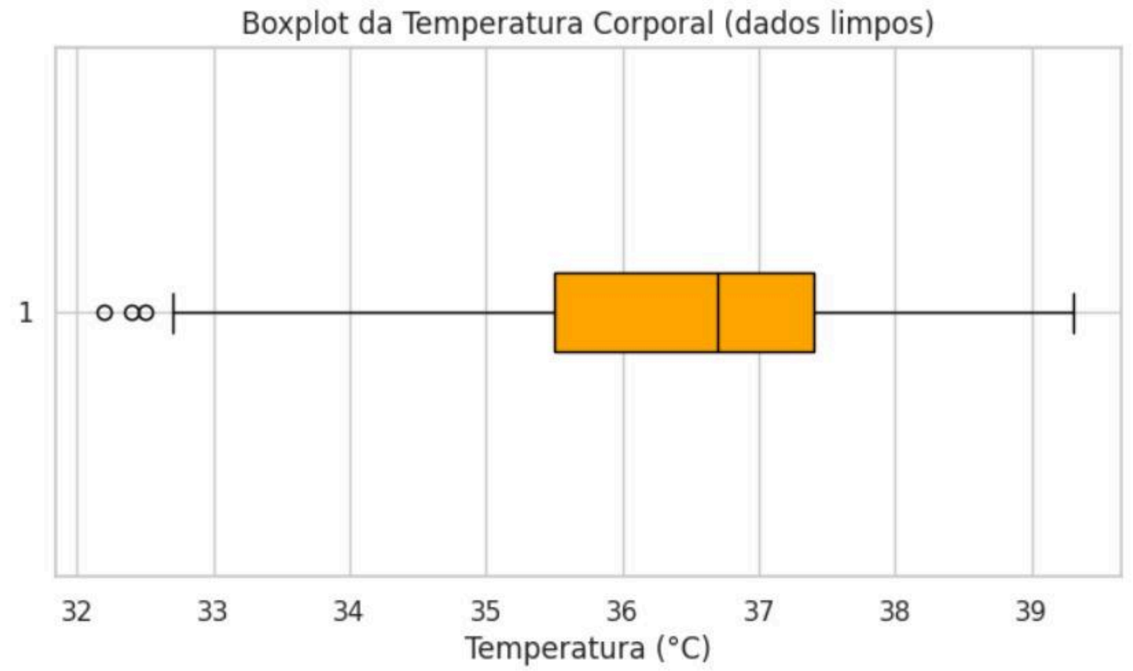
PySpark

Apache Spark was used as the main tool for data preprocessing and exploratory analysis due to its ability to process large volumes of information in a distributed and efficient manner. With Spark, it was possible to identify and handle missing values, apply techniques to remove or replace outliers based on percentile intervals or defined thresholds, and conduct a detailed exploration of relevant clinical variables such as heart rate, body temperature, and oxygen saturation. This robust processing helped improve the quality and reliability of the data used in subsequent analyses.

The body temperature data initially contained values outside the acceptable range for a human being, such as temperatures exceeding 9000 °C. These values are considered severe outliers and indicate recording errors or sensor failures.



After data cleaning, it was possible to remove these extreme values and retain only observations within a physiologically plausible range. This cleaning process made the data more realistic and suitable for further analysis, allowing for more reliable conclusions and reducing the impact of anomalies.



Preprocessing

Handling Missing Values

To address the issue of missing values in the clinical time series, we adopted a multi-step imputation strategy inspired by *Che et al. (2018)*, which demonstrated the effectiveness of this approach in medical time series with frequent missingness.

First, we applied **forward filling**, where each missing value is filled using the most recent previous observation for the same patient and variable. This technique preserves temporal coherence and is especially useful in ICU data, where measurements can be irregularly spaced.

If no previous value exists for a variable in a patient's time series, we then fill the missing entry using that patient's **individual-specific mean** for the same variable.

Finally, if a variable was never observed for a given patient, the missing value is filled with the **global mean** computed across the training set.

This hierarchical imputation strategy is summarized by Che et al. as follows:

“Variable values are first forward filled and then set to individual-specific mean if there are no previous values. If the variable is never observed for a patient, its value is set to training set global mean.”

Feature Extraction Window

To simulate realistic early-stage clinical decision-making, we restricted our feature extraction to the first 24 hours of each ICU stay. This temporal window is commonly used in prognostic modeling and aligns with the clinical imperative to make accurate predictions shortly after admission—when interventions are most actionable and impactful.

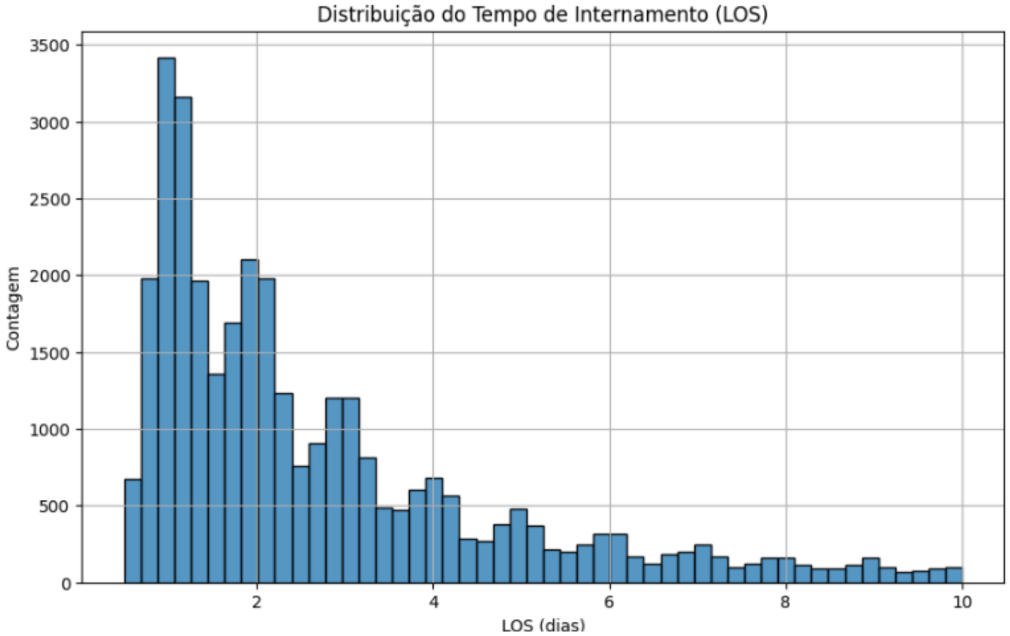
By limiting input variables to data collected within the initial 24 hours, we ensure that all predictive signals are available early in the patient's trajectory, making our models suitable for prospective deployment. Furthermore, this approach reduces data leakage from later clinical events and reflects real-world constraints where complete patient histories are not always accessible at the time of decision-making.

This choice also helps standardize input lengths across patients, allowing fairer comparison and more stable model training.

Target Variable

The target variable for the model is **LOS**, which represents the number of days between hospital admission (ADMITTIME) and discharge (DISCHTIME). This was calculated in advance and cleaned to remove outliers (e.g., patients with LOS exceeding 120 days).

The LOS was computed as the difference between discharge and admission timestamps and then converted to hours. This transformation facilitates finer-grained analysis and visualizations, especially when comparing short-term versus long-term stays.



Modelling

The primary objective of this phase was to predict the Length of Stay (LOS) for ICU patients using demographic, administrative, and early clinical features extracted from the MIMIC-III database. This task was framed as a supervised regression problem, with LOS (in hours) as the continuous target variable.

To address this, we implemented a machine learning pipeline based on the Random Forest Regressor, a non-parametric, ensemble-based algorithm that is well-suited to clinical data. It handles heterogeneous features and non-linear relationships effectively, requires minimal hyperparameter tuning, and offers built-in robustness to outliers. Additionally, Random Forests provide intrinsic feature importance scores, making the model interpretable and practical for real-world healthcare use.

We began by applying a standard 80/20 train-test split on all four datasets (endocrine, circulatory, respiratory, and full cohort), using Dask to efficiently manage large dataframes in parallel and to reduce memory usage. This allowed us to scale the workflow without compromising on performance or hitting resource limits.

We then performed a more thorough evaluation using 5-fold cross-validation, comparing the results with and without Dask to assess its impact on performance. While the predictive metrics remained relatively similar, we observed a notable reduction in execution time when using Dask, demonstrating its benefit in large-scale data processing.

Cross Validation

To obtain a robust and unbiased estimate of the model's generalization ability, we used 5-Fold Cross-Validation (CV). This technique splits the dataset into five equally sized partitions (four folds are used for training and the remaining fold is used for testing)

This process is repeated five times, ensuring each sample serves as a test instance once.

Evaluation Metrics

We evaluated model performance using the following standard regression metrics:

- Mean Absolute Error (MAE): Average magnitude of prediction errors, less sensitive to outliers.

- Root Mean Squared Error (RMSE): Penalizes large errors more heavily, reflecting error dispersion.

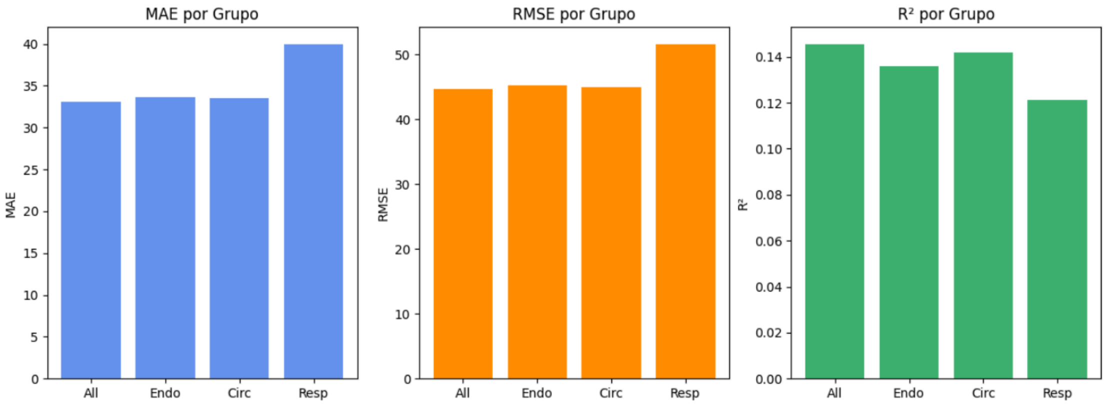
- R² Score (Coefficient of Determination): Indicates the proportion of variance in LOS explained by the model. Values near 1.0 reflect strong predictive power.

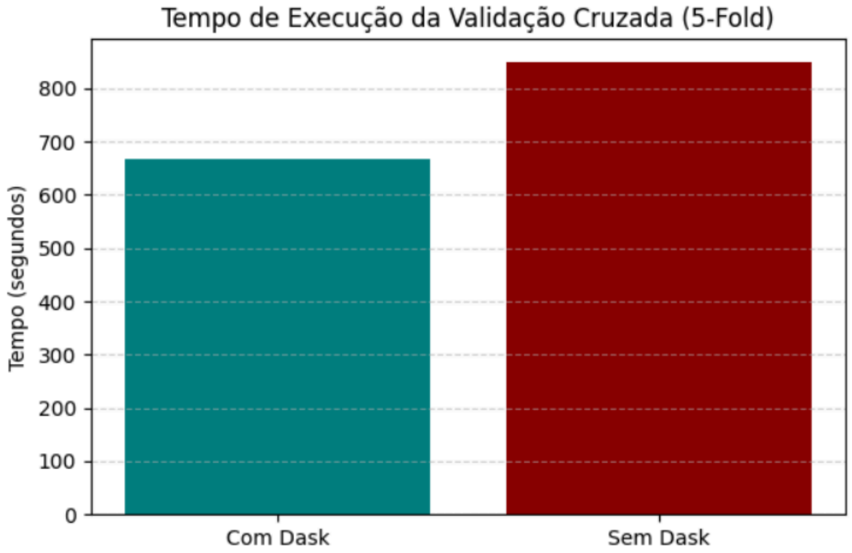
All metrics were computed for each fold and reported as mean ± standard deviation, allowing us to quantify both the accuracy and consistency of the model across different data splits

Results

The figures below present the performance of the Random Forest Regressor across four patient groups: the full dataset ("All"), endocrine diseases ("Endo"), circulatory diseases ("Circ"), and respiratory diseases ("Resp").

Additionally, we compare the execution time of 5-fold cross-validation with and without Dask. Dask significantly reduced computation time, demonstrating its usefulness in scaling data processing workflows for moderately large datasets.





Discussion

From the results, we observe that:

The overall performance is consistent across the different disease groups, with R² scores ranging between ~0.12 and 0.15, indicating that the model captures some signal, though a substantial portion of the LOS variability remains unexplained.

The Respiratory group shows slightly worse performance in all metrics, suggesting that less data leads to worse results

The Circulatory group yielded the best R² score, which may indicate that early features (demographics and vital signs) are more informative for these cases.

The MAE and RMSE values are fairly close, implying a relatively balanced error distribution without excessive outliers.

Regarding performance, the execution time of cross-validation dropped by nearly 22% with Dask, highlighting its efficiency even when working with datasets of ~30,000 observations.

These findings confirm the viability of applying machine learning to predict ICU length of stay using early-stage clinical data. While the model shows moderate predictive ability, further improvement could be achieved by incorporating richer features such as lab results, treatments administered, or detailed time-series signals.

Conclusion

In this study, we developed a scalable and interpretable machine learning pipeline to predict ICU Length of Stay (LOS) using the MIMIC-III dataset. By integrating demographic, administrative, and early clinical features with distributed data processing tools, we demonstrated the feasibility of LOS prediction in a real-world critical care setting. The Random Forest model achieved reasonable performance, offering valuable insights despite the inherent variability and complexity of clinical outcomes.

Our methodology aligns with the modular design principles of MIMIC-Extract and can serve as a solid baseline for future work. To enhance predictive accuracy and clinical utility, future research could explore the integration of longer temporal windows, dynamic time-series features (e.g., trends in vital signs), and more sophisticated architectures such as gradient boosting or deep learning models. These extensions have the potential to improve not only the precision of LOS predictions but also the broader understanding of patient trajectories in intensive care.

References

MIMIC-Extract: A Data Extraction, Preprocessing, and Representation Pipeline for MIMIC-III

Che, Zhengping, et al. "Recurrent Neural Networks for Multivariate Time Series with Missing Values." *Scientific Reports*, vol. 8, 2018.