Annex A - Presenting datasets in detail

To keep the main work concise, we have chosen to provide a more extensive description of the databases covered in this annex. Although reading it is not mandatory to understand the work conducted in the other chapters, this annex can clarify doubts and explain in more detail the decisions made to assemble the datasets and assess the predictive performance of the ML models.

As we believe that the assembled datasets are a contribution to this work, we have documented here a detailed description, hoping it may be useful for future research built upon this data. The first section of this appendix introduces the data-centric approach we adopted when assembling datasets and discusses the considerations at each stage of this process. We believe it can serve as a guideline for other data-centric works.

After the introductory section, the work is divided into four sections, each referring to a data source. The data sources have one thing in common: They contain health data collected routinely, that is, not for research purposes. In some cases, the same data source originated more than one predictive task. For each generated database, a secondary test database was also created, whenever possible, based on temporal criteria (data collected later) or physical criteria (data collected in another hospital).

Thus, Section introduces the approach used and details the considerations made at each stage of the database assembly process. Section A.2 presents three distinct predictive problems, assembled based on a municipal data source containing information regarding covid-19 tests. Section A.3 outlines a classification problem aimed at predicting a severe condition for a hospitalized covid case, based on the label tests collected on the first day of hospitalization. Section A.4 presents two predictive problems assembled from notifications of dengue cases. Finally, section A.5 describes a predictive problem related to authorization for specialized care in the public health system.

A.1 Data-centric considerations to assemble datasets

Building a classification model involves several crucial steps, encompassing various tasks and methodologies, including data collection, cleaning and labelling and models training, testing and evaluation. to construct effective training data and design proper inference data (FACELI *et al.*, 2021). Each procedure must be carefully considered to ensure the model's effectiveness and reliability (SEEDAT *et al.*, 2022). The emerging field of data-centric AI is still being developed and formalized, so there are no definitive guidelines for building models within this framework. However, there is a consensus that this approach emphasizes collaboration between domain specialists and data scientists at various stages (SEEDAT *et al.*, 2022; PAN *et al.*, 2022; ZHA *et al.*, 2023).

While automation in data preparation for ML models can undoubtedly reduce the chance of human errors, for certain tasks, human involvement is essential to ensure data consistency with the intended purpose. This human engagement is vital for ensuring that AI systems behave in a manner consistent with human intentions. Consequently, the extent of human participation hinges on the goal of aligning the data with human expectations, with relevant considerations varying based on the required level of reliability (SEEDAT *et al.*, 2022; ZHA *et al.*, 2023).

In this section, we describe how we have addressed the various difficulties encountered when dealing with health data. Drawing from both literature and practical experience, we adopted strategies tailored to address these challenges. We sought the assistance of a clinician in preprocessing decisions, which represents a distinguishing factor of these datasets and this work. The collaboration with a domain expert ensured preprocessing decisions based on data nature and collection, enhancing the quality and relevance of the datasets.

To encourage the utilisation of the datasets generated in ML studies, they are publicly accessible in a repository [1]. They can be found in our repository. Each dataset is accompanied by comprehensive documentation to facilitate understanding and usage.

Next, we briefly discuss the considerations in each step of data preparation, focusing on binary classifiers. While our primary focus was not to identify the optimal solution for each problem, we aimed to address them consistently, considering time and resource constraints and mainly seeking solutions that aligned with a data-centric perspective. As this work aims to develop ML predictive systems for health care, we also hope to situate health professionals with some basic ML concepts.

[1] https://github.com/gabivaleriano/HealthDataBR

A.1.1 Data collection

The first step in building a classification model involves gathering and preparing the training data. Raw data is typically not ready for model training due to issues such as noise, inconsistencies, and irrelevant information, which can hinder generalization. Data preparation often accounts for 80% of the time in a data science project, involving challenging cleaning and transformation processes tailored to the unique characteristics of each dataset. This process frequently requires laborious trial and error to achieve the desired quality (SEEDAT *et al.*, 2022).

From a data-centric perspective, data collection heavily relies on domain knowledge. A deep understanding of the application domain is critical for collecting relevant and representative data. Besides that, accurate labeling is essential for supervised learning, and human expertise is often necessary to ensure that labels align with clinical standards. Domain specialists also play a crucial role in identifying relevant data issues, particularly in raw data. They help pinpoint relevant features, potential failures, and the extent to which data values can be trusted, as well as methods to verify data consistency. This guidance is vital for the subsequent steps of data preparation (ZHA *et al.*, 2023).

Reducing the amount of training data while retaining its essential information can improve model accuracy, efficiency, and interpretability. Techniques of feature selection, guided by domain knowledge, help reduce complexity and produce cleaner data (ZHA *et al.*, 2023). In some cases, we adopted strategies to reduce the data size in number of features and instances, always guided by the expertise of specialists. Since decisions were made in the context of each database explored, we will detail the main decisions taken in the next sections.

A.1.2 Class imbalance

All ML predictive models employed in this study utilized a supervised learning approach, that requires a labeled dataset for training. Moreover, our focus primarily revolved around classification tasks, where the target label comprises a finite set of classes. However, in scenarios with imbalanced datasets, ML techniques often exhibit bias toward the majority class at the expense of the minority class (FERNÁNDEZ *et al.*, 2018), which is frequently of greater interest.

To address these instances, we categorized them based on the degree of class imbalance. Let $n_{\min}$ denote the number of elements in the minority class and $n_{\max}$ represent the number of elements in the majority class. Differents strategies were adopted according to the ratio $n_{\min}/n_{\max}$:

- When $n_{\min}/n_{\max} > 0.6$, no further adjustments were made.
- When $0.05 < n_{\min}/n_{\max} \leq 0.6$, examples from the majority class were randomly removed until both classes attained equal representation.
- When $n_{\min}/n_{\max} \leq 0.2$, the minority class was augmented until it comprised 20% of the size of the majority class. This augmentation process involved randomly selecting instances from the minority class with replacement and duplicating them. Subsequently, to balance the class distribution, elements from the majority class were randomly removed until both classes had an equal number of instances.

The strategy of equalizing class representation by randomly removing instances from the majority class offers a simple and effective way to mitigate class imbalance. Achieving a balanced dataset reduces the risk of the model being biased toward the majority class. However, randomly removing examples from the majority class can lead to a loss of potentially valuable information. To address this issue, we also included an oversampling step to prevent training models on significantly reduced datasets compared to the original.

Choosing random instances to duplicate or remove can also introduce bias into the dataset. To mitigate this, the best practice is to repeat the process, generating multiple versions of the dataset, to avoid sample bias. However, this may not be as problematic for our datasets, where the number of instances to be duplicated and removed was substantial only in very large datasets. This ensures that even with the sampling process, a diverse set of instances is still taken into account. In this way, given our goal was to provide a preliminary glimpse into the dataset's potential, we did not perform repeated sampling in this study.

A.1.3 Missing data

Beyond data entry errors, healthcare data often has unique characteristics that contribute to high rates of missing values which further complicates data analysis. For instance, variations in medical practices can lead to differences in the completeness of recorded information across patients. Additionally, patients might omit certain details, or data collection professionals may overlook specific information Kyono *et al.* (2021).

Handling missing values is crucial for ML algorithms, as they generally demand complete datasets for effective model building. These missing values can occur randomly or be dependent on other features, potentially introducing bias and further errors when imputation methods are employed. Such situations can trigger a cascading effect, where initial data issues propagate and adversely impact subsequent tasks.

For these reasons, in this study, we removed all instances containing missing values whenever possible. When removal was not feasible, we adopted a 3-nearest neighbors strategy for imputing missing values. This approach involves identifying the three nearest neighbors of each observation with missing values within the training dataset. The missing values are then imputed as the average of the feature values of these neighboring observations.

The decision to adopt this strategy aimed to ensure a simple mechanism with reduced complexity and computational costs. By considering only the local context of each instance, this approach helps mitigate the effects of outliers and noise present in the dataset. However, we must be aware that instances located near decision boundaries or in overlapping regions may pose additional challenges. In these cases, observations of the opposite class within their neighborhood will influence the imputed values which can further complicate the classification of those instances.

A.1.4 Models evaluation

There are various strategies for learning the approximation function that maps input data to output labels. In machine learning, algorithms are the specific methods used to learn the function $f(x)$ from data. The literature offers a wide range of classification techniques (OSISANWO *et al.*, 2017). However, no single algorithm consistently outperforms others across all types of tasks (WOLPERT; MACREADY, 1997). Consequently, we selected a set of algorithms known for their robust predictive performance on structured data similar to our datasets. Inspired by prior studies (PAIVA *et al.*, 2022), we evaluated our datasets using a pool of seven ML algorithms to compare a variety of models, including simpler alternatives (SEEDAT *et al.*, 2022). These algorithms were implemented using the Scikit-learn library (PEDREGOSA *et al.*, 2011) and are briefly described below.

Each algorithm has its own set of parameters, known as hyperparameters, which need to be set before training begins. Hyperparameters control aspects of the learning process and the model structure. Different combinations of algorithms and hyperparameters perform better for different problems depending on the complexity of the decision boundary required to separate the classes. As commonly practiced, algorithms are chosen based on the specific characteristics of the problem and the data, and hyperparameters are tuned following a search strategy. Given our aim was to provide a preliminary perspective on model performance, we did not perform hyperparameter optimization, as this step can be time-consuming. Instead, we adopted a range of ML models with varying biases to gain initial insights, and used the hyperparameters default values.

- **Support Vector Classifier (SVC)**: SVC is a powerful predictive method rooted in

statistical learning theory and widely utilized in various pattern recognition tasks (LORENA; CARVALHO, 2007). It operates by constructing a hyperplane in a high-dimensional feature space to maximize the margin between different classes, thereby enhancing generalization performance.

- **Gradient Boosting Classifiers (GB)**: GB is an ensemble technique that iteratively combines multiple weak learning models to create a robust predictor. In each iteration, observations misclassified by previous models are assigned higher weights, allowing subsequent models to focus on correcting these errors (NELSON, 2019). The final prediction is obtained through a weighted majority voting strategy across all models.

- **Random Forest (RF)**: RF is another ensemble technique that leverages multiple decision tree models. These trees are constructed using bootstrapped samples of the data and a subset of randomly selected features. The predictions from individual trees are aggregated using a simple majority voting scheme to produce the final prediction (YIU, 2019).

- **Logistic Regression**: Logistic Regression is a linear classification model used for binary classification tasks. It models the probability of a binary outcome using the logistic function and estimates the coefficients of the input features to make predictions (LAVALLEY, 2008). Despite its simplicity, logistic regression is widely used due to its interpretability and efficiency.

- **Bagging**: Bagging, short for Bootstrap Aggregating, is a technique that generates multiple bootstrap samples of the training data and trains a base classifier on each sample. The final prediction is then obtained by averaging the predictions of all base classifiers (LEE *et al.*, 2020). Bagging helps to reduce overfitting and improve the stability and accuracy of the model.

- **Multilayer Perceptron (MLP)**: MLP is a type of artificial neural network characterized by multiple layers of interconnected neurons. It is capable of learning complex relationships between inputs and outputs through a process of forward and backward propagation of errors (POPESCU *et al.*, 2009).

For each classification problem, we constructed a secondary dataset for model evaluation using external data. This approach allowed us to evaluate our models in two steps: first, using only the main dataset, and second, using the secondary dataset to assess the models' ability to generalize.

When splitting the same dataset for training and testing models, selection bias can be introduced. This issue can be mitigated using a cross-validation strategy (SEEDAT *et*

al., 2022). In this approach, both training and testing data are derived from the same dataset through a repeated process. Initially, the dataset is partitioned into five folds of approximately equal size, stratified by class. Four folds are used for training a predictive model, while the fifth fold is reserved for testing, simulating the scenario of presenting new data to the model. This process is repeated five times, with each fold taking a turn as the test set. The average predictive performance across these iterations is then assessed. For evaluating the main dataset alone, we employed a five-fold cross-validation.

Relying solely on performance metrics from training data can fail to capture important properties such as robustness, generalization, and decision-making logic. Therefore, testing models with external data is crucial to determine their generalizability. External evaluation is essential for ensuring model safety and effectiveness in real-world applications (SEEDAT *et al.*, 2022).

External validation is considered the best practice in healthcare (COLLINS *et al.*, 2021). This approach enabled us to determine whether datasets compiled from specific regions or time periods could be generalized to other contexts. For instance, a hospital in a particular city may serve a demographic with distinct social and financial characteristics, potentially affecting the health profiles of its patients compared to hospitals serving different populations. Consequently, models trained on data from one hospital may lack generalizability to others.

By using this two-step evaluation process, we aimed to ensure that our models were not only effective within the confines of the main dataset but also capable of generalizing well to external datasets, thereby assessing their applicability in real-world scenarios.

A.1.5 Data range transformation

Transforming the data range is an essential step in preparing data for machine learning models. It ensures that the features in the dataset are scaled to a specific range, typically between 0 and 1 or -1 and 1. This process is important because it can significantly enhance the performance of the model by ensuring that all features contribute equally to the outcome. Without this transformation, features with larger numerical ranges can dominate those with smaller ranges, leading to biased results and potentially causing the model to learn suboptimal patterns.

Each methods has its advantages and is chosen based on the nature of the data and the requirements of the machine learning model. In this work, before training models we have adopted an standardization step. This method scales the data to have a mean of zero and a standard deviation of one. This approach is particularly useful when the features have different units, it ensures that all features contribute equally, improves model

performance, and handles outliers.

A.1.6 Performance metrics

For each dataset, we report the average AUC (Area Under the ROC Curve) and per-class recall and precision for all prediction models. The AUC quantifies the overall discriminatory power of the model and ranges from 0 to 1, with higher values indicating better performance. An AUC of 0.5 corresponds to random predictions, while values closer to 1 indicate excellent discrimination. This metric is commonly used in medical research to measure the ability to distinguish between classes.

For binary classification problems, the per-class recall corresponds to sensitivity and specificity measures, two metrics widely adopted in clinical research. Recall of the positive class measures the proportion of true positive cases correctly identified by the model, while the recall for the negative class measures the proportion of true negative cases correctly identified.

Similarly, precision per class offers additional insight. While recall measures the quantity of instances predicted correctly, precision can be considered a measure of the quality of predictions. Precision for the positive class measures the proportion of predicted positive cases that are truly positive, while precision for the negative class measures the proportion of predicted negative cases that are truly negative. Both recall and precision range from 0 to 1, with higher values indicating better performance.

These metrics provide comprehensive insights into the predictive performance of the models. The AUC reflects overall discrimination ability, while precision and recall per class offer detailed performance assessments for binary classification tasks. The most important metric to consider depends on the primary objective of the model. When the positive class is smaller and detecting positive samples correctly is the main focus (with less emphasis on correct detection of negative examples), precision and recall values should be prioritized. Conversely, when both classes' detection is equally important, the ROC curve gives equal weight to the prediction ability of both classes.

A.1.7 Feature importance

In addition to the predictive outcomes, we adopted SHAP (SHapley Additive exPlanations) values to interpret the decisions made by our prediction models. SHAP is a widely recognized method employed to assess the importance of each feature in the predictions generated by ML models, providing insights into the primary parameters influencing classification predictions (LUNDBERG; LEE, 2017).

Based on the concept of Shapley values from cooperative game theory, SHAP values are measured by evaluating the contribution of each feature to the prediction, where each feature is considered a "player" in a game. The SHAP value for a feature represents the average contribution of that feature across all possible combinations with other features, offering distribution of importance (LUNDBERG; LEE, 2017).

One way to visualize SHAP values is through a a summary plot. In this representation each point corresponds to a single observation from the dataset. The distance along the x-axis indicates the extent to which a particular variable influenced the model's decision regarding the classification outcome. Variables are ordered based on their importance in the model's overall performance. However, it's important to note that in each instance, the model's output results from the combined influence of all variables.

We specifically chose the Random Forest algorithm to generate SHAP plots due to its robustness and popularity in classification tasks. RF is an ensemble method that combines multiple decision trees, which makes it well-suited for capturing complex interactions between features. By using RF, we can leverage its inherent ability to handle non-linear relationships and high-dimensional data, providing a comprehensive view of feature importance (YIU, 2019).

A.1.8 Data visualization

As visual beings, humans naturally tend to process and retain information presented graphically. Faithful and user-friendly representations of data significantly contribute to the involvement of data specialists in model construction (ZHA *et al.*, 2023). In this way, for binary features in general, and especially for the most important features of each dataset, we present graphical representations of distributions by class and in some cases the co-relation or co-occurrence among the most important features. By doing this we aimed to help understand the data and thereby facilitate comprehension of the model's decision-making process.

Moreover, results suggest that explaining the training data can influence how users assess a system's trustworthiness. Detailed explanations of data used to train the models help people develop an informed sense of trust in ML systems. When users understand the distributions and importance of features, they are better equipped to judge the performance and relevance of the model's predictions (ANIK; BUNT, 2021). This not only increases the model's transparency but also promotes more effective collaboration between data scientists and domain experts.

A.1.9 The covid pandemic in Brazil

Since the two databases adopted in this work are covid-19 related, we present some information to contextualize data collection within the progression of the disease in Brazil. Table A.1 presents the start and end dates of the three main waves, as well as the number of deaths in each wave (MOURA *et al.*, 2022). The table refers to the entire country, while the data we use were collected in two specific cities. Although both cities are located in the state of São Paulo, whose number of cases and deaths follows the same distribution as the country (MOURA *et al.*, 2022).

TABLE A.1 – Characteristics of covid-19 epidemiological waves in Brazil.

Characteristics	**First**	**Second**	**Third**
Start date	23 Feb 2020	08 Nov 2020	26 Dec 2021
End date	07 Nov 2020	25 Dec 2021	21 May 2022
Deaths (number in wave)	162 269	455 379	46 046

Another important piece of information is the starting date of the vaccination campaign. Vaccinations began in Brazil, as well as in the cities where data was collected, in 2021. In last week of January 2021 in the city of São Paulo and in the last week of February in the city of São José dos Campos. The vaccines were initially distributed in doses to a small segment of the population, comprising health workers, older people, and individuals with comorbidities.

A limitation of our work is that we did not have access to the vaccination information of the individuals whose data was collected. At most, we could speculate about the influence of vaccination on the collected data. Although part of the data collection extends through the vaccination period (May 2021), we do not believe that an immunization effect was already in effect. While we cannot entirely dismiss the potential impact of this phenomenon on the disease symptoms and effects on the human body. It is also possible that people sought health services not because of an infection but due to a secondary effect of the vaccines. This would be an example of data drift, when statistical properties of the input data change over time. These changes can lead to discrepancies between the training and testing data, potentially reducing the accuracy and reliability of the model's predictions (MALLICK *et al.*, 2022).

A.2 Severity and hospitalization prognosis using municipal data

The covid-19 pandemic promoted interdisciplinary efforts among scientists worldwide to address numerous critical questions. With the pandemic evolving over the course of more than a year, a substantial volume of data became available for analysis. Through a collaborative partnership with the private non-profit Research and Planning Institute (IPPLAN - Instituto de Pesquisa e Planejamento), we had access to a database containing information on individuals who underwent covid tests in São José dos Campos (SJC). Leveraging this data source, we designed three distinct prediction tasks aimed at optimizing medical care planning for the local population and strategically allocating resources to combat the disease more effectively.

A.2.1 Data source context

In Brazil, it was mandatory for citizens to complete a questionnaire when undergoing a covid-19 test, although not all fields are obligatory. This questionnaire, filled out by pharmacy, laboratory, or hospital staff, is based on patient self-declarations. Consequently, it is common to encounter missing information, typos, and inaccuracies in the data collected.

Since the beginning of the pandemic, the IPPLAN, in association with the City Hall of SJC, compiled this information along with other governmental management data on a daily basis. This database includes information from covid-19 test questionnaires, as well as data from health systems, such as hospitalization needs and reported deaths.

The raw dataset adopted in this study comprises tests conducted from March 1st, 2020, to May 14th, 2021, encompassing 255,815 tested cases and 44 attributes. It is categorized into five primary information groups:

1. Personal information (e.g., gender, date of birth, address, district, and postal code).
2. Covid-19 test results (positive or negative).
3. Disease progression information (e.g., date of first symptoms, hospitalization date, and location of hospitalization).
4. Symptoms (e.g., fever, cough, sore throat, vomiting).
5. Comorbidities (e.g., chronic cardiovascular disease, immunodeficiency).

In the subsequent sections, we outline the preprocessing steps applied to this raw data.

A.2.2 Data preprocessing

Initially, a series of procedures were implemented to clean, adjust, transform, and standardize the data. Some of these steps were recommended by IPPLAN professionals. The initial preprocessing involved removing attributes that were not relevant to our analysis, such as *address, place of admission, date of notification*, as well as eliminating duplicated rows or those containing inconsistent information (e.g., patients over 110 years old).

Additionally, we excluded cases where hospitalization occurred more than 15 days after the onset of symptoms to ensure the correlation between covid-19 positive diagnosis and hospitalization. Similarly, for cases resulting in death, only those occurring within 30 days after the onset of symptoms were considered, ensuring the correlation between covid-19 positive diagnosis and death. Cases where citizens were already hospitalized at the time of testing were also excluded.

In the raw database, a blank space could be interpreted in two ways: as the absence of the symptom or comorbidity, or as neglected information. To address this ambiguity, we implemented a step to clean the data by removing rows with fewer than three fields assigned, aiming to eliminate those that were not correctly filled out. Following this, all blank spaces were treated as the absence of the symptom or comorbidity. Consequently, no further steps were taken to handle missing values in any of the datasets. Additionally, any comorbidities and symptoms reported by less than 1% of the population included in each dataset generated were removed, as they were underrepresented in the sample.

After preparing the raw data, 43,579 observations and 36 attributes were kept, corresponding to 17% of the original number of tests recorded. Regarding gender, there are 19,484 (44.7%) male and 24,095 (55.3%) female individuals. Among the test results, 22,535 (51.7%) were negative and 21,044 (48.3%) were positive.

In the subsequent stage, we filtered the dataset to include only positive cases for covid-19. From this point, different steps were taken for each predictive task. The first task aimed to assist professionals in focusing their attention on particular cases with a higher likelihood of developing serious conditions. The other two tasks aimed to support bed demand planning in hospitals and health centers. In the following subsections, we describe the preprocessing steps specific to each problem.

A.2.2.1 Dataset 1: *severity_sjc*

The objective here is to predict whether a citizen with a positive diagnosis for covid-19 will develop a serious condition, requiring greater medical attention and the reserve of hospital resources.

The severity of covid-19 cases can be defined in multiple ways. The evolution to death

clearly indicates a higher severity of the case. But even if a case does not evolve to death, a long term hospitalization could leave sequels (SAGARRA-ROMERO; VIÑAS-BARROS, 2020) and also brings high demands of human and material resources. Therefore, the need for a long term hospitalization is also considered a criteria for severity in this study.

In this way, the adopted proxy for a severe condition is hospitalization stay longer than ten days or death. The decision of adopting ten days was based on the clinical experience of the doctor who collaborated in this study. The predictive model has two possible outcomes for a given case: *severe* or *non-severe*.

To assemble the dataset for this predictive task we adopted the following steps:

- Removing cases with date of hospitalization but no exit date, as we cannot access the hospitalization stay period for them.
- Removing cases with place of hospitalization but no dates of hospitalization.
- Any cases registering more than 100 days of hospitalization or negative were removed, being considered outliers and typos, respectively.

The final dataset comprises 18,995 cases and 18 attributes. Among these cases, 91% were categorized as *non-severe*, while 9% were classified as *severe*.

A.2.2.2 Dataset 2: *hospitalization_sjc*

The second predictive task aimed to address the question: Which patients, among those who tested positive for covid-19, will need to be hospitalized? The intention here is to predict whether a citizen positively diagnosed for covid-19 will require hospitalization in either Intensive Care Unit (ICU) and non-ICU beds. The possible outcomes are *hospitalized* or *non-hospitalized*. For this dataset we adopted the following the additional actions:

- Removing cases corresponding to citizens that died but were not hospitalized.
- Removing cases with place of hospitalization but no dates of hospitalization.

The dataset in its final version consists of 20,011 cases and 18 attributes. Among them, 79% did not need hospitalization, while 21% required hospitalization.

A.2.2.3 Dataset 3: *hosp_days_sjc*

The objective of the third task is to identify, among patients who were hospitalized, how long they will remain hospitalized. In this way, we aimed to predict the period of

hospitalization of a citizen. Two possible outcomes are considered: *short* (up to 7 days) and *long* (more than 7 days) term hospitalization. These values were adopted aiming to achieve three well-balanced classes. We performed the following specific additional actions:

- Removing cases without any hospitalization information: no hospitalization date nor date of entry into the ICU.
- Removing cases which have a date of hospitalization (either ICU and non-ICU beds), but no date of exit nor any information on case evolution.
- Removing cases with zero or one days of hospitalization.

The final dataset contains 2,828 cases and 22 features. Among these cases, 37.6% were hospitalized for a short term, 28.7% for a medium term, and 33.7% for a long term.

A.2.2.4 Creating train and test datasets

We employed a temporal criterion to create distinct training and test datasets. Data from the first year of the pandemic, spanning from March 1, 2020, to February 28, 2021, constituted our training set. Subsequently, the remaining records within the database, covering the period from March 1, 2021, to May 14, 2021, were designated as our test set. The rationale behind employing a temporal criterion was to assess whether the data collected during the initial year could effectively predict outcomes in subsequent cases. This approach allows us to evaluate the generalizability and predictive power of our model across different temporal phases of the pandemic.

A.2.3 Datasets description

A summary of the characteristics of the three datasets is presented in Table A.2. Symptoms and comorbidities are binary attributes, denoted by a value of *one* when the corresponding symptom/comorbidity is present and *zero* otherwise. Gender is also represented as a binary value, with *one* indicating female. Age is the unique continuous attribute in the datasets.

The *severity_sjc* and *hospitalization_sjc* datasets exhibit some similarities, such as having the same number of features and a comparable number of instances. Additionally, both datasets are imbalanced, with the majority class being *non-severe* and *non-hospitalized*, respectively. In contrast, the *hosp_days_sjc* dataset has approximately balanced classes and is relatively small, as it pertains specifically to hospitalized patients.

TABLE A.2 – Summary of the characteristics of the datasets built from municipal data. Obsev.: Number of observations; Feat.: Number of features; Fem.: percentage of female patients.

Dataset	♯ Observ.	♯ Feat.	% Positive class	% Fem.	Mean age [range]
severity_sjc	15,304	18	severe (07.3)	53.6	43.9 [0 - 108]
severity_sjc_test	3,691	18	severe (16.1)	49.4	45.6 [0 - 105]
hospitalization_sjc	15,963	18	hospitalization (16.9)	53.1	44.3 [0 - 108]
hospitalization_sjc_test	4,048	18	hospitalization (37.0)	48.6	45.6 [0 - 105]
hosp_days_sjc	1,790	22	long-term (48.7)	42.6	58.7 [0 - 108]
hosp_days_sjc_test	1,038	22	long-term (46.1)	43.9	55.3 [0 - 97]

Regarding the similarities across train and test datasets, they show comparable age statistics and percentages of female patients. However, there is a noticeable difference in the proportion of the positive class *severity_sjc* and *hospitalization_sjc* and large number of hospitalizations per period of time in *hosp_days_sjc*. The test sets refer to the period of the third wave, which was less lethal than the first two. Within this context, the most plausible explanation is that over time, the data became more structured and there was an improvement in data collection.

In each subsequent subsection, we provide more information about the dataset characteristics and the performance of ML models for each predictive task.

A.2.3.1 Dataset 1: *severity_sjc*

This dataset was compiled for the purpose of predicting severe cases of covid-19 from the information declared when undergoing the disease test. Severe patients are defined as individuals who either pass away within 30 days after testing positive or require hospitalization for more than 10 days.

Regarding gender distribution, female individuals are predominant inside the *non-severe* class (55%). The opposite trend is observed within the *severe* class, that presents a male predominance (57%). This distribution aligns with existing literature, indicating that severe cases of the disease are more prevalent among male (CAPUANO *et al.*, 2020).

Figure A.1 illustrates the distribution of binary columns, symptoms and comorbidities within the dataset *severity_sjc*. The proportion of symptom (or comorbidity) presence inside each class is adopted, since classes are imbalanced.

The classes show distinct patterns based on symptoms distribution. *Dyspnea*, *respiratory distress*, and *low oxygen saturation* are frequently observed in the *severe* class, with *dyspnea* also present in the *non-severe* class, albeit less frequently. Symptoms such as *sore throat* tend to be more prevalent in the *non-severe* class while *fever* and *cough* exhibit comparable prevalence across both classes. Similarly, certain comorbidities as *cardiovascular disease* and *diabetes* show a higher occurrence in the *severe* class, aligning with existing literature highlighting the significant impact of comorbidities on patient

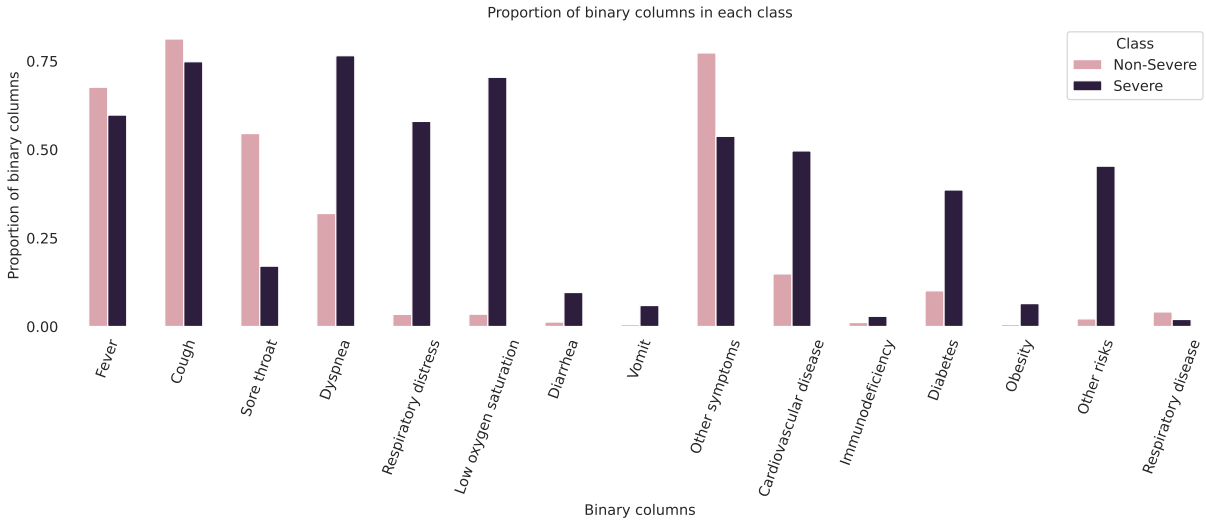


FIGURE A.1 – Distribution of symptoms and comorbidities across *non-severe* and *severe* classes in the *severity_sjc* dataset. Symptoms such as *dyspnea*, *respiratory distress*, and *low oxygen saturation* are more prevalent within the *severe* class, while *sore throat* is more common in the *non-severe* class. Comorbidities such as *cardiovascular disease* and *diabetes* also exhibit varying distributions between the two classes.

prognosis in this disease (SANYAOLU *et al.*, 2020).

Table A.3 displays the average predictive performance and standard deviation of models trained on the *severity_sjc* dataset assessed through a five-fold cross-validation. To address data imbalance, an over-under-sampling step was applied within each training set generated in the cross-validation process (see: A.1.2. The metrics adopted to measure models' performance are AUC, recall and precision per class.

TABLE A.3 – Average predictive performance and standard deviation of models trained on the *severity_sjc* dataset, evaluated using five-fold cross-validation. Classes imbalance was addressed through an over-under-sampling technique. The best value for each metric is highlighted in bold. AUC: Area Under the Curve; RCL: Recall; PRC: Precision; SVC: Support Vector Classifier; RF: Random Forest; GB: Gradient Boosting; LR: Logistic Regression; BAG: Bagging; MLP: Multilayer Perceptron.

	$\mathbf{SVC_{linear}}$	$\mathbf{SVC_{rbf}}$	**RF**	**GB**	**LR**	**BAG**	**MLP**
AUC	0.955 (0.006)	0.954 (0.011)	0.956 (0.009)	0.960 (0.006)	**0.960** (0.005)	0.943 (0.009)	0.957 (0.008)
$\mathbf{RCL_1}$	0.893 (0.023)	**0.920** (0.027)	0.918 (0.022)	0.913 (0.025)	0.884 (0.027)	0.892 (0.023)	0.919 (0.029)
$\mathbf{RCL_0}$	**0.945** (0.002)	0.936 (0.004)	0.915 (0.003)	0.932 (0.001)	0.942 (0.002)	0.913 (0.007)	0.924 (0.006)
$\mathbf{PRC_1}$	**0.559** (0.011)	0.530 (0.019)	0.459 (0.012)	0.514 (0.011)	0.543 (0.005)	0.448 (0.023)	0.489 (0.017)
$\mathbf{PRC_0}$	0.991 (0.002)	**0.993** (0.002)	0.993 (0.002)	0.993 (0.002)	0.990 (0.002)	0.991 (0.002)	**0.993** (0.002)

Overall, the metrics indicate strong predictive performance across all algorithms. However, precision in the *severe* class is consistently low. The recall in the *severe* class is high, indicating that the majority of *severe* cases are correctly classified. These two factors indicate a high number of false positives.

The highest precision for the *severe* class is achieved by the Support Vector Classifier with a linear kernel (0.559) and the best recall for this class is presented by Support Vector Classifier with RBF kernel (0.920). The best recall in the *non-severe* class is reached by Support Vector Classifier with linear kernel (0.945) while SVC with RFB kernel holds the highest precision in this class (0.993, same value presented by Multilayer Perceptrom). Finally, Logistic Regression exhibits the highest AUC (0.991).

Figure A.2 presents a summary plot of the SHAP values extracted from the *severity_sjc* dataset. The SHAP values were extracted using a balanced sample of the *severity_sjc* dataset and employing the Random Forest algorithm. Notably, all variables, except *age*, are binary, facilitating interpretation. Purple dots in the plot indicate the presence of a symptom or comorbidity, while pink dots mean their absence. Features are arranged based on their impact on model decisions.

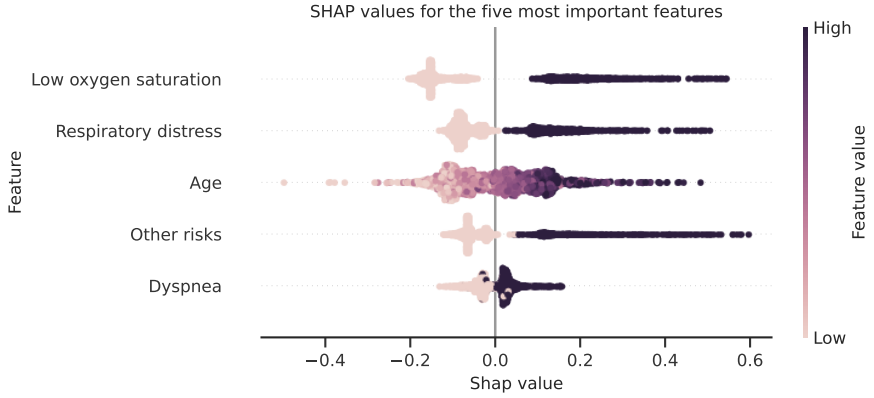
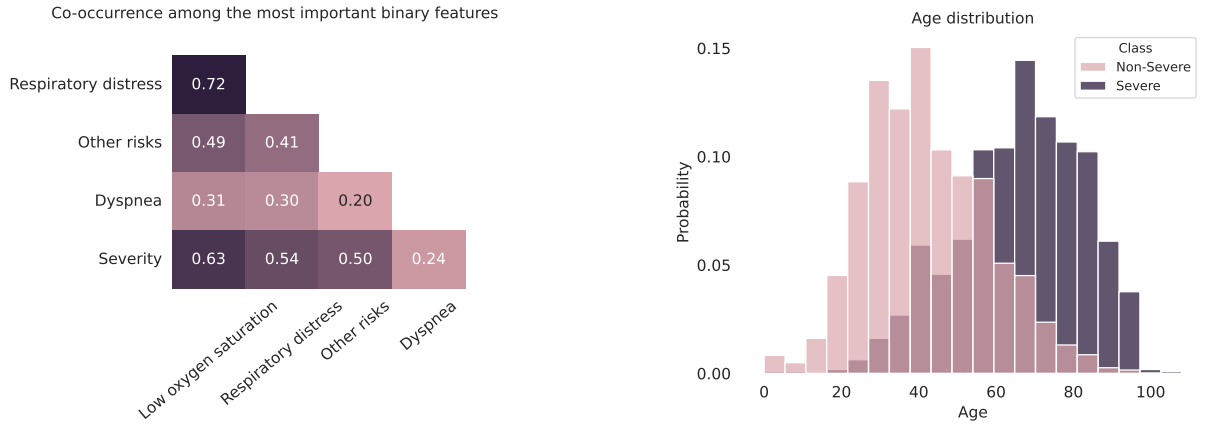


FIGURE A.2 – Summary plot illustrating SHAP values of the five most important features in the *severity_sjc* dataset. SHAP values were computed using a balanced sample of the dataset and the Random Forest algorithm. Each dot represents a single instance and the x-axis measure the impact of a feature in each prediction. Features within the plot are ranked according to their influence on model's decision-making process. For the *age* feature, low values, represented in pink and light purple, influenced the prediction of individual instances as *non-severe*. Binary variables, on the other hand, are represented by purple dots for the presence of symptoms or comorbidities and pink dots for their absence. *Low oxygen saturation* emerges as the most pivotal variable, exerting substantial influence on the predictive outcome. For the selected features, their presence consistently influenced the classification of instances as *severe*.

Instances where patients exhibit *low oxygen saturation* or *respiratory distress* tend to pull the model's output towards the *severe* class. Although the absence of those symptoms have little influence on model decision, which can be visualised in the graph by the dots condensed near the zero SHAP value. In the same way, elderly individuals, represented by dark-purple dots, are more likely to be predicted as *severe* cases. The inclusion of the feature *other risks* among the top-ranked factors suggests that comorbidities significantly influence predictions towards the *severe* class, despite not being explicitly specified in the data source.

We conducted further exploration on the five most critical features of the *severity_sjc* dataset, identified through their SHAP values. In Figure A.3, the co-occurrence between



(a) Co-occurrence among key features in the *severity_sjc* dataset. The strongest association is observed between *low oxygen saturation* and *respiratory distress* symptoms. A significant co-occurrence is noted between these symptoms and severity.

(b) Distribution of *age* by class in the *severity_sjc* dataset. The density probability function is used for normalization. Class *severe* distribution exhibits a shift to the right, indicating higher values of *age*.

FIGURE A.3 – Exploration of the five most important features within the *severity_sjc* dataset. Left: Co-occurrence between the most important binary features. Right: Distribution of *age* by class.

binary features and the distribution of *age* are depicted. The histogram uses the density probability function for normalization, enabling a comparative analysis between classes, despite the unbalanced number of samples.

Notably, in Figure A.3a a strong co-occurrence, exceeding 0.7, is observed between the two most pivotal symptoms: *low oxygen saturation* and *respiratory distress*. Furthermore, their co-occurrency with severity surpasses 0.5, explaining their significance in predicting outcomes and contributing to the dataset's high predictive performance. In Figure A.3b, the difference between *age* distributions is evident, with patients in the *severe* class exhibiting advanced ages.

Validating with external data

To perform external validation, we assembled a second dataset from the same data source. While the main dataset encompasses data from the first year of the pandemic, the *severity_sjc_test* dataset includes information from the following months, specifically tests collected from March to May 2021. Table A.4 presents the predictive performance of the models when trained on the main dataset and tested on the external data. A balanced sample of the main dataset was used to train the models.

The metrics for the external evaluation exhibit the same pattern as the cross-validation results. The predictive values are high, except for precision in the positive class, indicating a high rate of false positives. The Random Forest algorithm achieves the best recall for the *severe* class (0.887) and the best precision in the *non-severe* class (0.974, same value obtained by Gradient Boosting). Logistic Regression shows the highest precision in the

TABLE A.4 – Average predictive performance and standard deviation of models trained on the *severity_sjc* dataset and tested on the *severity_sjc_test* dataset. Class imbalance on training data was addressed through an over-under-sampling technique. The best value for each metric is highlighted in bold. AUC: Area Under the Curve; RCL: Recall; PRC: Precision; SVC: Support Vector Classifier; RF: Random Forest; GB: Gradient Boosting; LR: Logistic Regression; BAG: Bagging; MLP: Multilayer Perceptron.

	$\mathbf{SVC_{linear}}$	$\mathbf{SVC_{rbf}}$	**RF**	**GB**	**LR**	**BAG**	**MLP**
AUC	0.907	0.890	0.897	**0.913**	0.915	0.883	0.907
$\mathbf{RCL_1}$	0.862	0.881	**0.887**	0.884	0.861	0.854	0.881
$\mathbf{RCL_0}$	0.846	0.835	0.823	0.835	**0.849**	0.821	0.828
$\mathbf{PRC_1}$	0.519	0.506	0.491	0.508	**0.523**	0.478	0.496
$\mathbf{PRC_0}$	0.970	0.973	**0.974**	**0.974**	0.969	0.967	0.973

positive class (0.523) and the best recall in the negative class (0.849). Finally, Gradient Boosting presents the best AUC value (0.913).

A.2.3.2 Dataset 2: *hospitalization_sjc*

This dataset was compiled to predict the hospitalization of citizens in a fortnight period after testing positive for covid-19. It includes information provided in the form completed during the covid-19 testing process. Features distribution inside this dataset mirrors the patterns observed in the *severe_sjc* dataset (see Section A.2.3.1). Consequently, we will focus our description here on the points of differentiation between them.

Gender distribution follows a similar pattern to the one described in the previous subsection, where female individuals are more prevalent within the *non-hospitalized* class (55.3%), contrasting with their minority representation among hospitalized patients (42.7%).

Figure A.4 illustrates the distribution of binary features (symptoms and comorbidities) within the dataset *hospitalization_sjc*, adopting the proportion relative to the class size.

Once again, distinct patterns emerge between classes. Although, in this dataset, certain features, such as *respiratory distress*, *low oxygen saturation*, and *other risks*, are almost exclusively present in the *hospitalized* class.

In Table A.5, we present the average and standard deviation of the AUC, recall and precision per class of the predictive models evaluated on the *hospitalization_sjc* dataset. Models were evaluated through a five-fold cross-validation. To address data imbalance, an over-under-sampling step was applied to the majority class within each training set generated in the cross-validation process. Similar to the predictions for severity, the performances achieved here are notably high.

Interestingly, in this dataset, the precision in the positive class is notably high, which contrasts with the *severity* prediction. Despite the similarity between the datasets, here models could effectively differentiate between *hospitalized* and *non-hospitalized* patients without generating false positives. Among the evaluated models, Gradient Boosting

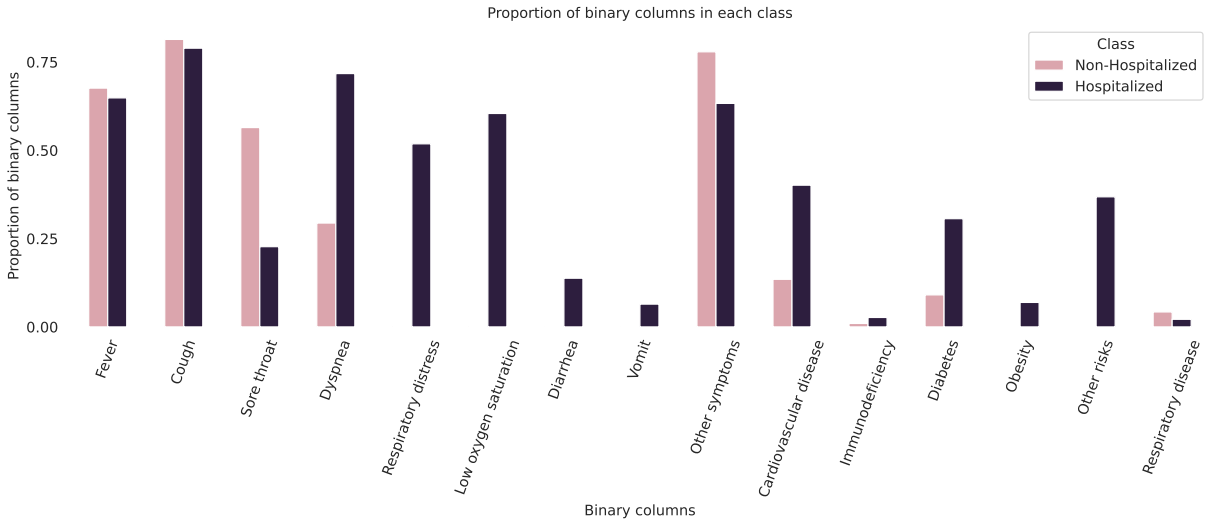


FIGURE A.4 – Distribution of symptoms and comorbidities among hospitalized and non-hospitalized patients. The *hospitalized* class shows a higher prevalence of features such as *respiratory distress*, *low oxygen saturation*, *other risks* that almost exclusively characterize this class. The presence of comorbidities like *cardiovascular disease* and *diabetes* are also more commonly observed in the positive class. Conversely, *sore throat* and *other symptoms* are more characteristic of non-hospitalized patients. The symptoms *fever* and *cough* exhibit comparable prevalence across both classes.

TABLE A.5 – Average predictive performance and standard deviation of models trained on the *hospitalization_sjc* dataset, evaluated using five-fold cross-validation. Class imbalance was addressed through an over-under-sampling technique. The best value for each metric is highlighted in bold. AUC: Area Under the Curve; RCL: Recall; PRC: Precision; SVC: Support Vector Classifier; RF: Random Forest; GB: Gradient Boosting; LR: Logistic Regression; BAG: Bagging; MLP: Multilayer Perceptron.

	$\mathbf{SVC_{linear}}$	$\mathbf{SVC_{rbf}}$	**RF**	**GB**	**LR**	**BAG**	**MLP**
AUC	0.940 (0.024)	0.955 (0.009)	0.959 (0.005)	**0.968** (0.004)	**0.968** (0.002)	0.952 (0.005)	0.966 (0.003)
$\mathbf{RCL_1}$	0.871 (0.015)	0.872 (0.014)	0.897 (0.013)	0.882 (0.016)	0.882 (0.017)	0.888 (0.011)	**0.890** (0.016)
$\mathbf{RCL_0}$	**0.996** (0.000)	0.993 (0.002)	0.940 (0.006)	0.988 (0.001)	0.987 (0.002)	0.941 (0.006)	0.974 (0.006)
$\mathbf{PRC_1}$	**0.978** (0.003)	0.961 (0.009)	0.752 (0.019)	0.936 (0.006)	0.934 (0.011)	0.756 (0.017)	0.876 (0.023)
$\mathbf{PRC_0}$	0.974 (0.003)	0.974 (0.003)	**0.978** (0.003)	0.976 (0.003)	0.976 (0.003)	0.976 (0.002)	0.977 (0.003)

demonstrates the highest AUC (0.968), the same value achivied by Logist Regression in this metric. Multilayer Perceptron exhibits the highest recall for the positive class (0.890). The best recall for the *non-hospitalized* class is presented by Support Vector Classifier with a linear kernel (0.996) that also holds the highest precision for the *hospitalized* class. Finally, Random Forest showcases the best precision for the negative class (0.978).

Figure A.5 illustrates SHAP values associated with the dataset *hospitalization_sjc*. Similar to the approach taken with the *severity_sjc* dataset, this plot was generated using a balanced sample of the dataset, adopting an over-under-sampling step, and the Random Forest algorithm.

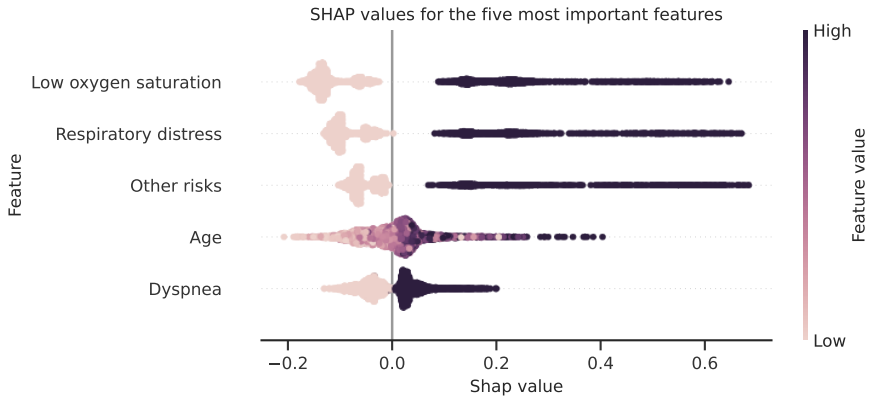
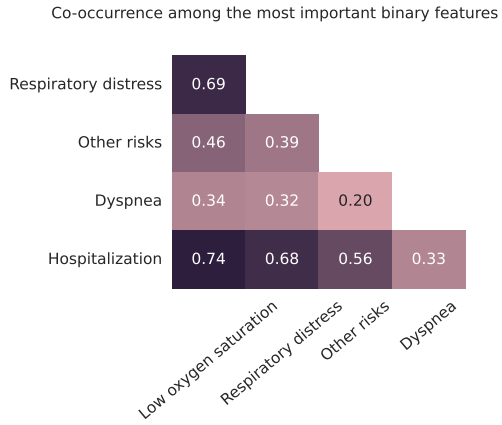


FIGURE A.5 – SHAP values visualization of the most important features when predicting hospitalization using the *hospitalization_sjc* dataset. SHAP values were generated through a balanced sample of the dataset and the Random Forest algorithm. Key influential attributes include *low oxygen saturation* and *respiratory distress*. The feature *age* ranks differently in importance compared to its role in predicting severity.

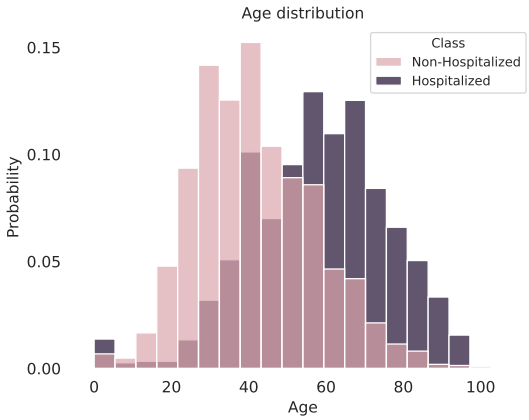
Comparable observations with the *severity_sjc* dataset can be drawn here as well. The attributes exerting the most significant influence on the need for hospitalization pertain to symptoms, particularly *low oxygen saturation* and *respiratory distress*. However, a difference arises in the importance of the feature *age*: while it ranks as the third most important feature for predicting severity, it holds the fourth position when predicting hospitalization. This suggests that age may play a more crucial role in predicting the development of *severe* cases, while its significance decreases when predicting hospitalization.

To gain further insights into the five most important features, we present two additional plots. Figure A.6 illustrates the co-occurrence between the most important binary features. The figure also presents the normalized distribution of *age* within each class, achieved through the probability density function normalization method.

Notably, the co-occurrence between *respiratory distress* and *hospitalization* in Figure A.6a appears more pronounced here compared to the *severity_sjc* dataset, reaching a correlation coefficient of 0.74. Moreover, strong correlations (> 0.5) are observed between



(a) Co-occurrence among key features in the *hospitalization_sjc* dataset. There is a pronounced correlation between *low oxygen saturation hospitalization*. *Respiratory distress* also presents high correlation with the target feature.



(b) Normalized distribution of *age* within each class of the *hospitalization_sjc*, normalization through the probability density function. There is a wider range of overlapping values when comparing with the severity prediction. The trend between classes persists with older patients more prevalent in the *hospitalized* class.

FIGURE A.6 – Exploration of the five most important features within the*hospitalization_sjc* dataset. Left: Co-occurrence between the most important binary features. Right: Distribution of *age* by class.

the target feature and low *oxygem saturation* as well as *other risks*.

In Figure A.6b, we visualize a wider range of overlapping distributions, potentially explaining the decreased importance of *age* in predicting hospitalization compared to severity. Despite this increased overlap, the trend between classes persist, with older patients more prevalent in the *hospitalized* class.

Validating with external data

A second dataset was assembled from the same data source, containing the information of covid tests undertaken in São José dos Campos from March to May of 2021, the months following the first year of the pandemic. A balanced sample of the main dataset was adopted to train models and their were now evaluated in the test set. Table A.6 presents models performance in this external evaluation.

All metrics in the external evaluation demonstrate high predictive performance, indicating the models' ability to generalize effectively. These results indicate that models built from the first year of the pandemic would be useful for predicting future hospitalization needs.

Bagging presents the highest recall for the negative class (0.889) and precision for the positive class (0.935), the same value achieved by Random Forest in this metric. Support Vector Classifier with a linear kernel shows the best precision for the hospitalized patients (0.980) and the best recall for the non-hospitalized class (0.989), the same matched by SVC with an RBF kernel. Finally, Gradient Boosting demonstrates the highest AUC

TABLE A.6 – Average predictive performance and standard deviation of models trained on the *hospitalization_sjc* dataset and tested on the *hospitalization_sjc_test* dataset. Class imbalance on training data was addressed through an over-under-sampling technique. The best value for each metric is highlighted in bold. AUC: Area Under the Curve; RCL: Recall; PRC: Precision; SVC: Support Vector Classifier; RF: Random Forest; GB: Gradient Boosting; LR: Logistic Regression; BAG: Bagging; MLP: Multilayer Perceptron.

	$\mathbf{SVC_{linear}}$	$\mathbf{SVC_{rbf}}$	**RF**	**GB**	**LR**	**BAG**	**MLP**
AUC	0.957	0.947	0.957	**0.969**	0.965	0.952	0.965
$\mathbf{RCL_1}$	0.874	0.874	0.888	0.878	0.878	**0.889**	0.882
$\mathbf{RCL_0}$	**0.989**	**0.989**	0.949	0.986	0.987	0.942	0.978
$\mathbf{PRC_1}$	**0.980**	0.979	0.912	0.974	0.976	0.900	0.959
$\mathbf{PRC_0}$	0.931	0.931	**0.935**	0.933	0.932	**0.935**	0.934

(0.969).

A.2.3.3 Dataset 3: *hosp_days_sjc*

This dataset was compiled to predict the length of hospitalization stay among individuals with covid-19 admitted to hospitals. We framed the duration of hospital stay as a classification task, aiming to predict whether the stay would be categorized as *short* or *long*.

The gender distribution remains consistent across each class, with female individuals comprising the minority in the long hospitalization (44.7%) and in the short hospitalization class (40.5)%.

Figure A.7 depicts the distribution of binary columns (symptoms and comorbidities) within each of the three classes.

While there is little variation in the distribution of symptoms between the classes, a few exhibit discernible patterns of either increase or decrease in frequency with prolonged hospitalization. Notably, *sore throat* frequency decrease with longer hospital stays, while *low oxygen saturation*, *cardiovascular disease*, *diabetes*, and *other risks* display trends of increasing prevalence in *long* hospitalization stays.

Table A.7 presents the average and standard deviation of recall and precision, per class, as well as the general AUC results for models constructed using the *hosp_days_sjc* dataset. Models were evaluated through a five-fold cross-validation.

Unlike the previous models, the predictive performance achieved by all ML algorithms here was not satisfactory, with AUC, precision and recall values only slightly exceeding 0.5. This indicates significant challenges in accurately predicting the duration of a patient's hospitalization. Particularly disappointing are the results for the medium-term stay class, where predictive sensitivity and specificity are notably lacking. Attempts were also made to develop regression models to predict the actual number of hospitalization days, but

FIGURE A.7 – Distribution of symptoms and comorbidities within each class of the *hosp_days_sjc* dataset. While there is minimal variation in the distribution of symptoms across the classes, some exhibit patterns of either increase or decrease in frequency with prolonged hospitalization. The frequency of *sore throat* decreases with longer hospital stays, whereas *low oxygen saturation*, *cardiovascular disease*, *diabetes*, and *other risks* show trends of increasing prevalence.

TABLE A.7 – Average predictive performance and standard deviation of models trained on the *hospitalization_days_sjc* dataset, evaluated using five-fold cross-validation. The best value for each metric is highlighted in bold. AUC: Area Under the Curve; RCL: Recall; PRC: Precision; SVC: Support Vector Classifier; RF: Random Forest; GB: Gradient Boosting; LR: Logistic Regression; BAG: Bagging; MLP: Multilayer Perceptron.

	$\mathbf{SVC_{linear}}$	$\mathbf{SVC_{rbf}}$	**RF**	**GB**	**LR**	**BAG**	**MLP**
AUC	0.595 (0.023)	0.569 (0.016)	0.531 (0.019)	0.551 (0.023)	0.598 (0.023)	0.529 (0.034)	0.536 (0.013)
$\mathbf{RCL_1}$	0.591 (0.037)	0.568 (0.028)	0.497 (0.030)	0.530 (0.032)	0.546 (0.021)	0.435 (0.045)	0.527 (0.020)
$\mathbf{RCL_0}$	0.543 (0.039)	0.541 (0.035)	0.542 (0.037)	0.532 (0.021)	0.587 (0.047)	0.586 (0.022)	0.530 (0.027)
$\mathbf{PRC_1}$	0.552 (0.025)	0.541 (0.019)	0.508 (0.016)	0.518 (0.017)	0.558 (0.025)	0.498 (0.022)	0.516 (0.007)
$\mathbf{PRC_0}$	0.583 (0.029)	0.569 (0.020)	0.531 (0.015)	0.544 (0.018)	0.576 (0.020)	0.522 (0.017)	0.542 (0.008)

these efforts yielded similarly low predictive performance.

The SHAP values were computed for this predictive task, providing insights into the importance of each feature for predicting each class. To identify the overall most important features, we calculated the mean SHAP values across the three classes and selected features with the highest average. Figure A.8 illustrates the mean SHAP values within each class for the five selected features. Since the number of instances in each class was balanced, we computed the SHAP values using the complete dataset, the values were computed using the Random Forest algorithm.

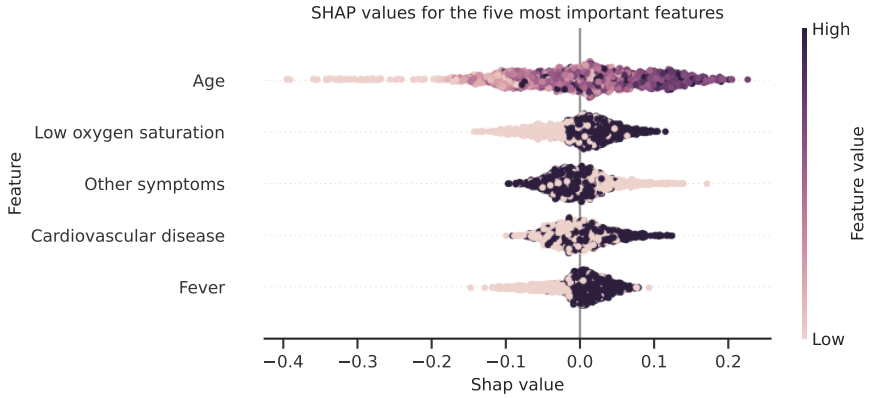


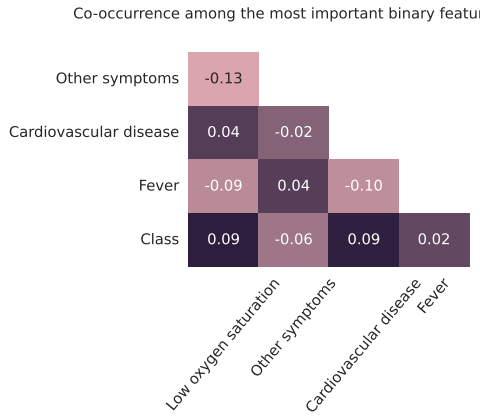
FIGURE A.8 – Mean SHAP values within each class for the five selected features in the predictive task of the *hosp_days_sjc* dataset. The SHAP values were computed using the Random Forest algorithm, since the number of instances in each class was balanced, the complete dataset was adopted. Among the selected features, *age* exhibited the highest mean SHAP values.

Among the selected features, *age* emerged with the highest mean SHAP values. Furthermore, it appears that nearly all selected features mainly influenced predictions for the *long* hospitalization classes, with the exception of *other symptoms*.

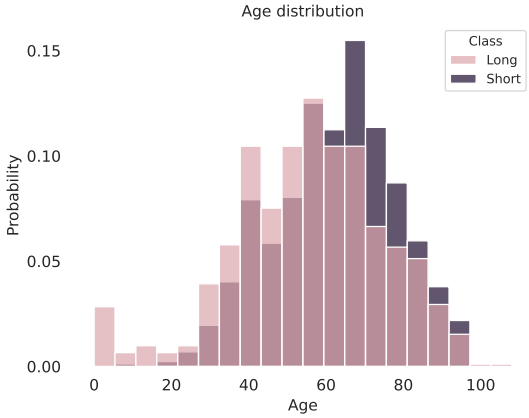
To a deeper comprehension of the five selected features, determined through mean SHAP values across different classes, we conducted an analysis of their correlation with hospitalization length. Additionally, we examined the distribution of the feature *age* within each class. Figure A.9 presents both plots.

In Figure A.9a is depicted the co-occurrence values between hospitalization and the chosen features. Hospitalization stay is referred to as *class* in this figure. Any selected feature presents a high co-occurrence with hospitalization stay. The features with greater level are *other symptoms* and *low oxygen saturation* that presents a negative co-occurrence (-0.13).

The distribution of *age* within each class is depicted in Figure A.9b. Given the balanced nature of classes within *hosp_days_sjc*, this histogram employs frequencies instead of probabilities as utilized in preceding sections. A slight leftward shift is notable in the *short* class, indicating a younger population, with the opposite for the *long* class.



(a) Co-occurrence among the five most important features and the hospitalization stay within the *hosp_days_sjc* dataset. The hospitalization stay, is denoted as *class* in this figure. The co-occurrence coefficients are generally small, with the highest value observed betwee *other symptoms* and *low oxygem saturation*.



(b) Distribution of *age* by class within the *hosp_days_sjc* dataset. The distribution of the *short* class shows a slight shift to the left, indicating lower values of *age*, whereas the *long* class exhibits the opposite trend. There is an overlap between classes across the entire range of values.

FIGURE A.9 – Exploration of the five most important features within *hosp_days_sjc* dataset. Left: Correlation between binary features. Right: Distribution of *age* by class.

However, there is overlap across the entire range of values in these distributions.

Validating with external data

From the same data source we compiled a second dataset, following the same preprocessing criteria. The *hosp_days_sjc_test* dataset contains the information regading hospitalizations that occurred from March to May of 2021, the second year of the pandemic. We adopted this dataset to assess the predictive performance of models built with the *hosp_days_sjc* dataset. The result of this evaluation is presented in Table A.8.

TABLE A.8 – Average predictive performance and standard deviation of models trained on the *hosp_days_sjc* dataset and tested on the *hosp_days_sjc_test* dataset. The best value for each metric is highlighted in bold. AUC: Area Under the Curve; RCL: Recall; PRC: Precision; SVC: Support Vector Classifier; RF: Random Forest; GB: Gradient Boosting; LR: Logistic Regression; BAG: Bagging; MLP: Multilayer Perceptron.

	$\mathbf{SVC_{linear}}$	$\mathbf{SVC_{rbf}}$	**RF**	**GB**	**LR**	**BAG**	**MLP**
AUC	0.585	0.575	0.568	0.589	0.589	0.540	0.557
$\mathbf{RCL_0}$	0.480	0.499	0.489	0.491	0.395	0.399	0.493
$\mathbf{RCL_0}$	0.621	0.596	0.598	0.626	0.683	0.620	0.587
$\mathbf{PRC_1}$	0.520	0.514	0.510	0.529	0.516	0.473	0.505
$\mathbf{PRC_0}$	0.582	0.581	0.577	0.589	0.568	0.546	0.574

We observe in Table A.8 the same patter observed in the cross-validation. The predictive performance is very low and did not reach 0.6 for most part of algorithms and metrics.

A.2.4 Discussion

From a database comprising the symptoms and comorbidities of patients who underwent a covid-19 test in the city of São José dos Campos, we generated three distinct datasets. The first dataset, *severity_sjc*, aims to predict whether a patient will develop a severe case of covid-19. The second dataset, *hospitalization_sjc*, was created to forecast the need of hospitalization. Finally, the *hosp_days_sjc* dataset presents a classification problem aiming to predict the length of a hospitalized patient's stay.

We provided a summary of feature distributions for each dataset to offer a broader understanding of their potential. Additionally, we assessed the performance of several ML models in the predictive tasks created. For the first dataset, models consistently exhibited low performance in terms of precision for the positive class, in both cross-validation and external evaluation. This indicates a high rate of false positives. This means that many *non-severe* patients were incorrectly classified as *severe*. However, the models achieved a high level of recall for the severe class, indicating that most *severe* cases were correctly identified.

Within the *hospitalization* problem the predictive results achieved by all models, in both evaluations, suggest that they can effectively support decision-making regarding the allocation of hospital resources. In both cases, *severity* and *hospitalization*, certain features such as *low oxygen saturation*, *respiratory distress* and *other risks*, related to comorbidities, influenced the models' predictions. Elderly individuals also skewed predictions towards severity and the need for hospitalization. These findings align with existing literature on covid-19.

Regarding the third dataset, the predictive performance did not reach desired levels neither in cross-validation or external evaluation. There may be two reasons for the challenge. First, there may be administrative factors affecting the length of hospitalization, such as the availability of hospital beds and the waiting list of patients with more severe conditions. Second, the patient's recovery time could vary depending on their response to the treatment provided, which may not always align with the attributes considered. These hypotheses were confirmed by a medical expert. Therefore, additional factors beyond just age, sex, initial symptoms, and comorbidities are needed to track whether a patient will require a short, medium, or long-term hospital stay. This could include blood tests and other monitoring results during hospitalization.

Interestingly, the findings indicate that routine information collected during a SARS-CoV-2 diagnosis test can be used to build accurate predictive models for determining whether a new case will require hospitalization. This highlights the importance of such information in supporting efficient management of hospital resources, which is beneficial for the general population and public authorities. It also encourages more diligent

completion of medical records and forms.

A.3 Severity prognosis for hospitalized patients

Since the onset of the covid-19 pandemic, numerous studies employed laboratory tests to build ML models for diagnosing the disease, predicting the progression to severe conditions and forecasting the need for intensive treatments (MORAES *et al.*, 2020; FERNANDES *et al.*, 2021). In Brazil, a large developing country significantly affected by the covid-19 pandemic, there was initiatives focused on collecting, storing, and making laboratory data publicly available.

One such initiative was the *Covid Data Sharing/BR* project, which provided laboratory and demographic information from covid-19 patients treated at several major hospitals in the São Paulo metropolitan area (MELLO *et al.*, 2020). This project was supported by FAPESP (São Paulo Research Foundation). Among the five available databases, we adopted two that contain information on patient outcomes: Hospital Sírio Libanês (HSL) and Hospital Beneficência Portuguesa (HBP). These databases were used to create two distinct datasets with identical features to predict the prognosis of severe conditions in hospitalized patients based on the blood tests performed on the day of admission.

A.3.1 Data source context

The *Covid Data Sharing/BR* repository is a publicly accessible digital database that contains laboratory test results collected in hospitals. For each patient who underwent a covid-19 test, whether hospitalized or receiving ambulatory care, all additional blood tests performed on the patient were recorded in the database. Data collection in both adopted databases starts in March 2020. Data collection finishes in different moments, in HSL, it concludes in May 2021, while in HBP, it concludes in January 2021.

The data presented notable differences in the lab tests collected in each patient, leading to a high rate of missing values. Additionally, there were variations in test nomenclature and measurement units across different health centers. There was also considerable variability in the type of data available per hospital, which may stem from differences in patient profiles at each center and the lack of a standardized protocol for covid-19 care and treatment.

Each hospital's raw database consisted of three files. Each one containing:

- Demographic patient information such as birth year and gender.
- Laboratory test results performed at the hospital

- Information about the patient's hospital stay and outcomes.

These databases were used to create two distinct datasets. The next subsection provides a brief overview of the data preprocessing. Subsequently, the first dataset, *severity_hsl*, is examined with an in-depth analysis of the most important features and some predictive results of ML models applied to this dataset under cross-validation assessment. Additionally, a cross-check study uses the second dataset, *severity_hbp*, as a test set to evaluate the ability of generalization of the models.

A.3.2 Data preprocessing

Next, we summarize the preprocessing decisions taken to prepare the data. These preprocessing steps were designed to clean, filter, and structure the dataset to train ML models, ensuring data quality, consistency, and relevance for predicting covid-19 severity outcomes. These steps were applied to each database, from HSL and HBP, resulting in the creation of two datasets. We emphasize that all decisions were made in consultation with a medical data specialist.

- Inconsistency identification and renaming: The laboratory test labels in the raw database had inconsistencies, which were identified and renamed with the help of a domain expert.

- Data merging: The three files in the database, containing patient demographic information, laboratory test results, and hospital stay/outcome information, were joined together to create a unified table.

- Filtering hospitalized patients: The raw database contained various types of attendance records. To focus specifically on hospitalized patients, we performed a filtering step and retained only those patients with a registered hospitalization type of attendance.

- Filtering positive covid-19 cases: Only the laboratory tests of patients who confirmed covid-19 diagnosis by *C-Reactive Protein* test within 15 days after hospital attendance were retained.

- Discarding late tests: Laboratory tests performed after the fourth day of hospitalization were discarded, as the focus was on capturing early prognosis.

- Retaining the first test result: When multiple results were available for a laboratory test for a single patient, only the first value was kept. This was done to capture the patient's admission status and provide an early prognosis.

- Filtering by age: Only data from young and adult individuals (age > 16) were kept, as laboratory test results for children may not be comparable to those of adults.
- Selecting hospitalized patients: Only data from hospitalized patients were retained, as the goal was to predict severity outcomes based on hospitalization days and death.
- Outcome selection: Only patients with outcomes of death, medical discharge (cured or improved), or long-term hospitalization were included.
- Adding a label to each instance: A severity column was added to label each patient as severe or not based on the following criteria: a severe patient has a hospitalization length of 14 days or more or progresses to death within 14 days; otherwise, it is labeled as a non-severe case.
- Handling multiple hospitalizations: In cases where a patient had multiple hospitalizations, only the last hospitalization record was kept to remove duplicates and dependency between instances.
- Outlier removal: The most extreme four values for each laboratory test were removed. This step aimed to address outliers that could be attributed to human or experimental errors.
- Retaining tests with sufficient values: Laboratory tests were retained if they had at least 50% of values filled for each class (*severe* and *non-severe*).
- Variance-based feature selection: Features with a variance lower than 0.05 were eliminated, as they were deemed to contribute less to the ML models' discrimination power.
- Domain expert examination and elimination of features: Laboratory tests that were not directly related to covid-19 were eliminated based on expert evaluation.
- Handling data redundancies: For feature pairs with a correlation index greater than 0.9, one of them was randomly removed to eliminate redundancies.
- Ensuring similarity between hospitals: Lab tests that are not present in one of the datasets were removed, so that we can generate models that generalize across the two hospitals considered.

In this work we have adopted as severity criteria 14 or more days of hospitalization or death. The value was adopted based on the median hospitalization time reported in the literature (WU *et al.*, 2020) and confirmed in the population contemplated in this study.

A.3.3 Datasets description

After completing all the preceding steps, we obtained two datasets. Table A.9 provides an overview of the main characteristics of the datasets, including the number of patients, the number of features (admission tests) included, the percentage of missing values, the percentage of severe cases, the percentage of male patients, and the age range of patients in each dataset.

TABLE A.9 – Summary of the datasets created, including number o patients, number of admission tests considered, percentage of missing values, percentage of severe cases, percentage of male patients and age statistics

Hospital	**♯ Patients**	**♯ Features**	**% NA**	**% Severe**	**% Male**	**Mean age [min - max]**
HSL	1,432	20	6.62	36.7	65.6	62.2 [19 - 89]
HBP	185	20	8.97	36.2	55.1	59.4 [24 - 86]

The *severity_hsl* dataset includes a larger sample size of 1,432 patients, whereas the *severity_hpb* dataset contains only 185 patients. The larger sample size in the *severity_hsl* dataset offers greater statistical power and robustness in the analysis, which is why it was selected for the primary analyses. The *severity_hbp* dataset will be utilized as a test set in a subsequent phase of the study.

Table A.10 presents the mean and standard deviation of several ML models assessed through a five-fold cross-validation. To achieve balanced samples, an undersampling step was applied to the majority class within each training set generated in the cross-validation process. Missing values were addressed using imputation, performed within each round of train and test split, adopting the mean value calculated across the three nearest neighbors. The models were evaluated based on three metrics: AUC, recall and precision per class.

TABLE A.10 – Average predictive performance and standard deviation of models trained on the *severity_hsl* dataset, evaluated using five-fold cross-validation. Class imbalance was addressed through an undersampling technique. The best value for each metric is highlighted in bold. AUC: Area Under the Curve; RCL: Recall; PRC: Precision; SVC: Support Vector Classifier; RF: Random Forest; GB: Gradient Boosting; LR: Logistic Regression; BAG: Bagging; MLP: Multilayer Perceptron.

	SVC_{linear}	**SVC_{rbf}**	**RF**	**GB**	**LR**	**BAG**	**MLP**
AUC	0.732 (0.022)	0.738 (0.016)	**0.747** (0.016)	0.736 (0.018)	0.741 (0.020)	0.709 (0.012)	0.730 (0.014)
RCL_1	0.679 (0.039)	**0.697** (0.055)	0.688 (0.033)	0.679 (0.022)	0.665 (0.028)	0.592 (0.016)	0.652 (0.034)
RCL_0	0.666 (0.046)	0.643 (0.046)	0.660 (0.054)	0.640 (0.039)	0.679 (0.036)	**0.710** (0.040)	0.677 (0.042)
PRC_1	0.545 (0.025)	0.535 (0.024)	0.545 (0.033)	0.526 (0.024)	**0.550** (0.026)	0.546 (0.033)	0.543 (0.040)
PRC_0	0.780 (0.016)	**0.785** (0.026)	0.783 (0.017)	0.773 (0.011)	0.776 (0.015)	0.748 (0.013)	0.768 (0.024)

The models presented an intermediate performance, specially low in the precision

of the positive class. The best AUC is presented by Random Forest. Support Vector Classifier with RFB kernel holds the highest recall in the *severe* class (0.697) while the best recall in the opposite class is exhibited by Bagging (0.710). The Logistic Regression demonstrates the best precision in the positive class (0.550), for the negative class Support Vector classifier presents the highest value of precision (0.785). The predictive performance observed here is significantly lower than the performance achieved with the symptom and comorbidity data recorded in the covid test (see: A.2.3.1).

Figure A.10 presents the SHAP summary plot for the *severity_hsl* dataset. SHAP values were extracted using the Random Forest algorithm on a balanced version of the dataset, which was achieved by undersampling the majority class.

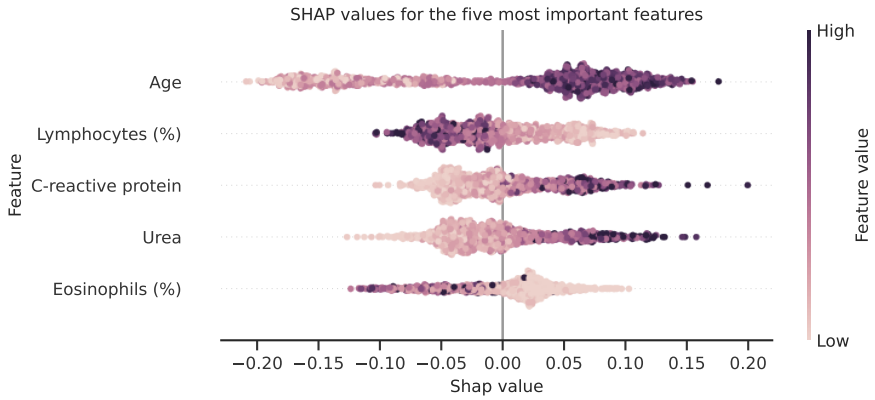


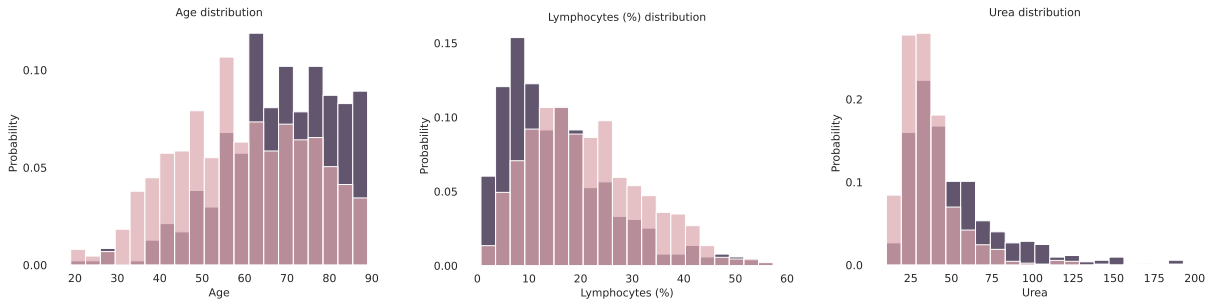
FIGURE A.10 – SHAP summary plot for the *severity_hsl* dataset, showing the impact of the five most important features on model predictions. The plot was generated using the Random Forest algorithm on a balanced version of the dataset. The most significant feature is *age*, with higher values lead the model to predict the instance as *severe*. Dark purple colors represent higher levels of the blood tests, while light purple and pink colors denote lower levels. For some features higher levels drive prediction toward the *severe*, while for other features, the opposite trend is observed.

The most important feature is *age*, where higher values increase the prediction likelihood of the *severe* class, while lower values increase the likelihood of the *non-severe*. High values in *lymphocytes percentage* (LYM (%)) and *eosinophils percentage* (EOS (%)) shift the prediction towards the *non-severe* while higher *c-reactive protein* (CRP) and *urea* levels, move the prediction towards the *severe* class.

To better understand the distribution of the five most important features within each class, Figure A.11 presents the histogram of feature values for each class separately. The histograms are normalized using the probability density function to account for the imbalanced classes. There is noticeable overlap across a wide range of values for all features.

Within the features *age* (A.11a) and *lymphocytes percentage* A.11b), a distinct pattern can be observed despite the overlap in distributions. The *severe* class tends to concentrate on one side of the distribution, while the *non-severe* class spans a wide range of values.

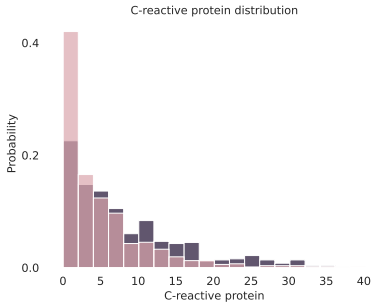
For *urea* levels, the *non-severe* class shows values concentrated on one side of the



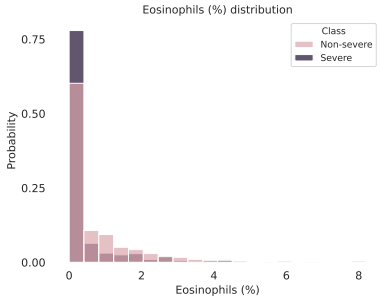
(a)

(b)

(c)



(d)



(e)

FIGURE A.11 – Histograms showing the distribution of the five most important features within each class. Distributions were normalized using the probability density function to account for class imbalance. The *severe* class is depicted in dark purple while the *non-severe* is shown in pink, with areas of overlap represented in light purple. While there is significant classes overlap across all values, certain patterns can still be observed, such as higher *age* in the *severe* class and the predominance, although not exclusivity, of lower *c-reactive protein* levels in the *non-severe* class.

distribution, whereas the *severe* class exhibits a broader range of values.

The other two selected features primarily present values around zero. In the case of *eosinophils percentage*, the *severe* class concentrates values around zero, while the *non-severe* class dominates values greater than 0.2. Conversely, for CRP levels, the *non-severe* class dominates values closer to zero, with significant overlap in other value ranges.

Validating with external data

A secondary dataset was assembled with data from the HBP, this dataset contains the same features and was employed as a test set for models trained with the *severity_hsl* dataset. This experiment aims to assess whether models trained on data from one hospital can maintain the same level of accuracy when applied to another dataset. This serves as a external validation for the trained models and helps to evaluate their broader applicability. The table A.11 presents the predictive performance of various ML models trained on the *severity_hsl* dataset and tested on the *severity_hpb* dataset.

TABLE A.11 – Average predictive performance and standard deviation of models trained on the *severity_hsl* dataset and tested on the *severity_hbp* test dataset. Class imbalance on training data was addressed through an under-sampling technique. The best value for each metric is highlighted in bold. AUC: Area Under the Curve; RCL: Recall; PRC: Precision. SVC: Support Vector Classifier; RF: Random Forest; GB: Gradient Boosting; LR: Logistic Regression; BAG: Bagging; MLP: Multilayer Perceptron.

	$\mathbf{SVC_{linear}}$	$\mathbf{SVC_{rbf}}$	**RF**	**GB**	**LR**	**BAG**	**MLP**
AUC	0.753	0.744	0.780	**0.785**	0.769	0.715	0.742
$\mathbf{RCL_1}$	0.642	0.672	0.672	**0.701**	0.687	0.567	0.687
$\mathbf{RCL_0}$	0.695	0.678	0.686	**0.771**	0.720	0.695	0.703
$\mathbf{PRC_1}$	0.544	0.542	0.549	**0.635**	0.582	0.514	0.568
$\mathbf{PRC_0}$	0.774	0.784	0.786	**0.820**	0.802	0.739	0.798

Models performance achieve only intermediate values, specially low concerning the precision in the positive class. The Gradient Boosting algorithm achieved the highest AUC score (0.785), recall and precision for the positive class (0.701 and 0.635) and recall and precision for the negative class (0.771, 0.820), demonstrating superior model performance in distinguishing between classes.

A.3.4 Discussion

From the data available in the *Covid-19 Data Sharing/BR* repository, we assembled two datasets to predict the severity of patients with covid-19. We described our decisions during data preprocessing and provided an overview of some ML models created from these datasets. We also explored the possibility of generalizing models across hospitals.

In our initial experiments, using cross-validation with the *severity_hsl* dataset, we found sensitivity and specificity levels under 0.7, which may be considered inadequate for practical application. This could be due to the complexity of the task. Predicting severity

in hospitalized patients is challenging, as they already have in some level an aggravated condition, making it harder to differentiate severe from non-severe cases.

Similar studies aimed at predicting severity in hospitalized patients have shown better predictive performance than our models, with AUC scores of up to 0.92 in (KIM *et al.*, 2020) and around 0.90 (FERNANDES *et al.*, 2021). However, these studies included more clinical variables not available in our cohort, such as the BRADEN scale and initial symptoms like dyspnea and body temperature.

Clinical information has significant predictive power for covid-19 clinical aggravation, as confirmed in our previous study with a database containing only symptoms and comorbidities. Therefore, we anticipate that integrating clinical information with lab test results can improve predictive performance. In the absence of this information, the results are suboptimal, though they reveal correlations between severity and routine lab tests conducted during the first day of hospital admission.

The cross-check study, using *severity_hbp* as test set, reveals values similar or superior to the performance observed during cross-validation within the *severity_hsl* dataset. Two factors likely contributed to the generalization across datasets: both datasets had similar percentages of severe cases and comparable age ranges. However, there was a notable difference in the percentage of male patients. This disparity may impact the generalizability of the trained model, as it could contain a more gender-biased sample, and covid-19 tends to affect male patients more severely.

Overall, these findings suggest that models trained on data from one hospital can generalize to other health centers, possibly due to the similar populations served by these hospitals, which share the same city and similar social and financial conditions.

A.4 Predicting outcomes from disease notification

The Brazilian Information System for Notifiable Diseases (Sinan) relies primarily on the reporting of cases involving diseases and conditions outlined in the national list of notifiable diseases. The objective of the Sinan is to collect, transmit, and disseminate data routinely generated by the Epidemiological Surveillance System of the three levels of government (Federal, State, and Municipal) through a computerised network. Professionals designated in each level of government, who are directly and indirectly involved in the notification and investigation of disease cases, are the intended users of this system.

One such disease is dengue, which is caused by the Aedes mosquito. We have compiled datasets containing information on dengue across the national territory. Multiple datasets can be assembled from this database. We present two distinct classification problem: one

aiming to predict death among severe cases, and another aiming to predict hospitalization among all notified cases.

A.4.1 Data source context

Sinan data management relies with the managers of the Unified Health System (SUS - Sistema Único de Saúde) and access authorization to the Sinan system is under the responsibility of the Municipal or State Secretary of Health. They identify the notifying health units and the healthcare professionals responsible for receiving, investigating, and recording notified cases. Notifying units are generally those that provide services under the SUS. Other units such as private hospitals and/or private clinics can be registered in the National Registry of Health Establishments (CNES) as sources of notification.

The Technical Unit of the Information System for Notifiable Diseases (UT-SINAN) is responsible for the federal management of SINAN. After gathering data from all notifiable units, data is made available in batches corresponding to each year and can be accessed through Information Technology Department of the SUS (DATASUS) website. The original databases, initially in .dbc format, were extracted from DATASUS website and then converted to .csv format.

Our primary data base was from the year 2023, while supplementary data from the beginning of the current year were adopted as a test set, with the objective of assessing models generalization.

The data source contains 1,517,551 records of notifications from the year 2023. Notifications are made using a standardized form, resulting in raw data with a total of 121 columns. While some fields are mandatory when reporting a disease occurrence, others can be left incomplete, leading to a substantial number of blank fields. The dengue data source includes information on the following aspects:

1. Personal details such as gender, professional occupation, race, and education;
2. Dates such as birth date, onset of symptoms, notification of the disease, disease progression, hospitalization, and death;
3. Symptoms, including general symptoms, warning signals, and severe signals;
4. Comorbidities;
5. Location details such as state and health unit.

We have meticulously compiled datasets containing information on dengue across the national territory. Our first task is to predict fatalities among severe cases of the disease,

and the second task is to predict which notifications will necessitate hospitalization. For each classification problem, we assembled a dataset for the year 2023 and a secondary dataset with information available from the beginning of 2024, a year marked by increased cases of the disease nationwide (BARCELLOS *et al.*, 2024). In the following sections, we describe the decisions made to preprocess the data, resulting in the creation of four distinct datasets.

A.4.2 Data preprocessing

Both predictive tasks shared some common preprocessing steps, which are briefly described below.

- Removal of columns: Since some columns in the form are not obligatory and are often left unfilled, they contain minimal information and were removed.
- Columns transformation: To prepare the data for the predictive task, certain information formats were transformed. For instance, race was converted into five binary columns corresponding to the categories in the form (white, black, parda, asian, and indigenous). Similarly, state was transformed into binary columns based on the region of the country the state belongs to.
- Imputation of education level: Education level, an indirect indicator of social and financial status, was imputed using the mean across groups sharing the same age, gender, race, and city. If an instance did not belong to any group, it was removed.
- Filtering of age: Age values lower than 12 were removed because the disease on children may have different characteristics compared to adults. Age values over 110 years were also removed as they were considered typos.
- Removal of missing values: Since columns kept in the dataset had a low rate of missing values, rows containing missing data were removed. In way the datasets does not contain missing values.
- Removal of columns with low frequency: Columns such as symptoms, comorbidities, or location within a region, that had low frequency in each dataset were removed.

Besides the described steps, some extra preprocessing tasks were applied to assemble each predictive task, they are described in the following subsections.

A.4.2.1 Dataset 1: *death_dengue_23*

The objective of this classification task is to predict whether a severe case of the disease will result in death within a 40-day period. The information of severity is one of the fields in the notification form. The two classes in this dataset are *survivor* and *dead*. Below, we outline the specific preprocessing steps taken to prepare the data for this task:

- Filtering severe cases: The raw data includes a field classifying the severity of the disease. Only patients with severe dengue were included in this dataset.
- Excluding other outcomes: The dataset was limited to patients who either survived or died from the disease. Outcomes such as "unknown" or death due to other causes were excluded.
- Removal of missing values: Since columns kept in the dataset had a low rate of missing values, rows containing missing data were removed. In way the datasets does not contain missing values.
- Removal of the first sympthoms: in this dataset we want to predict death from the moment of the aggravated condition, in this way, the initial sympthoms were removed.
- Limiting death timeframe: To confirm that severe dengue was the cause of death, patients who died more than 40 days after the beginning of the severe condition were removed from the dataset.

The final dataset contains 1,179 cases and 35 attributes. Of these cases, 45% were categorized as *survivor*, while 55% had the *dead* label.

The same preprocessing methodology was applied to the data from the current year (2024). This dataset will serve as a test set to assess the generalization ability of models trained on the main dataset. The *death_dengue_24* dataset contains 405 cases, with 56% classified as *dead* and 44% classified as *survivor*.

A.4.2.2 Dataset 2: *hospitalization_dengue_23*

This dataset was assembled to predict the need for hospitalization based on the initial symptoms declared in the notification form. Both severe and non-severe dengue cases were considered for this task. The two labels in the dataset are *hospitalized* and *non-hospitalized*. To create the dataset, several preprocessing steps were carried out, which are described briefly below:

- Exclusion of unknown hospitalization: The dataset includes only patients with known hospitalization status, either *hospitalized* or *non-hospitalized*.
- Filtering of outcomes: The dataset includes only patients who survived or died from dengue. Cases with outcomes such as "unknown" or deaths from other causes were excluded.
- Exclusion of severe *non-hospitalized* patients: Patients classified as severe cases or who died from the disease but did not have a hospitalization record were removed.
- Limitation on hospitalization timeframe: To ensure that dengue was the reason for hospitalization, patients hospitalized more than 15 days after the onset of symptoms were excluded from the dataset.
- Removal of duplicates: Duplicate rows were removed from the dataset.
- Removal of row with less information: Since the database contains a large number of notifications we removed those with less than five fields completed (including symptoms and comorbidities). In this way we could reduce the size of the dataset generated and keep those notifications that contain more information.
- Reducing the *non-hospitalized* class: At this stage, we still had a very large dataset (almost 400,000 cases) with a highly imbalanced distribution (less than 0.5% in the positive class). To address this imbalance, we reduced the size of the negative class by removing instances with identical symptomatic states.

The final dataset contains 29,920 cases, and 37 features. *Hospitalized* patients consist in 47.1% of the cases included and 52.9 % are *non-hospitalized* cases. A large number of cases were included in the dataset as it encompasses both severe and non-severe cases of dengue. For this dataset only the first symptoms were considered, excluding subsequent aggravation signals. The goal is to predict hospitalization based on the initial presentation of the disease.

Using data from the current year (2024), we created a test set to evaluate the generalization of models trained on the main dataset. The *death_dengue_24* dataset was assembled following the same preprocessing steps as the main dataset. This test set contains 14,345 cases, of which 39.7% are infected patients requiring hospitalization, while 60.3% did not require hospitalization.

A.4.3 Datasets description

After preprocessing, four datasets were created. Two of these datasets, compiled with data from 2023, were explored further. These datasets were analyzed for feature distri-

bution and assessed using a cross-validation experiment. To evaluate their generalization ability, they were tested with external data from the current year (2024). The following subsections provide a detailed discussion of data exploration and evaluation.

A.4.3.1 Dataset 1: *death_dengue_23*

This dataset was assembled to predict death among severe dengue cases. To understand how features behave inside each class, Figure A.12 illustrates the distribution of binary columns, including comorbidities, severe signals, race, and location, within each class.

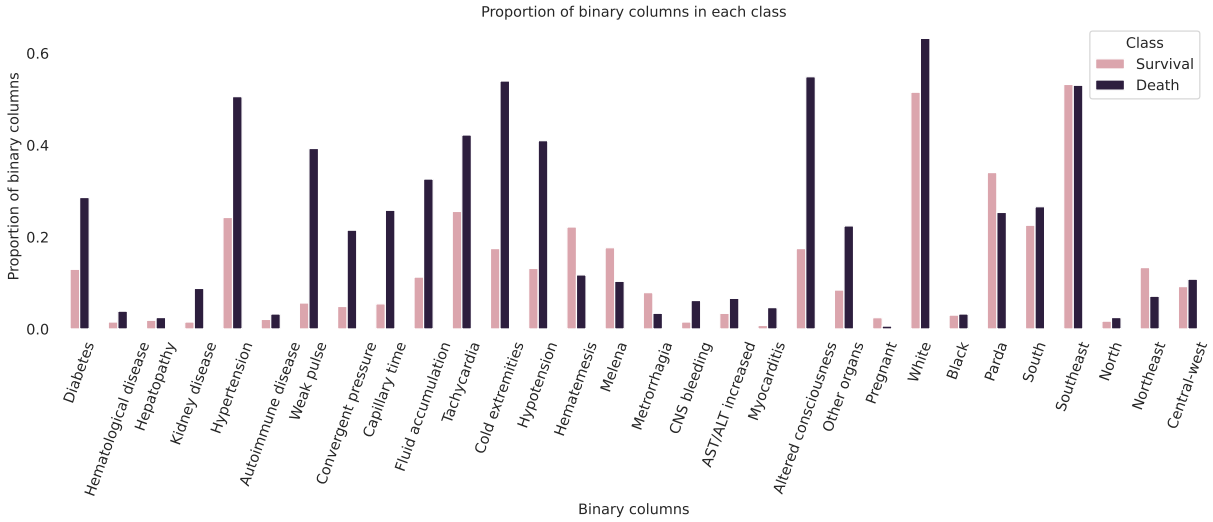


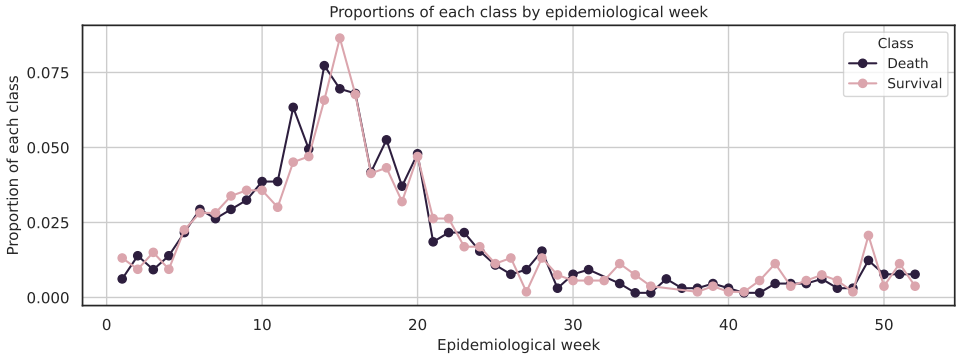
FIGURE A.12 – Distribution of binary columns—comorbidities, severe signals, race, and location—within each class of the dataset *death_dengue_23*. The comorbidities more frequent are *diabetes* and *hypertension* with a prevalence of them in the *dead* class. The severe signals *cold extremities* and *altered consciousness* are the most frequent, also differentiating classes. The racial composition is dominated by the *white* group, and the regional distribution is concentrated mainly in the *south* and *southeast* regions.

In terms of comorbidities, *diabetes* and *hypertension* are the most prevalent, with a significant difference between classes, being notably more common in the *dead* class.

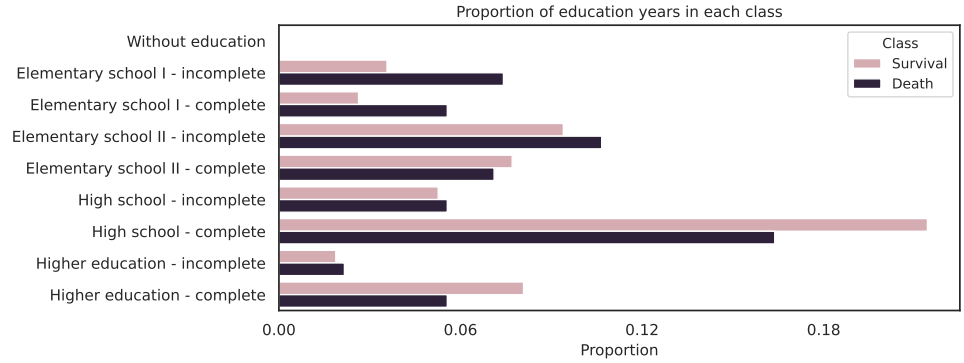
Most severe signals are more prevalent within the *dead* class, with *cold extremities* and *altered consciousness* standing out as the most common and exhibiting the greatest discrepancy in distribution across classes.

Regarding the distribution of race, the dataset is dominated by the *white* group, which is associated with a higher prevalence of *dead* cases. The *parda* population is the second most frequent group, showing a higher prevalence of *survivor cases*.

The data distribution across the country reveals a high concentration of notifications in the *southeast* region, followed by the *south* region, with a small variation in distribution among classes. It is important to note that a high number of notifications does not necessarily indicate a high number of cases, as these regions, known for their higher levels



(a) Proportion of severe cases in each epidemiological week of the *death_dengue* dataset. There is a noticeable peak around week 15. Both classes present similar distributions.



(b) Proportion of education level within each class of the *death_dengue* dataset. The most frequent education level is *High School - complete*. Class *survivor* is predominant inside higher levels of education.

FIGURE A.13 – Distribution of two non-binary features within each class of the *death_dengue* dataset. Top: Proportion of severe cases per epidemiological week in each class. Bottom: Proportion of education levels by class.

of Human Development Index (HDI) (NASCIMENTO *et al.*, 2022), likely have a more robust healthcare system that enables greater notification capacity.

To complete the dataset description, Figure A.13 shows the distribution of two non-binary features. The first plot represents the proportion per class of severe cases in each epidemiological week, the second graph represents the proportion of level of education by class.

Figure A.13a illustrates the number of notifications registered in each week. The highest number of cases occurs around week 15, corresponding to April. This is the end of summer in Brazil and a period of frequent rain, which creates ideal breeding conditions for the Aedes mosquito due to stagnant water. This is typically the time of year with the most cases and consequently most severe cases. The proportion of the classes *dead* and *survivor* appears similar in both classes.

Figure A.13b shows the proportion of education level within each class. The most common education level is *High School - complete*, with most notifications falling under

the *survivor* class. For education levels below *High School - incomplete*, the proportion of *dead* cases surpasses the proportion of *survivor* cases.

Table A.12 presents the average predictive performance and standard deviation of various ML learning models trained on the *death_dengue_23* dataset using a five-fold cross-validation. The metrics assessed include sensitivity, specificity, and AUC.

TABLE A.12 – Average predictive performance and standard deviation of models trained on the *death_dengue_23* dataset, evaluated using five-fold cross-validation. The best value for each metric is highlighted in bold. AUC: Area Under the Curve; RCL: Recall; PRC: Precision; SVC: Support Vector Classifier; RF: Random Forest; GB: Gradient Boosting; LR: Logistic Regression; BAG: Bagging; MLP: Multilayer Perceptron.

	$\mathbf{SVC_{linear}}$	$\mathbf{SVC_{rbf}}$	**RF**	**GB**	**LR**	**BAG**	**MLP**
AUC	**0.869** (0.012)	0.864 (0.012)	0.868 (0.012)	0.856 (0.016)	0.866 (0.013)	0.829 (0.008)	0.839 (0.008)
$\mathbf{RCL_1}$	0.811 (0.018)	0.822 (0.026)	**0.836** (0.035)	0.813 (0.034)	0.802 (0.021)	0.747 (0.027)	0.796 (0.040)
$\mathbf{RCL_0}$	0.767 (0.025)	0.744 (0.026)	0.752 (0.034)	0.761 (0.042)	**0.773** (0.030)	0.763 (0.029)	0.739 (0.022)
$\mathbf{PRC_1}$	0.809 (0.014)	0.797 (0.015)	0.804 (0.019)	0.807 (0.024)	**0.812** (0.016)	0.794 (0.016)	0.788 (0.012)
$\mathbf{PRC_0}$	0.770 (0.011)	0.776 (0.023)	**0.792** (0.033)	0.771 (0.028)	0.763 (0.013)	0.713 (0.018)	0.750 (0.033)

The overall performance of the models is good. The best AUC is presented by Support Vector classifier with a linear kernel (0.869). Random Forest holds both the highest recall in the positive class and precision in the negative class (0.836 and 0.792). On the other hand, Logistic Regression shows the best Recall for the *survival* class and precision for the *death* class (0.773 and 0.812).

Figure A.14 presents the SHAP summary plot of the five most important features for the *death_dengue_23* dataset. SHAP values were computed using the Random Forest algorithm.

The most significant feature is *age*, with both high and low values strongly influencing predictions. High *age* values are associated with predictions of dead cases, while low *age* values lead to opposite predictions.

The remaining important features are all binary, which simplifies interpretation. Their presence, indicated by purple dots, consistently influences predictions towards dead cases, while their absence, indicated by pink dots, has a lesser impact in the opposite direction. All the binary features selected, *alterated conciousness, weak pulse, cold extremities* and *hypotension*, are severe signals of the disease.

To better understand the selected features, Figure A.15 presents two plots. The first plot is the shows the co-occurrence among the chosen severe signals, as well as their co-occurrence with the target feature. The second plot illustrates the distribution of *age*

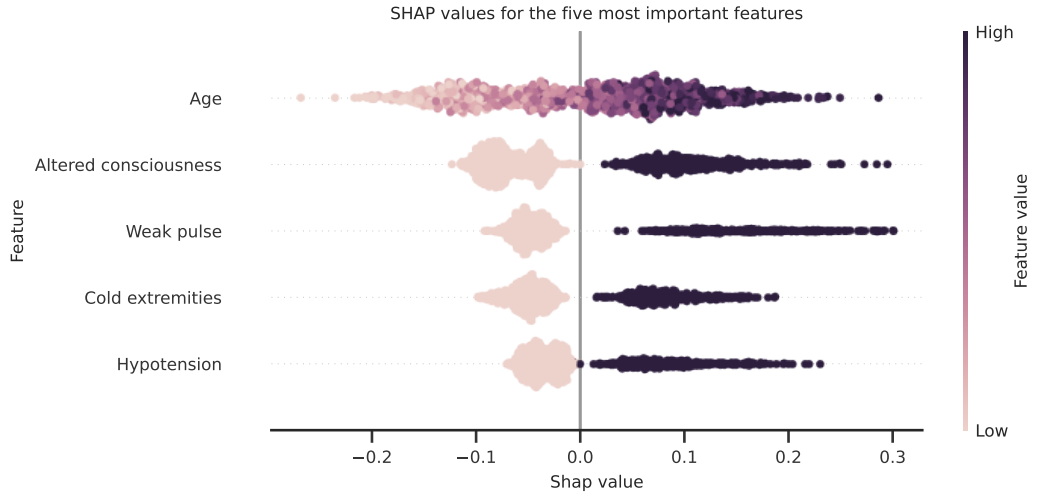


FIGURE A.14 – SHAP summary plot for the five most important features in the *death_dengue_23* dataset. The SHAP values were calculated using the Random Forest algorithm. The plot highlights *age* as the most significant feature, with high and low values having a strong influence on predictions. The other important features are binary, their presence, represented by dark purple dots, impacts instances predictions as *dead* cases, while their absence has a lower impact in predicting the instances as *survivor* cases.

inside each class.

In Figure A.15a, we observe co-occurrence greater than 0.3 between *death* and each of the selected severe signals. The plot shows the strongest co-occurrence (0.4) between *weak pulse* and *cold extremities*. Another pair of severe signals with a notable co-occurrence is *weak pulse* and *hypotension* (0.36).

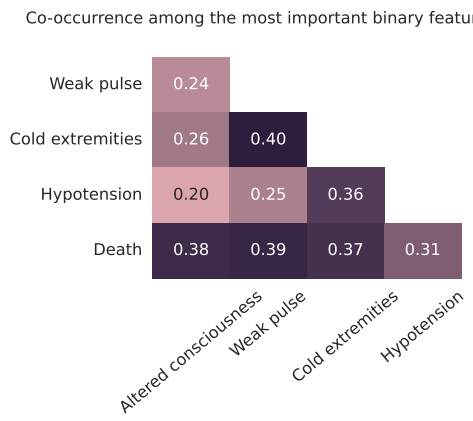
Figure A.15b shows the distribution of age within each class. Although there is overlap across the entire age range, a clear trend is evident: the *dead* class, shown in dark purple, dominates the higher end of the *age* distribution, while the *survivor* class, depcted in pink, is more prevalent at the lower end.

Validating with external data

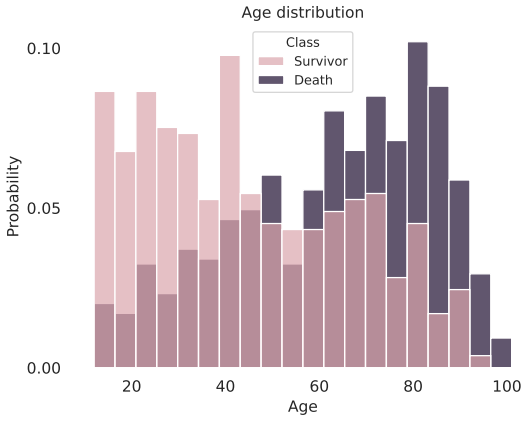
As outlined earlier (refer to Section A.4.2), two datasets were assembled using identical preprocessing steps to predict mortality among severe cases of dengue. The primary dataset consists of data from 2023, while the test set was created using data from the current year, 2024. We now present an evaluation of models trained on the primary dataset and tested on the secondary set.

Table A.13 provides a summary of the two datasets compiled. The table includes the number of patients, the number of features considered, the percentage of deceased patients, the percentage of male patients, and age statistics.

The proportion of *dead* patients is comparable in both datasets, with 54.9% in 2023 and 56.5% in 2024, suggesting that mortality rates among severe patients have remained relatively stable across the two years. Similarly, the percentage of male patients and



(a) Co-occurrence between *death* and each of the selected severe signals in the *death_dengue_23* dataset. Co-occurrence values exceed 0.3 for all selected signals. The plot also highlights a correlation of 0.4 between *weak pulse* and *cold extremities*, and a correlation of 0.36 between *weak pulse* and *hypotension*.



(b) Distribution of *age* within each class in the *death_dengue_23* dataset. There is a large overlap between the two classes, but the *dead* class, shown in purple, dominate the right side with high *age* values. The opposite trend is presented by the *survivor* class, represented in pink.

FIGURE A.15 – Exploration of the five most important features within *death_dengue_23* dataset. Left: Co-occurrence between binary features. Right: Distribution of *age* by class.

TABLE A.13 – Summary of the datasets assembled from the data base of to predict mortality among severe cases of dengue, including number o patients, number of features, percentage of dead cases, percentage of male patients and age statistics

Year	**♯ Patients**	**♯ Features**	**% Death**	**% Female**	**Mean age [min - max]**
2023	1,179	35	54.9	53.7	54.6 [12 - 101]
2024	405	35	56.5	52.1	55.8 [12 - 105]

age statistics are closely matched, with males representing the majority in both datasets (around 55%). The median age in both datasets is approximately 54 years, with comparable age ranges. These consistencies indicate that the models have potential for generalization to data from the following year.

The table A.14 presents the average predictive performance of multiple ML models trained on the *death_dengue_23* dataset and evaluated on the *death_dengue_24* dataset.

TABLE A.14 – Average predictive performance of models trained on the *death_dengue_23* dataset and tested on the *death_dengue_24* dataset. The best value for each metric is highlighted in bold. AUC: Area Under the Curve; RCL: Recall; PRC: Precision; SVC: Support Vector Classifier; RF: Random Forest; GB: Gradient Boosting; LR: Logistic Regression; BAG: Bagging; MLP: Multilayer Perceptron.

	$\mathbf{SVC_{linear}}$	$\mathbf{SVC_{rbf}}$	**RF**	**GB**	**LR**	**BAG**	**MLP**
AUC	0.849	0.839	0.839	0.840	**0.852**	0.775	0.801
$\mathbf{RCL_1}$	0.769	**0.795**	0.799	0.782	0.747	0.642	0.729
$\mathbf{RCL_0}$	**0.784**	0.733	0.739	0.761	**0.784**	0.761	0.756
$\mathbf{PRC_1}$	**0.822**	0.795	0.799	0.810	0.818	0.778	0.795
$\mathbf{PRC_0}$	0.723	0.733	**0.739**	0.728	0.704	0.620	0.682

The models show good overall performance. The best AUC value was achieved by the Logistic Regression (0.852). Random Forest demonstrated the best precision for the *survival* class (0.739), while Support Vector Classifier with linear kernel holds the best precision for the positive class (0.822). Finally, the Support Vector Classifier presented the highest recall for the *death* class (0.795) and SVC with linear kernel the best recall for the negative class (0.784, the same value achieved by Linear Regression).

A.4.3.2 Dataset 2: *hospitalization_dengue_23*

The objective of this classification task is to predict the need for hospitalization using information available at the time of disease notification. Figure A.16 examines the distribution of binary features within each class, including symptoms, comorbidities, race, and country region.

The distribution proportions of binary columns are quite similar between the two classes. However, certain symptoms distinguish the classes: *conjunctivitis*, *arthritis*, and *lace* are more prevalent in the *non-hospitalized* class, while *fever*, *myalgia*, and *headache* are more common in the hospitalized class, though also frequently observed in *non-hospitalized* patients.

Regarding comorbidities, they are more prevalent in the *non-hospitalized* class, which is unexpected. *White* patients are more common in the *non-hospitalized* class, while *parda* (mixed-race) patients are more common in the hospitalized class. Finally, individuals living in the *South* and *Southeast* regions are more prevalent in the dataset, with most belonging to the non-hospitalized class. Conversely, other regions of Brazil have smaller

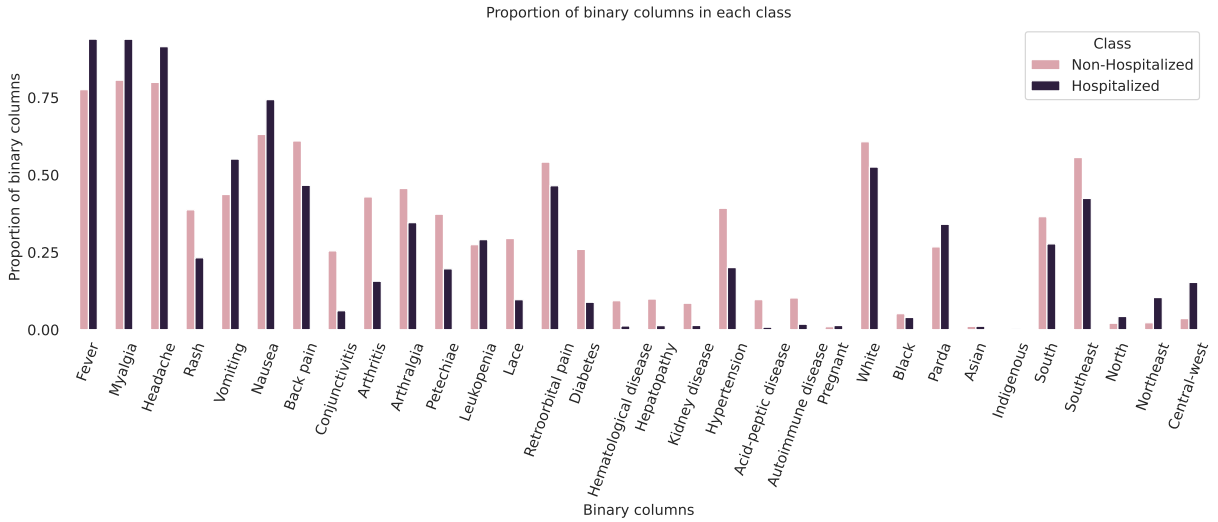


FIGURE A.16 – Proportion of binary features within the two classes the *hospitalization_dengue_23* dataset. Including: symptoms, comorbidities, race and country region. Features such as *conjunctivitis*, *arthritis*, and residence in the *Diabetes* region are more frequent inside the *non-hospitalized* class. In contrast, *fever*, *myalgia*, and *headache* are some of the features more common the *hospitalized* class.

representations and are dominated by *hospitalized* patients.

The *hospitalization_dengue_23* dataset includes two non-binary features that are examined in the next. Figure A.17 shows the distribution of notifications across epidemiological weeks, and the education levels.

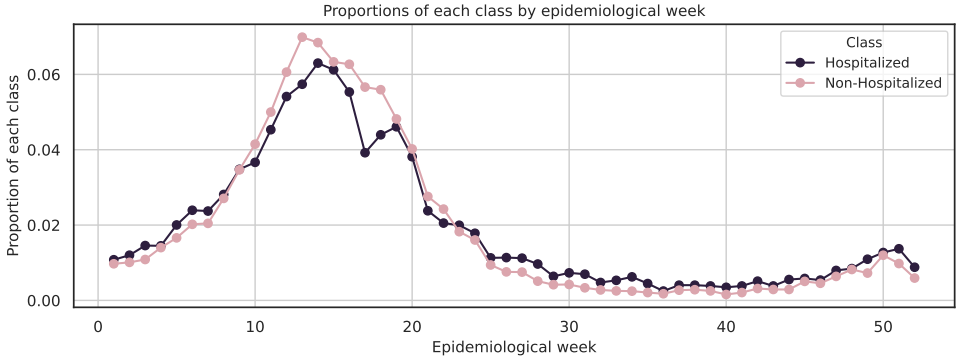
The proportions of classes in each week, as illustrated in Figure A.17a, follow an expected pattern. Notifications increase during the summer months when Brazil, a tropical country, experiences frequent rain. The peak of notifications occurs around April, with both classes showing similar distributions.

The proportions of education levels within each class. Both classes present the same proportions of the distributions by class. The majority of notifications fall within the *high school - complete* level of education. Most notifications with higher education levels are concentrated in the Southeast region, likely reflecting the region's higher social and financial conditions and consequently the presence of healthcare infrastructure, as mentioned earlier.

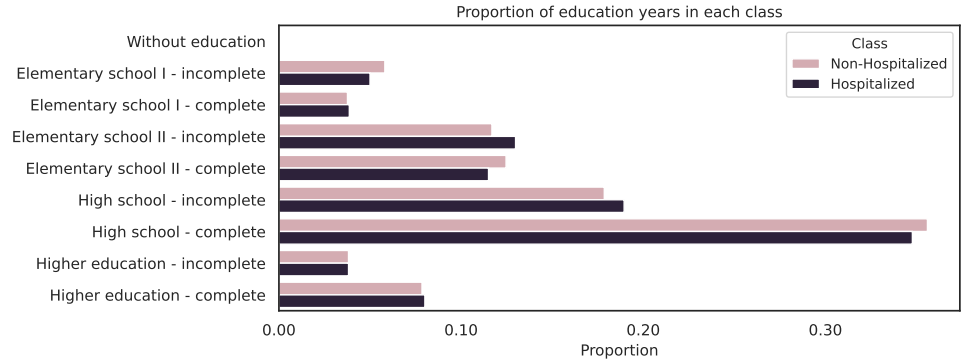
Table A.15 presents the mean and standard deviation of various ML models evaluated using five-fold cross-validation on the *hospitalization_dengue_23* dataset.

Overall, the models exhibit strong predictive performance. The Multilayer Perceptrom algorithm holds the best AUC (0.923). Gradient Boosting presents the highest recall in the positive class (0.832) and precision in the negative class (0.856). Finally the best recall for *non-hospitalized* patients is showed by Support Vector Classifier with RFB kernel (0.912) as well as the best precision for the *hospitalized* class (0.893).

The five most important features for predicting hospitalization in the *hospitalization_-*



(a) Proportion of notifications by class in each epidemiological week in the *hospitalization_dengue_23* dataset. The proportion is the same between the two classes along the year. There is rising in number of notifications from the first week with a maximum around week fifteen. It starts rising again in the last weeks of the year.



(b) Proportion of education level within each class of the *hospitalization_dengue_23* dataset. The most frequent education level is *High School - complete*. Both classes are equally present in each level of education

FIGURE A.17 – Distribution of two non-binary features within each class of the *hospitalization_dengue_23* dataset. Top: Proportion of notifications per epidemiological week in each class. Bottom: Proportion of education levels by class.

TABLE A.15 – Average predictive performance and standard deviation of models trained on the *hospitalization_dengue_23* dataset, evaluated using five-fold cross-validation. The best value for each metric is highlighted in bold. AUC: Area Under the Curve; RCL: Recall; PRC: Precision; SVC: Support Vector Classifier; RF: Random Forest; GB: Gradient Boosting; LR: Logistic Regression; BAG: Bagging; MLP: Multilayer Perceptron.

	$\mathbf{SVC_{linear}}$	$\mathbf{SVC_{rbf}}$	**RF**	**GB**	**LR**	**BAG**	**MLP**
AUC	0.919 (0.002)	**0.935** (0.002)	0.923 (0.003)	0.928 (0.003)	0.918 (0.002)	0.896 (0.003)	0.923 (0.003)
$\mathbf{RCL_1}$	0.816 (0.005)	0.827 (0.007)	0.798 (0.004)	**0.832** (0.004)	0.828 (0.005)	0.771 (0.007)	0.827 (0.009)
$\mathbf{RCL_0}$	0.886 (0.004)	0.912 (0.003)	**0.914** (0.003)	0.888 (0.004)	0.866 (0.005)	0.896 (0.004)	0.874 (0.012)
$\mathbf{PRC_1}$	0.864 (0.004)	**0.893** (0.003)	0.891 (0.003)	0.869 (0.004)	0.847 (0.005)	0.868 (0.005)	0.854 (0.011)
$\mathbf{PRC_0}$	0.844 (0.004)	0.855 (0.005)	0.835 (0.003)	**0.856** (0.003)	0.850 (0.004)	0.815 (0.005)	0.851 (0.005)

dengue_23 dataset were identified using SHAP values. These values were calculated with the Random Forest algorithm. Figure A.18 shows the SHAP summary plot, providing a visual representation of how different feature values impact the predictions.

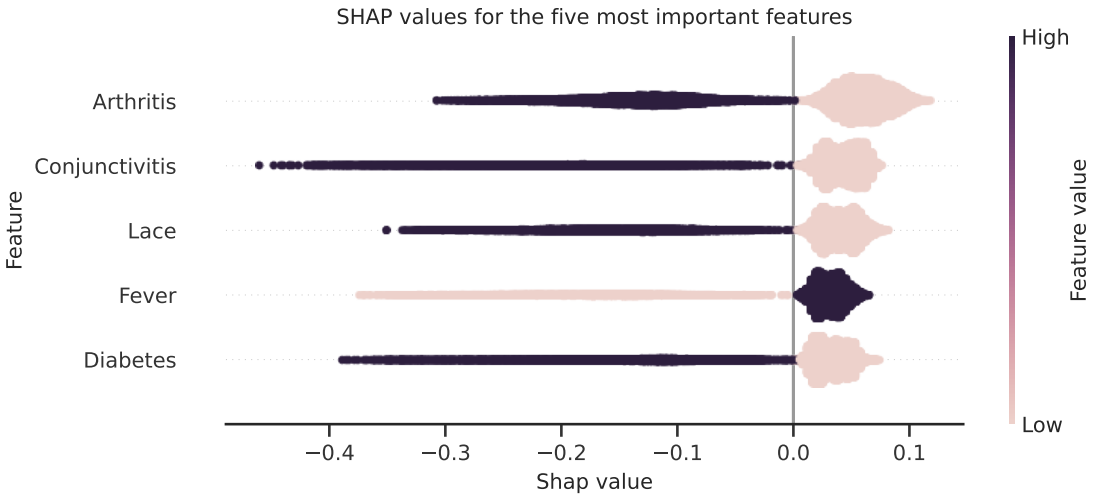


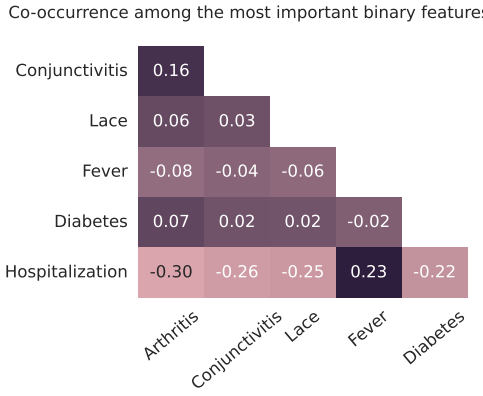
FIGURE A.18 – SHAP summary plot of the five most important features identified in the *hospitalization_dengue_23* dataset. SHAP values were calculated from the Random Forest algorithm. *Artritis* stands out as the most significant feature, with its presence strongly linked to *non-hospitalization* predictions. The presence of other selected feature also has strong influence towards *non-hospitalization*. The exception is fever that when is absent influence predictions to *non-hospitalization*.

In binary features, dark purple dots indicate the presence of a symptom or comorbidity. All selected features have a strong influence on models prediction to *non-hospitalized*, with minimal impact on hospitalization, evident from the dots clustering near the y-axis. The presence of *arthrits*, *conjunctivitis*, *lace* and *diabetes* as well as the absence of *fever*, inclines the model towards predicting the instance as *non-hospitalized*.

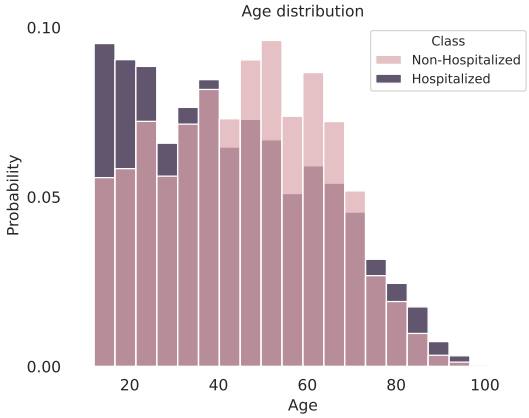
The unexpected prediction trends observed, where the presence of symptoms and even comorbidities pull predictions towards the negative class, requires consideration of factors underlying dengue virus infection. Dengue presents a broad clinical spectrum, from asymptomatic cases to severe and potentially fatal manifestations. The initial presentation typically includes a sudden onset of symptoms including high fever.

While some progress to severe disease, marked by alarm signals preceding hemorrhagic manifestations, it's noteworthy that seeking healthcare with symptoms doesn't always result in hospitalization. The decision to hospitalize primarily hinges on platelet volume, with hospitalization recommended when it drops below <20,000/mm3.

For a deeper understanding of the five features identified as the most important for predicting hospitalization in the *hospitalization_dengue_23* dataset, we present Figure A.19. The first plot inside the figure illustrates the co-occurrence of the binary features and the target class *hospitalization*. The second image shows the proportion of *age* within each class. The plot is normalized using the probability density function to account for



(a) Co-occurrence among the most important binary features selected in the *hospitalization_dengue_23* dataset. The plot highlights a intermediate co-occurrence among the selected binary features and the target variable. The highest co-occurrence level is observed between *hospitalization* and *arthritis*.



(b) Proportion of *age* by class within *hospitalization_dengue_23*, distributions were normalized through the probability density function. The plot highlights the overlap of the *age* across the whole range of values. There a slight predominance of *hospitalized* cases in the advanced years and also among the most young notifications. The *non-hospitalized* class dominate range 40-70 years old.

FIGURE A.19 – Exploration of the five most important features within *hospitalization_dengue_23* dataset. Left: Co-occurrence between binary features. Right: Proportion of *age* by class.

class imbalance.

The first plot in Figure A.19a displays an intermediate co-occurrence between the target feature, *hospitalization*, and the most significant features selected. The highest co-occurrence level is between *hospitalization* and *artritis* (-0.30), while the co-occurrence of *hospitalization* and other binary features are around the same level. .

The proportion of *age* presented in Figure A.19b indicates a wide range of overlap across classes. On the right side of the plot, there is a predominance of the *hospitalized* class, suggesting a higher probability of hospitalization among older individuals, the same happens with the youngest group. While *non-hospitalized* individuals slightly dominate the *age* range of 40-70 years old group, although with a higher presence of *hospitalized* individuals as well.

The relation between *age* and *hospitalization* mirrors findings in the disease literature. As BURATTINI *et al.* (2016) reports, the percentage of hospitalizations per age group in Brazil, from 2000 to 2014. There is not a linear trend akin to covid-19, where older individuals face a higher risk of severe cases. Instead, a notable number of hospitalizations among children (which are not included in our study) precede those aged over 65 years (12.1%). Following this, the age groups of 11-20 years (8.1%) and 51-65 years (7.6%) also show significant rates of hospitalization.

Validating with external data

To predict hospitalization we also assembled a secondary dataset, with the data available concerning the year 2024. This dataset will be adopted as a test set to assess the generalization of models trained in the primary dataset, composed with data of the 2023 year.

Next, Table A.16 presents a comparison across the two dataset *hospitalization_dengue_23* and *hospitalization_dengue_24*, concerning the number of notifications included, number of features, percentage of hospitalized patients, percentage of male patients and age statistics.

TABLE A.16 – Summary of the datasets assembled from the data base of disease dengue, to predict hospitalization, including number o patients, number of features, percentage of missing values, percentage of hospitalized cases, percentage of male patients and age statistics

Year	♯ Patients	♯ Features	% Hospitalizations	% Female	Mean age [min - max]
2023	29,920	37	47.4	60.5	39.7 [12 - 106]
2024	14,345	37	39.7	60.4	44.2 [12 - 99]

The percentage of hospitalized cases is slightly lower in 2024 (39.7%) compared to 2023 (47.4%). The proportion of female patients remains consistent across both datasets, and the patients mean *age* and interval present similar values. These similarities between the two datasets make it easier for models to generalize effectively.

Table A.17 presents the cross-check assessment results. Models were trained on a balanced version of the *hospitalization_dengue_23* dataset and tested on the *hospitalization_dengue_24* dataset.

TABLE A.17 – Average predictive performance and standard deviation of models trained on the *hospitalization_dengue_23* dataset and tested on the *hospitalization_dengue_24* dataset. The best value for each metric is highlighted in bold. AUC: Area Under the Curve; RCL: Recall; PRC: Precision; SVC: Support Vector Classifier; RF: Random Forest; GB: Gradient Boosting; LR: Logistic Regression; BAG: Bagging; MLP: Multilayer Perceptron.

	$\mathbf{SVC_{linear}}$	$\mathbf{SVC_{rbf}}$	RF	GB	LR	BAG	MLP
AUC	0.843	**0.878**	0.872	0.865	0.838	0.844	0.858
$\mathbf{RCL_1}$	0.835	0.842	0.815	0.837	**0.847**	0.793	0.860
$\mathbf{PRC_0}$	0.704	0.738	**0.764**	0.723	0.678	0.760	0.690
$\mathbf{PRC_1}$	0.650	**0.679**	0.694	0.665	0.633	0.685	0.646
$\mathbf{PRC_0}$	0.866	0.877	0.862	0.871	0.871	0.848	**0.882**

Random Forest and Bagging did not achieve the same level of performance observed during cross-validation. Despite this, most metrics in the external evaluation indicate good predictive performance. The exception is the the precision for the positive class that shows predictive that are not satisfactory.

The Support Vector Classifier with RFB kernel demonstrated the best performance in terms of AUC (0.878) and precision in the *hospitalized* class (0.679). Random Forest achieved the highest recall in the negative class, while logistic regression showed the best

recall in the positive class. Additionally, the Multilayer Perceptron demonstrated the highest precision for the *non-hospitalized* class. Overall, the predictive performance is lower than that of predicting death among severe cases (see Table **??**), suggesting that initial symptoms may not be as reliable for predicting hospitalization.

A.4.4 Discussion

From the data available at Sinan, we developed two classification problems related to to the disease dengue. The first predictive task focused on disease progression to death among severe cases, based on reported alarm signals. The second task assessed hospitalization among confirmed cases of the disease, based on initial symptoms. Both cases included information about comorbidites, race, the time of year, the region of the country.

For each dataset, we provided descriptive statistics to gain a better understanding of feature distribution within each class. We observed a peak in notifications, hospitalizations, and deaths occurring in April, which aligns with the rainy season in Brazil when the Aedes mosquito lays eggs in stagnant water.

There is a notable presence of cases in the southeast region and among individuals with higher education levels, particularly those with a complete high school education. We believe this region's better infrastructure and robust health system lead to more reported cases.

We also evaluated the performance of several ML models using five-fold cross-validation. Models performed well in predicting deaths and hospitalizations. This trend was also observed when assessing models with external data, using datasets from the following year (2024) to evaluate trained models.

The results underscore the potential of leveraging routinely collected Dengue notification data to inform predictive models, thereby enhancing healthcare systems by forecasting resource allocation needs in specific areas. The effectiveness of these models could be further improved with more consistent and reliable data collection practices. For instance, the database contains columns that are not described in the data dictionary, and the extensive nature of the notification form may reduce the likelihood of complete data entries. Addressing these issues would lead to more robust and accurate predictive models.

A.5 Authorizing specialized care

In Brazil, healthcare is provided free of charge to the entire population through the Unified Health System (SUS). The system's accessibility and the country's large popula-

tion lead to high demand, requiring prioritization of patients and cases. When a patient needs specialized care, general practitioners (GPs) submit a request, which is then assessed for approval or denial based on the severity of the case and the availability of human resources.

Evaluating these requests requires significant time and effort from specialized clinicians, who must review each case and decide whether it warrants attention. Automating part of this process with a ML system capable of authorizing or denying straightforward cases would alleviate the burden on healthcare teams, saving valuable time and effort.

The access to this database was obtained through a partnership with Telessaúde-BR, a research center affiliated with the Postgraduate Program in Epidemiology at the Faculdade de Medicina da Universidade Federal do Rio Grande do Sul. With the assistance of the Telessaúde-BR team, who possess in-depth knowledge of the data, a large dataset was assembled from the raw database. The classification task involves predicting which patients would be authorized to receive specialized care.

A.5.1 Data source context

The primary goal of data collection was to automate the process of authorizing or denying requests for specialized care, ultimately aiming to support clinical decision-making and improving the efficiency of the public healthcare system. The data source consists of real-world data, including requests for specialized care submitted by GPs working for the SUS. These requests were reviewed and evaluated by a team of specialists, and the evaluations are also recorded in the database.

The dataset contains descriptions of the clinical condition in Portuguese, alongside patient gender, year of birth, the type of the required specialized care, the Internacional Disease Classification (IDC) and the final decision made by a clinician, which can either be an authorization or denial. Clinical evaluations were performed by doctors specialized in the type of care requested by the GP. Requests may be denied if the clinical condition does not necessitate specialized care and can be managed by the GP or if there is insufficient information provided to justify the request.

The raw data was preprocessed to create a dataset suitable for a classification problem. In the following subsection, we describe the preprocessing steps involved.

A.5.2 Data preprocessing

The clinical dataset underwent various preprocessing steps to prepare it for analysis. They are briefly described next.

- Standardization: The textual data in the clinical descriptions was standardized by converting it to lowercase and removing stopwords.
- Removing incomplete requests: Patients with clinical descriptions of ten words or fewer were excluded from the dataset to focus on cases with sufficient clinical detail, thus ensuring the robustness and reliability of the analysis.
- Filtering ICD: To create a homogeneous dataset, we filtered the most frequent ICDs, selecting twenty of them for inclusion.
- Columns Transformation: Initially, the dataset encompassed various types of International Classification of Diseases (ICD) codes. To streamline the data for predictive tasks, we generated binary columns, each corresponding to a specific specialization present in the database. Additionally, we created a column indicating the number of words in the clinical description.
- Text transformation: the clinical descriptions were transformed into numerical vectors using a pre-trained BERT model Brazilian Portuguese. The adopted pre-trained model is a sentence-transformers model and maps sentences and paragraphs to a 1024 dimensional dense vector space (SOUZA *et al.*, 2020).

We also created a column with patient *age* and kept the feature *sex* available in the raw data.

In this data source that is none information about period of time. In this way it was not possible to split a sample to compose a test set, based on a chronological criteria. For this reason, we adopted a random split, stratified by class, selecting a tenth of instances to compose the test set. The final dataset contains 1,047 features and 19,355 instances, 56.7% being authorized requests and 43.3% denied requests. The test set contains 2,151 examples.

A.5.3 Describing the dataset

The *sus_authorization* dataset was assembled to forecast the authorization of specialized care within the Brazilian Unified Health System (SUS). It includes a range of features, initially represented as textual descriptions of clinical conditions, which were subsequently converted into numerical representations. Additionally, the dataset contains demographic information such as gender and age, and the ICD code corresponding to the patient's condition.

To explore the relationship between the ICD codes and the target feature, Figure A.20 illustrates the distribution of requests across each class in the dataset. For certain ICDs,

a notable disproportion between the two classes is evident. This observation suggests that while some diseases consistently require specialized care, there are conditions that admit varying degrees of severity.

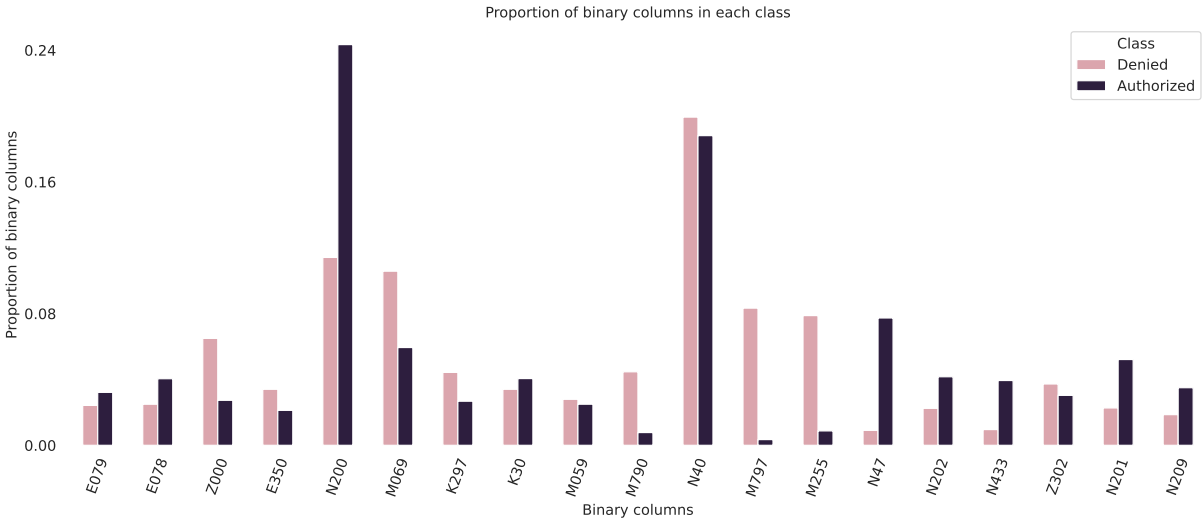


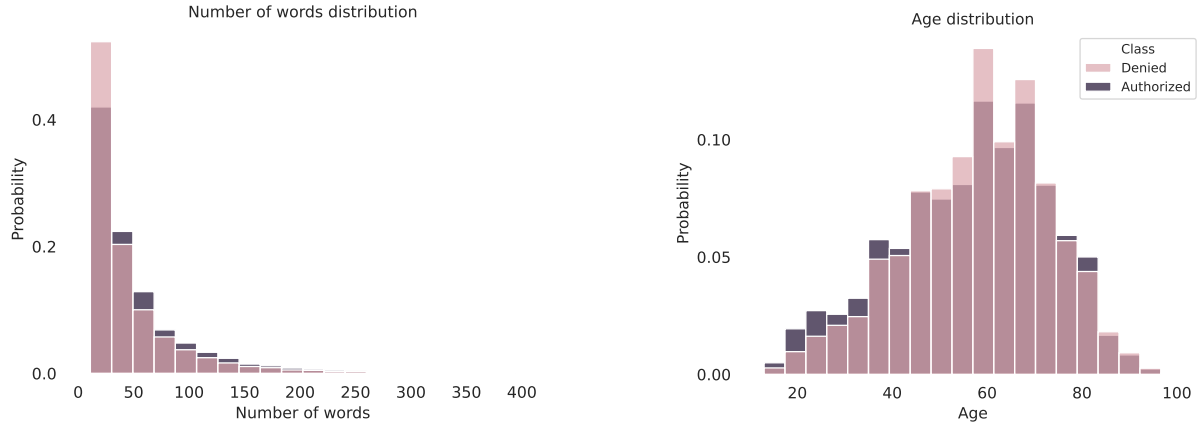
FIGURE A.20 – Distribution of binary columns — International Disease Codes within the *sus_authorization* dataset. Among the IDCs present, two are more prevalent: *N200* and *N40*. While some IDCs exhibit a similar proportion of acceptances and denials, others demonstrate unequal distributions.

The *gender* distribution remains consistent across the two classes, with male patients representing 57.4% of authorized requests and 55.9% of denied requests. As a substantial portion of the features correspond to the numerical transformation of clinical conditions, we further examined the distribution of *age* and *number of words*. Figure A.21 illustrates the distribution of these two variables within each class.

In Figure A.21a, a similar distribution of classes can be observed in relation to the number of words. The *denied* class predominantly occupies the range of 10 to 25 words, although a significant number of *authorized* cases also fall within this word count range. Additionally, the proportions of *age* in Figure A.21b by class also reveals a considerable overlap between the two categories. The *authorized* class slightly predominates in the *age* range under 40 years old, while there is a slight prevalence of *denied* requests in the *age* range of 50-70 years old. The two features appear to offer limited discriminatory power between the two classes.

Table A.18 presents the mean and standard deviation performance of diverse ML models in the *sus_authorization* dataset. The performance was measured in a five-fold cross-validation strategy.

The models exhibit strong performance across various metrics, with notable exceptions in recall for the negative class, which falls below 0.7. This suggests that the models struggle to accurately identify nearly half of the *denied* cases. Among the classifiers, the Support Vector Classifier stands out with the highest AUC value (0.798) and precision for both



(a) Proportion of the feature *number of words* within the *sus_authorization* dataset. Classes overlap across the complete range of values, with a predominance of *Denied* cases associated within the shorter descriptions.

(b) Proportion of *age* by class within *sus_authorization* dataset. The plot highlights the great overlap between the two classes. There is a small predominance of *authorized* cases in the lower ages, and of *denied* cases in age range of 50 - 70 years old.

FIGURE A.21 – Distribution of two non-binary features within each class of the *sus_authorization* dataset. Distributions normalized through the density probability function. Left: Proportion *Number of words* by class. Right: Proportion of *age* in each class.

TABLE A.18 – Average predictive performance and standard deviation of models trained on the *authorization_sus* dataset, evaluated using five-fold cross-validation. The best value for each metric is highlighted in bold. AUC: Area Under the Curve; RCL: Recall; PRC: Precision. ; SVC: Support Vector Classifier; RF: Random Forest; GB: Gradient Boosting; LR: Logistic Regression; BAG: Bagging; MLP: Multilayer Perceptron.

	$\mathbf{SVC_{linear}}$	$\mathbf{SVC_{rbf}}$	**RF**	**GB**	**LR**	**BAG**	**MLP**
AUC	0.777 (0.007)	**0.798** (0.007)	0.744 (0.006)	0.776 (0.007)	0.782 (0.007)	0.719 (0.004)	0.746 (0.006)
$\mathbf{RCL_1}$	0.795 (0.007)	0.820 (0.007)	0.804 (0.003)	**0.830** (0.006)	0.790 (0.007)	0.682 (0.007)	0.726 (0.016)
$\mathbf{RCL_0}$	0.607 (0.009)	0.610 (0.014)	0.549 (0.014)	0.548 (0.010)	0.617 (0.005)	**0.637** (0.005)	0.624 (0.011)
$\mathbf{PRC_1}$	0.726 (0.006)	**0.734** (0.008)	0.700 (0.007)	0.706 (0.005)	0.730 (0.004)	0.711 (0.003)	0.716 (0.005)
$\mathbf{PRC_0}$	0.693 (0.009)	**0.721** (0.010)	0.681 (0.008)	0.711 (0.009)	0.691 (0.007)	0.604 (0.005)	0.635 (0.012)

classes (0.734 and 0.721). Gradient Boosting achieves the highest recall in the positive class (0.830) , while Bagging demonstrates the best recall for the negative class (0.637).

SHAP values were adopted to identify the five most influential features within the *sus_authorization* dataset, employing the Random Forest algorithm for values extraction. Figure A.22 showcases the SHAP summary plot, highlighting these pivotal features.

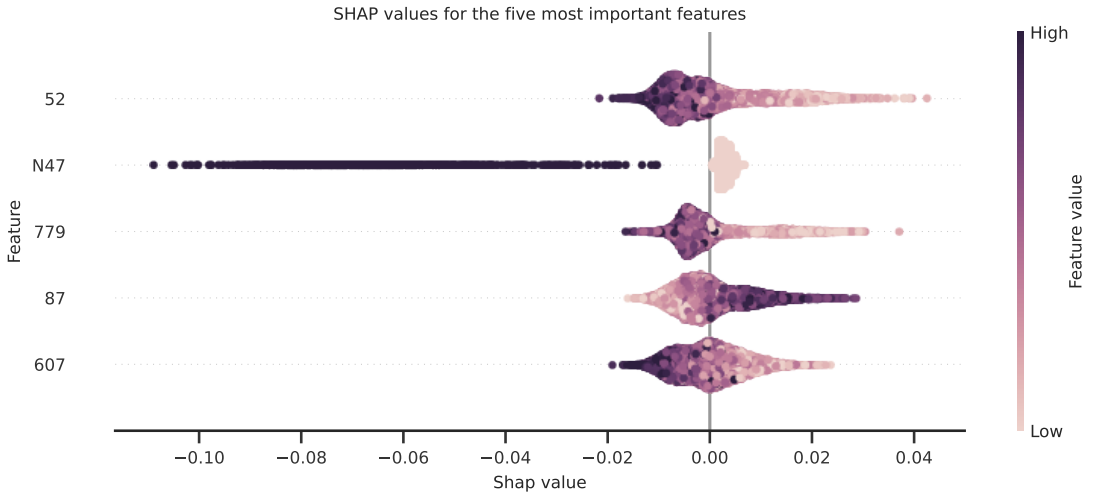


FIGURE A.22 – SHAP summary plot illustrating the five most influential features identified using SHAP values in the *sus_authorization* dataset. The features were extracted using the Random Forest algorithm. Four features correspond to components of the numerical vector representing the clinical condition, information initially in text format. The IDC *N47* was also selected with its presence influencing predictions towards authorization.

Notably, four selected features are elements of the numerical vector derived from clinical descriptions. While their direct interpretation may be challenging, the plot underscores their correlation with the target feature. Among these features is the ICD *N47*, whose presence notably influences predictions as an *authorized* case.

Validating with external data

In this data source we could not assembly a second dataset based on temporal criteria. In this way, we randomly select a tenth of the dataset, stratified by class, to compose a test set. Here we evaluate the predictive performance of models when trained in the *authorization_sus* dataset and tested in the *authorization_sus_test* dataset. Table A.19 presents the result of this assessment.

A.5.4 Discussion

We compiled a dataset sourced from GPs working in the Brazilian SUS, consisting of requests for specialized care. Our goal was to predict the authorization or denial of these requests. The dataset comprises the clinical description, converted into a numerical vector, along with the patient's age, gender, and IDC-related information.

TABLE A.19 – Average predictive performance and standard deviation of models trained on the *authorization_sus* dataset and tested on the *authorization_sus_test* dataset. The best value for each metric is highlighted in bold. AUC: Area Under the Curve; RCL: Recall; PRC: Precision; SVC: Support Vector Classifier; RF: Random Forest; GB: Gradient Boosting; LR: Logistic Regression; BAG: Bagging; MLP: Multilayer Perceptron.

	$\mathbf{SVC_{linear}}$	$\mathbf{SVC_{rbf}}$	**RF**	**GB**	**LR**	**BAG**	**MLP**
AUC	0.764	**0.784**	0.732	0.765	0.766	0.701	0.729
$\mathbf{RCL_1}$	0.778	0.803	0.809	**0.824**	0.774	0.650	0.736
$\mathbf{RCL_0}$	0.613	0.612	0.544	0.544	0.606	**0.627**	0.600
$\mathbf{PRC_1}$	0.725	**0.731**	0.699	0.703	0.720	0.696	0.707
$\mathbf{PRC_0}$	0.678	**0.704**	0.685	0.702	0.671	0.578	0.635

To assess the dataset's potential, we trained various ML models using a five-fold cross-validation strategy. The models excelled in AUC, recall for the negative class and precision for the positive class, while other metrics did not reach the same levels. Notably, denials appeared to be more predictable than authorizations. This insight suggests that weighting model predictions differently for each class could alleviate the system's burden by automating certain predictions.

The transformation of clinical conditions into numerical vectors facilitated computational processing and modelling. Although language models may offer superior performance metrics, they often require significant time and resources for training. Considering our objective of evaluating and comparing multiple models, we chose numerical representations to streamline the analysis. However, exploring the potential of language models remains a potential avenue for future research.

A.6 Conclusion

In this annex, we presented the main characteristics of the datasets compiled for this study and offered an initial performance assessment of various ML algorithms trained on these datasets. This evaluation includes cross-validation techniques and, in some instances, the incorporation of external validation with a secondary dataset. While there is room for improvement, these results underscore the potential of ML models to support decision-making across diverse healthcare scenarios with data collected routinely.

Many studies in Artificial Intelligence create solutions that are less usable, as they are disconnected from clinical practice by not involving experts in the data (TONEKABONI *et al.*, 2019). Despite producing accurate models, they may be of limited utility because they fail to understand the needs of the medical team. With the assistance of a medical professional during the dataset assembly process, we ensured the consistency and relevance of the cleaned datasets to healthcare scenarios. We consider the support of an expert in the data was indispensable, and with this collaboration, our contribution gained value by

offering more reliable and coherent databases.

To cite a few examples, involving an expert was crucial in validating when a prognosis is relevant to the medical team, comprehending the dynamics of the real-life disease notification process and validating other decisions made, such as removing infrequent variables and the imputation strategy for missing values. We were also assisted in preprocessing tasks, such as identifying medical exams with different labels but representing the same analysis. Finally, it was essential in results interpretation, such as understanding why the performance in some cases is higher than in others.

The unique attributes of our datasets make them invaluable resources for a wide range of ML applications, particularly in the realm of public health. Additionally, We observed varying performance levels across different studies, highlighting the importance of understanding the specific problem context. For example, in *severity_sjc* and *hospitalization_sjc* (see Section A.2.3.1 and Section A.2.3.2) we achieved high-performance metrics, while models did not address the problem set by *hosp_days_sjc* (see Section A.2.3.3). This diversity is intriguing because it allows us to explore how strategies impact problems with diverse characteristics.

In this way, our datasets offer valuable resources for ML researchers seeking real-world data to test novel models or strategies. These datasets encompass diverse characteristics, including missing data, imbalanced classes and the data types adopted. The initial performances provided baseline ML models as benchmarks, serving as starting points for researchers to explore the potential of these datasets. By making them publicly available in our repository, we foster collaboration within the ML community and encourage researchers to tackle critical health challenges.

Bibliography

ANIK, A. I.; BUNT, A. Data-centric explanations: Explaining training data of machine learning systems to promote transparency. *In*: **Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems**. **Proceedings** [...]. [*S.l.: s.n.*], 2021. p. 1–13.

BARCELLOS, C.; MATOS, V.; LANA, R. M.; LOWE, R. Climate change, thermal anomalies, and the recent progression of dengue in brazil. **Scientific reports**, Nature Publishing Group UK London, v. 14, n. 1, p. 5948, 2024.

BURATTINI, M. N.; LOPEZ, L. F.; COUTINHO, F. A.; SIQUEIRA-JR, J. B.; HOMSANI, S.; SARTI, E.; MASSAD, E. Age and regional differences in clinical presentation and risk of hospitalization for dengue in brazil, 2000-2014. **Clinics**, SciELO Brasil, v. 71, p. 455–463, 2016.

CAPUANO, A.; ROSSI, F.; PAOLISSO, G. COVID-19 kills more men than women: An overview of possible reasons. **Frontiers in Cardiovascular Medicine**, v. 7, n. 131, p. 1–7, July 2020. Available at: https://doi.org/10.3389/fcvm.2020.00131.

COLLINS, G. S.; DHIMAN, P.; NAVARRO, C. L. A.; MA, J.; HOOFT, L.; REITSMA, J. B.; LOGULLO, P.; BEAM, A. L.; PENG, L.; CALSTER, B. V. *et al.* Protocol for development of a reporting guideline (tripod-ai) and risk of bias tool (probast-ai) for diagnostic and prognostic prediction model studies based on artificial intelligence. **BMJ open**, British Medical Journal Publishing Group, v. 11, n. 7, p. e048008, 2021.

FACELI, K.; LORENA, A.; GAMA, J.; ALMEIDA, T. A.; CARVALHO, A. Inteligência artificial–uma abordagem de aprendizado de máquina. **Rio de Janeiro: LTC**, 2021.

FERNANDES, F. T.; OLIVEIRA, T. A. de; TEIXEIRA, C. E.; BATISTA, A. F. de M.; COSTA, G. D.; FILHO, A. D. P. C. A multipurpose machine learning approach to predict covid-19 negative prognosis in são paulo, brazil. **Scientific reports**, Nature Publishing Group, v. 11, n. 1, p. 1–7, 2021.

FERNÁNDEZ, A.; GARCÍA, S.; GALAR, M.; PRATI, R. C.; KRAWCZYK, B.; HERRERA, F. **Learning from imbalanced data sets**. [*S.l.*]: Springer, 2018.

KIM, H.-J.; HAN, D.; KIM, J.-H.; KIM, D.; HA, B.; SEOG, W.; LEE, Y.-K.; LIM, D.; HONG, S. O.; PARK, M.-J. *et al.* An easy-to-use machine learning model to predict the prognosis of patients with covid-19: Retrospective cohort study. **Journal of medical Internet research**, JMIR Publications Inc., Toronto, Canada, v. 22, n. 11, p. e24225, 2020.

KYONO, T.; ZHANG, Y.; BELLOT, A.; SCHAAR, M. van der. Miracle: Causally-aware imputation via learning missing data mechanisms. **Advances in Neural Information Processing Systems**, v. 34, p. 23806–23817, 2021.

LAVALLEY, M. P. Logistic regression. **Circulation**, Am Heart Assoc, v. 117, n. 18, p. 2395–2399, 2008.

LEE, T.-H.; ULLAH, A.; WANG, R. Bootstrap aggregating and random forest. **Macroeconomic forecasting in the era of big data: Theory and practice**, Springer, p. 389–429, 2020.

LORENA, A. C.; CARVALHO, A. C. de. Uma introdução às Support Vector Machines. **Revista de Informática Teórica e Aplicada**, XIV, n. 2, p. 43–67, 2007. Available at: https://www.seer.ufrgs.br/rita/article/viewFile/5690/3543.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. *In*: **31st Conference on Neural Information Processing Systems (NIPS 2017)**. Long Beach, CA, USA: [*s.n.*], 2017. Available at: https://arxiv.org/abs/1705.07874.

MALLICK, A.; HSIEH, K.; ARZANI, B.; JOSHI, G. Matchmaker: Data drift mitigation in machine learning for large-scale systems. **Proceedings of Machine Learning and Systems**, v. 4, p. 77–94, 2022.

MELLO, L. E.; SUMAN, A.; MEDEIROS, C. B.; PRADO, C. A.; RIZZATTI, E. G.; NUNES, F. L.; BARNABÉ, G. F.; FERREIRA, J. E.; SÁ, J.; REIS, L. F. *et al.* Opening brazilian covid-19 patient data to support world research on pandemics. **Zenodo**, 2020.

MORAES, B. A. F. de; MIRAGLIA, J.; DONATO, T.; FILHO, A. Covid-19 diagnosis prediction in emergency care patients: a machine learning approach. **https://www. medrxiv. org/content/medrxiv/early/2020/04/07/2020.04. 04.20052092. full. pdf**, 2020.

MOURA, E. C.; CORTEZ-ESCALANTE, J.; CAVALCANTE, F. V.; BARRETO, I. C. d. H. C.; SANCHEZ, M. N.; SANTOS, L. M. P. Covid-19: evolução temporal e imunização nas três ondas epidemiológicas, brasil, 2020–2022. **Revista de Saúde Pública**, SciELO Brasil, v. 56, p. 105, 2022.

NASCIMENTO, E. R. do; ALBUQUERQUE, M. A. de; BARROS, K. N. N. de O.; BARROS, P. S. N. Cluster analysis applied to the human development index (hdi) of brazilian states. **Research, Society and Development**, v. 11, n. 2, p. e18011225747–e18011225747, 2022.

NELSON, D. **Gradient Boosting Classifiers in Python with Scikit-Learn**. aug 2019. Available at: https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-learn.

OSISANWO, F.; AKINSOLA, J.; AWODELE, O.; HINMIKAIYE, J.; OLAKANMI, O.; AKINJOBI, J. Supervised machine learning algorithms: classification and comparison. **International Journal of Computer Trends and Technology (IJCTT)**, v. 48, n. 3, p. 128–138, 2017. Available at: https://doi.org/10.14445/22312803/ijctt-v48p126.

PAIVA, P. Y. A.; MORENO, C. C.; SMITH-MILES, K.; VALERIANO, M. G.; LORENA, A. C. Relating instance hardness to classification performance in a dataset: a visual approach. **Machine Learning**, Springer, p. 1–39, 2022.

PAN, I.; MASON, L. R.; MATAR, O. K. Data-centric engineering: Integrating simulation, machine learning and statistics. challenges and opportunities. **Chemical Engineering Science**, Elsevier, v. 249, p. 117271, 2022.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V. *et al.* Scikit-learn: Machine learning in python. **Journal of machine learning research**, v. 12, n. Oct, p. 2825–2830, 2011.

POPESCU, M.-C.; BALAS, V. E.; PERESCU-POPESCU, L.; MASTORAKIS, N. Multilayer perceptron and neural networks. **WSEAS Transactions on Circuits and Systems**, World Scientific and Engineering Academy and Society (WSEAS) Stevens Point ..., v. 8, n. 7, p. 579–588, 2009.

SAGARRA-ROMERO, L.; VIÑAS-BARROS, A. Covid-19: Short and long-term effects of hospitalization on muscular weakness in the elderly. **International journal of environmental research and public health**, Multidisciplinary Digital Publishing Institute, v. 17, n. 23, p. 8715, 2020.

SANYAOLU, A.; OKORIE, C.; MARINKOVIC, A.; PATIDAR, R.; YOUNIS, K.; DESAI, P.; HOSEIN, Z.; PADDA, I.; MANGAT, J.; ALTAF, M. Comorbidity and its impact on patients with covid-19. **SN comprehensive clinical medicine**, Springer, v. 2, p. 1069–1076, 2020.

SEEDAT, N.; IMRIE, F.; SCHAAR, M. van der. Dc-check: A data-centric ai checklist to guide the development of reliable machine learning systems. **arXiv preprint arXiv:2211.05764**, 2022.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. *In*: **9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)**. **Proceedings** [...]. [*S.l.: s.n.*], 2020.

TONEKABONI, S.; JOSHI, S.; MCCRADDEN, M. D.; GOLDENBERG, A. What clinicians want: contextualizing explainable machine learning for clinical end use. *In*: PMLR. **Machine learning for healthcare conference**. **Proceedings** [...]. [*S.l.: s.n.*], 2019. p. 359–380.

WOLPERT, D. H.; MACREADY, W. G. No free lunch theorems for optimization. **IEEE transactions on evolutionary computation**, IEEE, v. 1, n. 1, p. 67–82, 1997.

WU, Y.; HOU, B.; LIU, J.; CHEN, Y.; ZHONG, P. Risk factors associated with long-term hospitalization in patients with COVID-19: A single-centered, retrospective study. **Frontiers in Medicine**, v. 7, n. 315, p. 1–10, 2020. Available at: https://doi.org/10.3389/fmed.2020.00315.

YIU, T. **Understanding Random Forest**. June 2019. Available at: https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

ZHA, D.; BHAT, Z. P.; LAI, K.-H.; YANG, F.; JIANG, Z.; ZHONG, S.; HU, X. Data-centric artificial intelligence: A survey. **arXiv preprint arXiv:2303.10158**, 2023.