Introduction

This document provides an overview of the variables contained within the *hospitalization* dataset compiled from the Chikungunya disease database (1). Data was extracted from DATASUS, the TI department of the Unified Health System (SUS) in Brazil. The original database was collected by the Notifiable Diseases Information System (SINAN). The objective of the classification task is to predict hospitalization from the information contained in the moment of the disease notification. The document also provides an overview of the preprocessing steps took to assemble the dataset (2). For more detailed information about the datasets, please visit the project repository on GitHub: `https://github.com/gabivaleriano/HealthDataBR`.

Dictionary of Variables

Feature name	Description	Values	Observations
fever myalgia headache exanthema conjunctivitis petechiae leukopenia retro_orbital_pain nausea vomiting back_pain arthritis arthralgia	Symptom.	[0,1]	
diabetes arterial_hypertension	Comorbidity.	[0,1]	
age	Patient age.	Integer.	
hospitalization	Patient outcome.	[0,1]	Hospitalized = 1.
sex	Patient sex.	[0,1]	Female = 1.
id_state	State where the unit is located of health who carried out the notification.	Integer.	Table with Codes and Acronyms by the Brazilian Institute of Geography and Statistics (IBGE).
epidemiological_week	Epidemiological week in which the first symptoms occurred	[1:52]	Weeks of the standardized epidemiological calendar.
race	Color or race declared by the person.	[1:5]	1: White 2: Black 3: Yellow 4: Parda 5: Indigene
schooling_years	Grade and degree that the person is attending, or attended, considering the last series completed with approval, at the time of notification.	[0:8]	0: Without schooling 1: Elementary school I incomplete 2: Elementary school I complete 3: Elementary school II incomplete 4: Elementary school II complete 5: High school incomplete 6: High school complete 7: Higher school incomplete 8: Higher school complete

Table 1: Dictionary of Variables of the *hospitalization* dataset assembled from Chikungunya database.

Filters applied

Filter	**Size**
Initial dataset	269960
Filter 1: remove duplicates	269823
Filter 2: remove notifications without state information	269823
Filter 3: remove patients without year of birth	267576
Filter 4: remove patients older than 110 and younger than 12 years old	241772
Filter 5: remove patients without sex information	241514
Filter 6: remove patients without race information	207769
Filter 7: remove patients without schooling information	207667
Filter 8: keep only patients in the acute stage of the disease	138861
Filter 9: keep only patients with information about hospitalization	95014
Filter 10: remove patients hospitalized more than 15 days after the first symptoms	94821
Filter 11: remove non-hospitalized patients that died	91788
Delete columns without relevant information, filled mostly with NA's or redundant	

Table 2: Preprocessing steps to assemble the dataset.