1 Datasets

Here we describe the datasets employed in our experiments.

1.1 *Hospitalization* and *Severity*

Both datasets were extracted from the same data source containing the symptoms and comorbities declared when a citizen took a COVID test in the city of São José de Campos located in São Paulo state in Brazil. This data source was obtained through a partnership with the health Secretariat of São José dos Campos. The dataset encompasses the tests between March of 2020 and May of 2021. This database contains sensitive information and for ethical reasons can not be shared.

The *Hospitalization* dataset has three balanced classes: short (1:5 days), medium (6:10 days) and long (greater than 10 days) hospitalization. Table 2 summarize dataset information according to each class.

Table 1. Distribution of attributes per class in *Hospitalization* data set.

	Short	Medium	Long
Number of instances	1062	812	954
Fever	0.573	0.596	0.591
Cough	0.748	0.736	0.768
Sore throat	0.247	0.225	0.205
Dyspnea	0.707	0.778	0.752
Respiratory distress	0.529	0.499	0.534
Low oxygen saturation	0.569	0.638	0.666
Diarrhea	0.134	0.109	0.113
Vomit	0.065	0.578	0.055
Other symptoms	0.652	0.639	0.600
Chronic cardiovascular disease	0.034	0.298	0.0437
Immunodeficiency immunodepression	0.028	0.030	0.021
Chronic kidney disease	0.030	0.020	0.019
Diabetes mellitus	0.266	0.284	0.334
Asthma	0.034	0.026	0.035
Chronic neurological disease	0.024	0.039	0.026
Other chronic pneumopathy	0.034	0.030	0.034
Obesity	0.059	0.070	0.073
Other risks	0.328	0.372	0.410
Chronic respiratory disease	0.019	0.021	0.016
Sex (female)	0.455	0.432	0.413
Age (mean - sd)	53.3 - 19.7	58.5 - 16.5	61.3 - 14.7

Dataset *Severity* has two classes: severe and non-severe patients. Severe patients are those who progressed to death within days 30 of hospitalization or

were hospitalized for ten or more days. Non severe patients were not hospitalized or were hospitalized for less than 10 days. This dataset was very imbalanced, with only 0.9% of instances in the severe class. Since the class imbalance affects the instance hardness analysis, in this work we randomly removed part of the majority class before the experiments. Table 2 summarizes dataset information according to each class.

1.2 *Hospital 1* and *Hospital 2*

Both datasets were extracted from the same data source containing the laboratorial test results during the COVID pandemic for patients admitted at two private hospitals in the city of São Paulo. This data source is open and publicly available at [1]. The preprocessed datasets are available in the Supplementary material. In these datasets all patients are hospitalized and severity is assigned to patients who progress to death (within a period of 30 days from hospitalization) or for which the hospitalization period was larger than 14 days. Both datasets were imbalanced, and passed had instances randomly removed before the experiments. These datasets also contain missing values, which are input by the mean of the three nearest neighbors values. Table 4 summarizes some datasets information according to each class, after the imputation step.

[1] https://repositoriodatasharingfapesp.uspdigital.usp.br/

Table 2. Distribution of attributes per class in *Severity* dataset.

	Severe	Non-severe
Number of instances	1710	1710
Fever	0.551	0.680
Cough	0.735	0.791
Sore throat	0.184	0.546
Dyspnea	0.784	0.308
Respiratory distress	0.568	0.041
Low oxygen saturation	0.692	0.046
Diarrhea	0.091	0.012
Vomit	0.056	0.007
Other symptoms	0.556	0.750
Chronic cardiovascular disease	0.470	0.150
Immunodeficiency immunodepression	0.030	0.008
Diabetes mellitus	0.367	0.115
Obesity	0.074	0.009
Other risks	0.438	0.029
Chronic respiratory disease	0.017	0.036
Sex (female)	0.414	0.536
Age (mean - sd)	64.7 - 15.7	42.0 - 16.0

Table 3. Summary of the datasets *Hospital 1* including number o patients, period of data collection, percentage of male patients, age statistics and blood tests.

	Severe	Non-severe
Period	03/2020 - 05/2021	
Number of instances	529	529
Sex (female)	0.357	0.340
Age (mean - sd)	68.9 - 12.9	59.4 - 15.4
Blood tests	sodium, potassium, C-reactive protein, GOT, GPT, urea, D-dimer, troponin, creatine kinase, blood count, creatinine	

Table 4. Summary of the datasets *Hospital 2* including number o patients, period of data collection, percentage of male patients, age statistics and blood tests.

	Severe	Non-severe
Period	03/2020 - 01/2021	
Number of Instances	67	67
Sex (female)	0.328	0.537
Age (mean - sd)	65.8 - 14.6	55.2 - 17.0
Blood tests	sodium, potassium, C-reactive protein, GOT, GPT, urea, D-dimer, troponin, creatine kinase, blood count, creatinine	