Revista Colombiana de Estadística - Applied Statistics
January 2025, volume 48, issue 1, pp. 99–115
http://doi.org/10.15446/rce.v48n1.113819

Small Samples, New Viruses, Inputs for Decision-Making and Methodology: Bootstrap and Smote

Muestras pequeñas, nuevos virus, insumos para la toma de decisiones y metodología: Bootstrap y SMOTE

Martha Misas-Arango[1,a], Catherine Pereira-Villa[1,b], Wilson F. Rodríguez-Gómez[1,c], Jose E. Gomez-Gonzalez[2,d]

[1]International School of Economic and Administrative Sciences, Universidad de La Sabana, Chía, Colombia

[2]Department or School, Department of Finance, Information Systems and Economics, City University of New York - Lehman College, New York, United States

Abstract

This study presents a comprehensive methodology that combines resampling and oversampling techniques to address the challenges of limited and unbalanced data, specifically in the context of viral emergencies such as the COVID-19 pandemic. Utilizing advanced statistical techniques like Bootstrap and SMOTE, the study conducts a retrospective analysis of COVID-19 patients, identifying those at higher risk of mortality. The proposed methodology not only enhances the accuracy of predictions in scenarios with limited data but also facilitates better decision-making in clinical triage systems. By applying these methods, the study achieves early and accurate identification of high-risk individuals, optimizing resource allocation and timely medical interventions. The results demonstrate that this combination of statistical techniques effectively improves health systems and responses to new viral threats, providing a robust foundation for informed decision-making in medical emergencies.

Key words: Death predictors; Early warning systems; New viruses; Small sample methodologies; SMOTE.

[a]. E-mail: martha.misas@unisabana.edu.co
[b]. E-mail: catherine.pereira@unisabana.edu.co
[c]. E-mail: wilson.rodriguez1@unisabana.edu.co
[d]. E-mail: jose.gomezgonzalez@lehman.cuny.edu

Resumen

Este estudio presenta una metodología integral que combina técnicas de remuestreo y sobremuestreo para abordar los desafíos de datos limitados y desbalanceados, específicamente en el contexto de emergencias virales como la pandemia de COVID-19. Utilizando técnicas estadísticas avanzadas como Bootstrap y SMOTE, el estudio realiza un análisis retrospectivo de pacientes con COVID-19, identificando a aquellos con mayor riesgo de mortalidad. La metodología propuesta no solo mejora la precisión de las predicciones en escenarios con datos limitados, sino que también facilita una mejor toma de decisiones en los sistemas de triaje clínico. Al aplicar estos métodos, el estudio logra una identificación temprana y precisa de individuos de alto riesgo, optimizando la asignación de recursos y las intervenciones médicas oportunas. Los resultados demuestran que esta combinación de técnicas estadísticas mejora de manera efectiva los sistemas de salud y las respuestas ante nuevas amenazas virales, proporcionando una base sólida para la toma de decisiones informadas en emergencias médicas.

***Palabras clave*:** Bootstrapping; Metodologías para muestras pequeñas; Nuevos virus; Predictores de mortalidad; Sistemas de alerta temprana; SMOTE.

1. Introduction

The World Health Organization's declaration of COVID-19 as a pandemic in March 2020 highlighted the global susceptibility to emerging infectious diseases and underscored the pressing need for robust, adaptable methodologies to confront such crises. As of October 5th, 2022, the pandemic has resulted in over 624 million reported cases and more than 6.5 million deaths worldwide (Lupei et al., 2022). This staggering toll not only reflects the virulence of the SARS-CoV-2 virus but also highlights systemic challenges within healthcare infrastructures, particularly the ability to rapidly process and interpret limited, imbalanced data during the nascent stages of an outbreak.

Emerging viruses, by their very nature, present a conundrum: the urgency to understand and respond is met with the reality of sparse initial data and significant class imbalances, especially when identifying severe cases or mortality risk factors. Traditional statistical models, such as logistic regression, are lauded for their interpretability and straightforward implementation (Hosmer et al., 2013). However, their efficacy diminishes in the face of small sample sizes and skewed class distributions, often leading to models that are biased towards the majority class and insufficiently sensitive to minority class nuances—an issue critically pertinent when the minority class represents high-risk patient outcomes (Cornilly et al., 2023; Morgenthaler, 2023).

To mitigate these challenges, the literature has explored various methodological innovations. The Synthetic Minority Over-sampling Technique (SMOTE), introduced by Chawla et al. (2002), offers a means to address class imbalance by generating synthetic examples of the minority class, thereby facilitating a more balanced training dataset. While SMOTE has demonstrated efficacy in enhancing model performance (Fernández et al., 2018; Neptune.ai, 2023), it is not without

limitations, particularly when synthetic samples do not adequately capture the complex variability inherent in real-world data.

Bootstrap methods, rooted in resampling with replacement, provide a framework for estimating the sampling distribution of a statistic and have been instrumental in improving model robustness (Efron & Tibshirani, 1994; Le Thi & Nguyen, 2023). Breiman (1996) further advanced this domain with the development of bagging techniques, which aggregate predictions from multiple bootstrap samples to reduce variance. Nonetheless, bootstrap approaches alone may falter when confronted with pronounced class imbalances.

Cost-sensitive learning emerges as another pivotal strategy, wherein misclassification costs are explicitly integrated into the modeling process (Elkan, 2001; Analytics India Magazine, 2023). By assigning higher penalties to errors on the minority class, models are coerced into achieving better sensitivity. However, determining appropriate cost matrices is often challenging and context-dependent, potentially limiting the generalizability of such models.

Recognizing the individual merits and limitations of these techniques, this study proposes a hybrid methodology that synergistically combines SMOTE and bootstrap methods. By initially applying SMOTE, we aim to rectify class imbalances through the generation of synthetic minority class instances. Subsequently, bootstrap resampling is employed to enhance the stability and robustness of the predictive models. This integrative approach aspires to harness the strengths of both methodologies, mitigating their standalone weaknesses, particularly in scenarios characterized by limited and skewed data.

The applicability and efficacy of this hybrid methodology are examined through a retrospective analysis of COVID-19 patient data from a region in Colombia. While numerous studies have dissected COVID-19 mortality determinants across various geographies (Banik et al., 2020; Toya & Skidmore, 2021; Upshaw et al., 2021; Yalaman et al., 2021), there exists a relative paucity of research focusing on emerging and Latin American nations (Cifuentes et al., 2021; De la Hoz-Restrepo et al., 2020; Fernández-Niño et al., 2020; Laajaj et al., 2021). By concentrating on this demographic, the study not only fills a critical gap in the literature but also underscores the versatility of the proposed methodology across diverse epidemiological landscapes.

Moreover, the insights gleaned from this analysis have tangible implications for clinical decision-making. Rapid and accurate identification of mortality risk factors can inform triage protocols, optimize resource allocation, and ultimately enhance patient outcomes. The integration of Natural Language Processing algorithms to extract nuanced patient clinical histories further enriches the data quality, facilitating more precise modeling.

In structuring this paper, we commence with the present introduction, delineating the study's rationale and situating it within the broader scholarly discourse. The subsequent section details the methodology, explicating the integration of SMOTE and bootstrap techniques alongside data extraction processes. This is followed by the case study, wherein the proposed methodology is applied to the Colombian COVID-19 dataset, and the resultant findings are discussed. The paper

culminates with the conclusions, encapsulating the study's contributions, limitations, and avenues for future research.

2. Methodology

The methodological development starts from a research question that is related to a binary answer, y, about the realization of an event under study. This response can be defined as a Bernoulli random variable,$y \sim \text{Bernoulli}(p)$; $p = \text{prob}(y = 1)$. Let $\boldsymbol{Y} = \{y_1, y_2, \ldots, y_T\}$, be a random and representative sample of size T from the population, $\{p_1, p_2, \ldots, p_T\}$; $\forall i = 1, \ldots, T \quad p_i = \text{prob}(y_i = 1)$.

From the Bernoulli distribution, we have that $\forall i = 1, \ldots, T \quad E(y_i) = p_i$ and $y_i = E(y_i) + e_i$ where e_i satisfies basic regression assumptions. That is, $\forall i = 1, \ldots, T \; y_i = p_i + e_i$.

At this point, the objective question of the research focuses on what are the determinants of probability $p_i \;\; \forall i = 1, \ldots, T$. The boundary of $p_i \in (0, 1)$ requires defining probability as a nonlinear function of its determinants (The function F refers to the cumulative distribution function of the standard normal or logistic). Thus:

$$\forall i = 1, \ldots, T \quad p_i = F(\mathbf{X}_i^{'}\boldsymbol{\beta}) \tag{2.1}$$

$$y_i = F(\mathbf{X}_i^{'}\boldsymbol{\beta}) + e_i$$

Let

$$\mathbf{X}_i^{'} = (x_{1_i}, x_{2_i}, \cdots, x_{K_i}) \tag{2.2}$$

where K attributes correspond to the i-th individual associated with the i-th response y_i, which may help elucidate the likelihood of the event under consideration occurring.

Thus, the design matrix $\mathbf{X}$ is defined as:

$$\mathbf{X} = \begin{bmatrix} x_{1_1} & \cdots & x_{K_1} \\ \vdots & \ldots & \vdots \\ x_{1_T} & \cdots & x_{K_T} \end{bmatrix}_{T \times K} \tag{2.3}$$

and the information set, $\mathbf{CI}$, is formed by combining the vector $\boldsymbol{Y}$ and design matrix $\mathbf{X}$.

$$\mathbf{CI} = [\boldsymbol{Y} \sim \mathbf{X}]$$

$$= \begin{bmatrix} y_1 & x_{1_1} & \cdots & x_{K_1} \\ \vdots & \vdots & \vdots & \vdots \\ y_T & x_{1_T} & \cdots & x_{K_T} \end{bmatrix}_{T \times (K+1)}$$

It should be noted that in many cases, there is a population imbalance in the responses that leads to obtaining a similar sample imbalance. This fact produces:

(i) prediction bias in favour of the majority class, (ii) inability to recognize the characteristics of the minority class, (iii) inadequate generalization on unobserved data, especially in the minority class, (iv) inaccuracy in goodness of fit measures that can lead to risky conclusions, especially in situations where it is crucial to correctly identify the minority class, (v) high rates of false-positives or false-negatives and (vi) affection of the optimization process via minimization of errors associated with the majority class.

To overcome the imbalance problems and optimally select the set of determinants of the probability of the event under study, this research proposes a methodology that combines resampling methods or bootstrapping on the majority class and oversampling techniques or ($SMOTE$) on the minority class.

Thus, in the first instance, it is necessary that the matrix X is made up of attributes that have a differential impact on the vector Y, when it takes the value of 1 than when it takes the value of 0, and that they are not correlated with each other. Therefore, an exhaustive review of the attributes considered is necessary, and this review requires that the sample, and therefore the set of information, be balanced with respect to the vector Y of binary responses.

2.1. Selection Set of Possible Explanatory Variables

The first step uses bootstrapping techniques to form B random samples, equal in size to the number of observations from the minority group, from the dataset that make up the majority group. Thus, joining these groups will obtain B perfectly balanced samples that we will denote as MB^j, $\forall j = 1, \ldots, B$.

Step 1. Under the assumption of sample imbalance, it is necessary to form a set of balanced samples from *Bootstrap* (resampling with replacement). □

Step 1.1. In a vector form, the subsets $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are defined in such a way that $\boldsymbol{Y}_1 \subseteq \boldsymbol{Y}$ such that $y_i = 1, \forall i = 1, \ldots, T$ in size $T_{\boldsymbol{Y}_1} \times 1$, and $\boldsymbol{Y}_2 \subseteq \boldsymbol{Y}$ such that $y_i = 0, \forall i = 1, \ldots, T$ in size $T_{\boldsymbol{Y}_2} \times 1$, being $T_{\boldsymbol{Y}_1}/T \neq T_{\boldsymbol{Y}_2}/T$. Similarly, there are $\mathbf{X}_1$ and $\mathbf{X}_2$ matrices of characteristics associated with the responses included in $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, respectively. Conforming, therefore: $\mathbf{CI}_1 = [\boldsymbol{Y}_1 \sim \mathbf{X}_1]$ of dimension $T_{\boldsymbol{Y}_1} \times (K+1)$, and $\mathbf{CI}_2 = [\boldsymbol{Y}_2 \sim \mathbf{X}_2]$ of dimension $T_{\boldsymbol{Y}_2} \times (K+1)$. Being $\mathbf{X}^R = \mathbf{X}_1$ or $\mathbf{X}_2$ depending on the case.

□

Step 1.2. It is determined:

$$Min_T = \text{Minimum}(T_{\boldsymbol{Y}_1}, T_{\boldsymbol{Y}_2}) = \#\text{Obs. of the minority group}$$
$$Max_T = \text{Maximum}(T_{\boldsymbol{Y}_1}, T_{\boldsymbol{Y}_2}) = \#\text{Obs. of the majority group}$$

□

Step 1.3. From bootstrapping on the set $\mathbf{CI}_{i=1 \text{ or } 2}$ in size $Max_T \times (K+1)$, B random samples of size $Min_T \times (K+1)$ are generated, $\mathbf{M}^1, \mathbf{M}^2, \ldots, \mathbf{M}^B$. □

Step 1.4. Thus, B balanced samples can be built:

$$\forall j = 1, \ldots, B \quad \text{is defined as } \mathbf{MB}^j = \left[\mathbf{M}^j \sim \mathbf{N}\right] \quad \text{of dimension } 2Min_T \times (K+1);$$
$$\text{where } \mathbf{N} = \{\text{the subset of } \boldsymbol{Y}(\boldsymbol{Y}_1 \text{ or } \boldsymbol{Y}_2) \text{ of size} = Min_T \times (K+1)\}$$

□

Step 2. Difference of means. In each of the balanced samples, $\mathbf{MB}^j\ \forall j = 1, \ldots, B$, a mean difference test is carried out on each of the variables (or columns) that make up the matrix of characteristics $\mathbf{X}^R$, according to the implicit classification in the vector $\boldsymbol{Y}$. It should be noted that this test will depend, in its construction, on the type of variable under analysis, continuous or discrete. Thus, from each sample, those variables are selected where the test concludes that there is evidence to reject the hypothesis of equality of means:

$$\forall i = 1, \ldots, K \quad \forall j = 1, \ldots, B \quad H_0 : \text{mean } \mathbf{X}_i | \mathbf{X}_i \in \mathbf{M}^j = \text{mean } \mathbf{X}_i | \mathbf{X}_i \in \mathbf{N}$$

Let $\mathbf{MBL}^j =$
$\mathbf{MB}^j - \{\mathbf{X}_i | \forall i = 1, \ldots, K \quad \text{There is no evidence to reject } H_0 \text{ in } \mathbf{X}_i\}$
$\forall j = 1, \ldots, B$, be balanced and clean matrices of irrelevant variables.

Later, in step 3, we proceed to estimate the correlations between those variables of the set of attributes that are part of $\mathbf{MBL}^j\ \forall j = 1, \ldots, B$, to eliminate those that carry the same information. □

Step 3. Debugging correlated information.

In order to eliminate variables that carry the same information, the correlations between variables in each balanced sample (Brown & Benedetti, 1977; Roscino & Pollice, 2006), $\mathbf{MBL}^j\ \forall j = 1, \ldots, B$, are estimated. In the event of discovering, for a particular sample $\mathbf{MBL}^j, m(m \geq 2)$, variables correlated with each other, $C = \{x_{1_i}, \ldots, x_{m_i}\}$, we proceed to estimate m logistic regressions defined as $\forall i = 1, \ldots, 2 \times \text{Min}_T, \forall l = 1, \ldots, m \quad y_i = F(\beta_0 + \beta_1 x_{l_i}) + \vartheta_i;\ y_i, x_{l_i} \in \mathbf{MBL}^j, x_{l_i} \in C$. Thus, from the set C, the variable that presents the coefficient $\hat{\beta}_1$ of greater magnitude in absolute value will be selected. The remaining variables that make up the set C are removed from the sample $\mathbf{MBL}^j$.

In this way, $\mathbf{MD}^j\ \forall j = 1, \ldots, B$ balanced samples were obtained, cleaned of irrelevant variables (Step 2) and correlated with each other (Step 3). □

Step 4. Selection of the final set of likely explanatory variables of the binary decision.

To select the likely set of explanatory variables of the binary response model, the value a is established as the required percentage of samples $\mathbf{MD}^j\ \forall j = 1, \ldots, B$ in which the variable in question must be found. Thus, if a particular variable meets the criteria established by $(a\%)$, it will be part of the final set of explanatory variables. □

2.2. Establish the final balanced information set

Once the set of possible explanatory variables is obtained, we return to a sample of the original sample size, T, and we proceed to obtain the necessary sample balance to overcome the problems described above.

Step 5. "*SMOTE*" size balancing L

Let **CI**, $\mathbf{CI} = \boldsymbol{Y} \sim \mathbf{X}$ be a new information set, where $\boldsymbol{Y}$ corresponds to the binary response variable of dimension $T \times 1$ and $\mathbf{X}$ be the design matrix of dimension $T \times J$, $J \leq J$, where each column is one of the relevant and uncorrelated attributes (steps 2-3). Therefore, this set of information, **CI**, presents the same initial degree of imbalance. To achieve a balanced sample, according to the binary variable, the oversampling methodology or SMOTE is used (Chawla et al., 2002; Hanifah et al., 2015; He & Garcia, 2009). □

Step 5.1. From **CI**, according to Min_T (Step 4.2), the matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ and the corresponding design matrix $\mathbf{X}^E$ are defined, where $\mathbf{X}^E = \mathbf{X}_1$ or $\mathbf{X}^E = \mathbf{X}_2$.

$$\mathbf{X}^E = \begin{bmatrix} x_1^1 & \cdots & x_J^1 \\ \vdots & \cdots & \vdots \\ x_1^{\text{Min}_T} & \cdots & x_J^{\text{Min}_T} \end{bmatrix} = \begin{bmatrix} atrib_1^1 & \cdots & atrib_J^1 \\ \vdots & \cdots & \vdots \\ atrib_1^{\text{Min}_T} & \cdots & atrib_J^{\text{Min}_T} \end{bmatrix}$$

□

Step 5.2. Increasing the number of observations of $\mathbf{X}^E$ in such a way that a balance of size L is achieved (Step 4.2). In perfect balance (50, 50) $L = 2(T - \text{Min}_T)$. □

Step 5.2.1. For each observation in $\mathbf{X}^E$, the nearest neighbouring set, $\mathbf{W}$ (which contains $obs.w$ elements), is identified by topological distance, one of them is the Euclidean distance. □

Step 5.2.2. For each observation in the set $\mathbf{X}^E$, then w synthetic individuals are built. For example, for $\mathbf{X}_1^E = (atrib_1^1, atrib_2^1, \ldots, atrib_J^1)$, the first observation of $\mathbf{X}^E$, their $\mathbf{W}$ synthetic individuals are determined as follows:

$$\forall w = 1, \ldots, obs.w$$

$$\begin{aligned} \mathbf{X}_1^{E^k} = (&\boldsymbol{atrib}_1^1 + \text{ranuni}(0,1) \times (atrib_1^w - atrib_1^1), \\ &atrib_2^1 + \text{ranuni}(0,1) \times (atrib_2^w - atrib_2^1), \ldots, \\ &atrib_J^1 + \text{ranuni}(0,1) \times (atrib_J^w - atrib_J^1)) \end{aligned}$$

Be $\mathbf{X}^{EA} = \mathbf{X}^E \sim$ synthetic set and $\mathbf{C}_{\mathbf{X}}^E$, ($\mathbf{X}_1$ or $\mathbf{X}_2$), the complement set of $\mathbf{X}^E$. Thus, the final set of information that makes up the design matrix will be $\mathbf{X}^F = \mathbf{C}_{\mathbf{X}}^E \sim \mathbf{X}^{EA}$. □

In the final set of information, $\mathbf{CI}^F = \boldsymbol{Y} \sim \mathbf{X}^F$, the i-th observation corresponds to the binary selection response Y_i and its set of $\mathbf{X}_i^F$ attributes.

2.3. Determinants of the Event Under Study

Once the balanced sample is reached, we proceed to the estimation of the logistic model that will allow the determinants of the probability of the event under study to be found through tests of significance. Additionally, the results on the quality of the prediction and the marginal effects associated with the selected determinants will be presented.

Step 6. The selection of the determinants of the event under study is based on the results of the *stepwise* estimation of a logistic regression model (Bhandari et al., 2022):

$$\forall i = 1, \ldots, L \quad y_i = F(\mathbf{X}_i^{F'} \boldsymbol{\beta}) + e_i$$

The set of determinants will be made up of those relevant variables in the explanation of the binary decision at a given level of input and output significance, a_1 and a_2, respectively. Once the set of determinants has been defined, we proceed to the analysis of the confusion matrix, the ROC curve and the marginal effects. □

The methodology proposed in this work and that was presented in detail previously is reflected in the scheme of Figure 1.

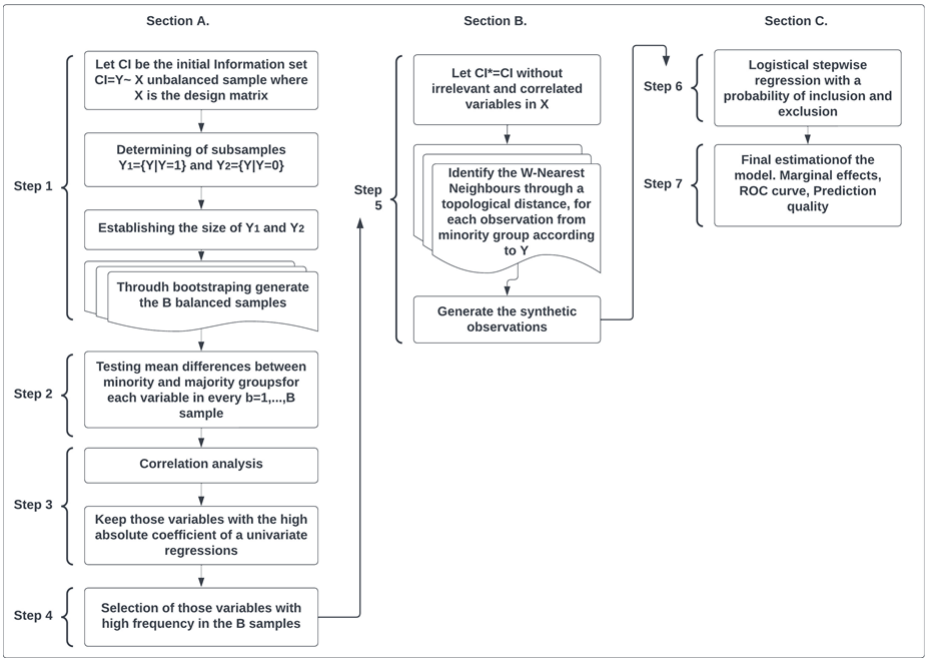

FIGURE 1: Methodology Scheme.

3. Case Study in the Application of the Methodology

This methodology can be applied in cases where due to the nature of the phenomenon studied, there are small and unbalanced samples (Dal Pozzolo et al., 2015). A recent case that meets these characteristics is the outbreak of SARS-COV2 in the first half of 2020. Therefore, at that time, it was important to understand the determinants that made a person, who is infected by the virus, die or live (Castro et al., 2021; Li et al., 2020).

For this purpose, a random and representative sample of $T = 354$ patients diagnosed with COVID-19 who entered the Clinic of the University of La Sabana between September 4 and 30, 2020, was used. Of these, 296 were recovered patients (majority group Max_T) and 58 were deceased (minority group Min_T). The methodology proposed above is pertinent to identify these determinants of death since, in particular, the sample is small and unbalanced (84% -16%).

In this case, the dependent variable is $Y = 1$ if the person dies and $Y = 0$ if the person lives, and the possible determinants ($K = 32$) consigned in the design matrix X are obtained from the medical records of the patients and a numerical assignment derived from natural language processing (see Annex 1). This set of explanatory variables includes patient characteristics such as age, sex, initial symptoms, comorbidities, treatment, and complications during hospitalization.

Subsequently, the most relevant attributes are identified, discarding those that provide the same information. For this, we take $B = 150$ balanced samples without repetition (step 1). For each subsample, a mean difference analysis is performed by attribute and according to the categorization of Y (step 2) discarding those that are not relevant. Then, a correlation analysis is performed (step 3.1), and among those with a correlation greater than 0.6, the significant variables are selected according to the magnitude of the coefficient by univariate logistic regression (step 3.2).

In this case, of the variables filtered in each sample (steps 2 and 3), 11 appear in at least the $a = 30\%$ samples (step 4). Thus, a design matrix is formed X° of dimension 354×11 where each column represents a possible determinant.

At this stage of the process, the set of possible determinants is made up of sex, age, neurological symptoms, skin symptoms, respiratory symptoms, heart disease, previous respiratory disease, thyroid disease, dyslipidaemia, dementia, and ventilatory failure.

Once the above determinants have been identified, we proceed to construct a synthetic sample that presents a balance, for which we have considered the relationship 65-35. To do this, the synthetic minority oversampling technique (SMOTE) process described in Section B of the methodology is carried out, identifying the W closest neighbours using Euclidean distance (step 5.2) and constructing the synthetic patient (step 5.3). Thus, a final information set is obtained CI^F, balanced in size 455×12.

Subsequently, the estimation of a logistic regression is performed using the *stepwise* methodology (step 6). In this regression, the dependent variable is the

binary response Y between patients who die and those who live, and the explanatory variables are those significant with the inclusion criteria (5%) and exclusion criteria (10%). These were sex, age, thyroid disease, skin symptoms, respiratory symptoms, and ventilatory failure.

Table 1 shows the results of the logistic regression estimated using the *stepwise* methodology. All variables were statistically significant at conventional levels. Parameter signs were the expected ones. Specifically, men are more likely to die from the COVID-19 virus than women. The likelihood of death increases with age. The age variable for the results reported here used a threshold value of 60 years, but the results are robust to different cut-off values. A patient presenting skin or respiratory symptoms at triage or suffering from thyroid disease is more likely to experience health complications and die. Ventilation complications are also good predictors of death. Marginal effects are shown below.

Table 1: Logit regression.

y	Coef.	Robust Std. Err.	z	$p > \|a\|$	95% Conf.	Interval
Gender	-1.1570	0.3924	-2.95	0.003	-1.9262	-0.3879
Age	2.9072	0.3985	7.30	0.000	2.1261	3.6883
Thyroid disease	1.2695	0.5672	2.24	0.025	0.1577	2.3812
Skin symptoms	2.8104	0.4111	6.84	0.000	2.0048	3.6161
Respiratory symptoms	1.8443	0.7655	2.41	0.016	0.3441	3.3446
Ventilatory failure	3.2106	0.3877	8.28	0.000	2.4507	3.9705
cons	-5.7987	0.8324	-6.97	0.000	-7.4302	-4.1672

Prediction Quality - ROC

The threshold value of the predicted probability for identifying a patient classified as dead is 65%. Table 2 shows the quality of prediction. Almost 90% of individuals were correctly classified. The corresponding ROC curve, presented in Figure 2, shows evidence supporting this claim. It is remarkable that the ROC reached a value of 1, meaning that in general, the model had a very good classification ability. Moreover, looking at the specificity and sensitivity results of the previous table, the selected classification cut point was appropriate, as it results in one of the closest dots to the upper left corner of the ROC curve, meaning that it returns one of the best possible classifications. results.

Approximately 90% of the patients were correctly classified. A sensitivity measure of approximately 80% is reached, that is, those who were classified as dead and those who actually died. Likewise, a specificity measure of 94% was obtained for those patients who were classified as survivors and those who survived. On the other hand, false-positives, those that were classified as alive when they actually died, reached 6%. Finally, the false-negatives, those that are classified as dead when they actually lived, reach 20%.

Table 2: Prediction Quality.

	True		
Classified	$Y = 1$	$Y = 0$	Total
$Y = 1$	125	17	142
$y = 0$	35	279	314
Total	160	296	456
Correctly classified			88.60%

* Results from STATA.

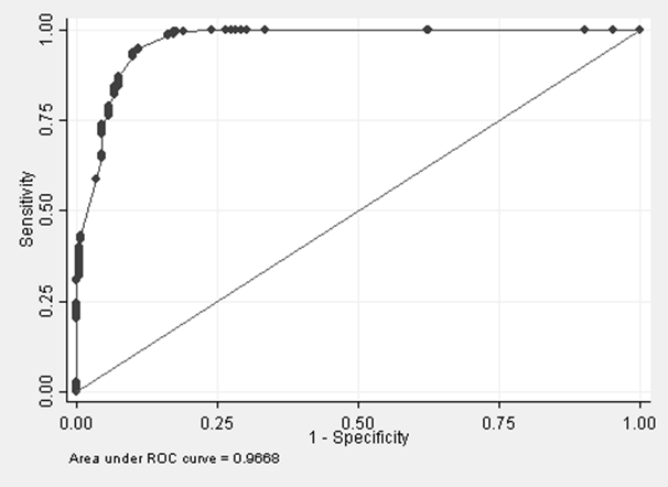

Figure 2: ROC Curve.

Marginal Effects

Since SMOTE generates continuous values in dichotomic variables, the analysis of marginal effects is achieved under average marginal effects by re-estimating the model with the same databases but approximating the continuous numbers to the nearest dichotomous neighbour with a cut of 0.5. The results do not vary in sign and change marginally in magnitude. Table 3 presents the average marginal effects of each variable. The standard error of each marginal effect was estimated under the delta method.

The results indicate that sex, age, skin and respiratory symptoms, thyroid disease, and ventilatory failure are the most relevant predictors of death. Specifically, women are, on average, 7.94% less likely to die than men. Population over 60 are 19.25% more likely to die of COVID-19. Presenting skin or respiratory symptoms at the triage procedure increases the probability of death by 18.90% or 11.23%, respectively. A person suffering from thyroid disease has an extra 7.09% chance of dying. Finally, patients presenting with a ventilatory failure complication have an extra 21.79% chance of death.

Table 3: Marginal Effects.

Variables	dy/dx	Delta-Method Std. Err.	z	$p > \|z\|$	95% Conf.	Interval
Gender	-0.0794	0.0252	-3.15	0.002	-0.1287	-0.0301
Age	0.1925	0.0241	8.00	0.000	0.1453	0.2397
Thyroid disease	0.0709	0.0370	1.92	0.055	-0.0016	0.1434
Skin symptoms	0.1890	0.0248	7.61	0.000	0.1403	0.2376
Respiratory symptoms	0.1123	0.0474	2.37	0.018	0.0194	0.2051
Ventilatory failure	0.2179	0.0148	14.77	0.000	0.1890	0.2468

4. Conclusions

This research introduces a mixed methodological approach that effectively combines resampling techniques (bootstrapping) with oversampling methods (SMOTE) to address the challenges posed by small and imbalanced datasets. This approach proved particularly useful in analyzing a sample of patients infected with SARS-CoV-2, treated at the Clínica Universidad de La Sabana during September 2020. Despite the limited size and inherent imbalance of the dataset, the findings are consistent with those from larger studies, underscoring the robustness and reliability of the proposed methodology in identifying key determinants of mortality.

The success of this approach in such a constrained setting highlights its potential applicability in other regions and contexts, where similar challenges with small and imbalanced samples may arise. Extending the application of this methodology could provide valuable insights across different geographical and demographic populations, helping to validate and refine the determinants identified in this study. This would ensure that the findings are not only contextually relevant but also broadly applicable, reinforcing their generalizability.

Given the prevalence of imbalanced datasets in epidemiological and public health research, continued exploration of techniques like SMOTE and its variants, such as SMOTE Bagging, is recommended. These techniques have proven effective in enhancing the accuracy and generalizability of predictive models, particularly when dealing with minority classes in the data. Future studies should consider not only applying but also fine-tuning these methods, potentially comparing them with other data balancing techniques to determine the most effective approach in different research scenarios.

Additionally, further investigation into the use of advanced modeling and statistical analysis techniques, such as machine learning models, could offer enhanced predictive capabilities, particularly in the context of COVID-19 mortality. Incorporating longitudinal analyses through time series and panel data would also provide a more dynamic understanding of the pandemic's evolution and associated risk factors, enabling more precise model adjustments over time.

The integration of more diverse data sources, including genomic data, biomarkers, and electronic health records, would enrich future analyses, offering deeper insights into individual susceptibility and responses to the virus. Such comprehensive data could lead to more accurate predictions and more informed decision-making in clinical and public health contexts.

In summary, while this study confirms the effectiveness of combining bootstrapping and SMOTE techniques in addressing the challenges of small and imbalanced datasets, there is significant potential for these methods to be refined and applied in broader contexts. Enhancing data infrastructure and improving interoperability between health systems will be critical for enabling more comprehensive and collaborative research. These efforts are crucial for advancing public health strategies and ensuring preparedness for future viral threats.

Acknowledgements

This work was supported by Universidad de La Sabana grant number EICEA-144-2021. The authors gratefully acknowledge the support of Juan Andrés Bernal, Pedro Caballero, Saulo Linares, Nicolás Sarmiento and Andrés Roncancio, students belonging to the Econometrics Study Group, for developing their quantitative skills and contributing to this research with the guidance of the authors.

[Received: February 2024 — Accepted: November 2024]

References

Analytics India Magazine (2023), 'Handling imbalanced data with class weights in logistic regression'. https://analyticsindiamag.com/handling-imbalanced-data-with-class-weights-in-logistic-regression/

Banik, A., Nag, T., Chowdhury, S. R. & Chatterjee, R. (2020), 'Why do covid-19 fatality rates differ across countries? an explorative cross-country study based on select indicators', *Global Business Review* **21**(3), 607–625.

Bhandari, S., Shaktawat, A., Tak, A., Patel, B., Shukla, J., Singhal, S., Gupta, K., Kakkar, S. & Dube, A. (2022), 'Logistic regression analysis to predict mortality risk in covid-19 patients from routine hematologic parameters', *Ibnosina Journal of Medicine and Biomedical Sciences* **12**, 123–129.

Breiman, L. (1996), 'Bagging predictors', *Machine Learning* **24**(2), 123–140.

Brown, M. B. & Benedetti, J. K. (1977), 'On the mean and variance of the tetrachoric correlation coefficient', *Psychometrika* **42**(3), 347–355.

Castro, M. C., Gurzenda, S., Macário, E. M. & França, G. V. A. (2021), 'Characteristics, outcomes and risk factors for mortality of 522,167 patients hospitalised with covid-19 in brazil: a retrospective cohort study', *BMJ Open* **11**(5).

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002), 'Smote: Synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research* **16**, 321–357.

Cifuentes, M., Rodriguez-Villamizar, L., Rojas-Botero, M., Alvarez, C. & Fernández-Niño, J. (2021), 'Socioeconomic inequalities associated with mortality for covid-19 in colombia: A cohort nationwide study', *Journal of Epidemiology and Community Health* **75**, jech–2020.

Cornilly, D., Van Aelst, S. & Verdonck, T. (2023), 'Robust inference and modeling of mean and dispersion for generalized linear models', *Journal of the American Statistical Association* . Disponible en: https://link.springer.com/article/10.1080/01621459.2022.2140054.

Dal Pozzolo, A., Caelen, O. & Bontempi, G. (2015), When is undersampling effective in unbalanced classification tasks?, *in* 'Proceedings of the International Conference on Data Mining'.

De la Hoz-Restrepo, F., Alvis-Zakzuk, N. J., De la Hoz-Gomez, J. F., De la Hoz, A., Gómez Del Corral, L. & Alvis-Guzmán, N. (2020), 'Is colombia an example of successful containment of the 2020 covid-19 pandemic? a critical analysis of the epidemiological data, march to july 2020', *International Journal of Infectious Diseases* **99**, 522–529.

Efron, B. & Tibshirani, R. J. (1994), *An Introduction to the Bootstrap*, Chapman & Hall/CRC.

Elkan, C. (2001), The foundations of cost-sensitive learning, *in* 'Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)', pp. 973–978. https://www.ijcai.org/Proceedings/01/Papers/145.pdf

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B. & Herrera, F. (2018), *Learning from Imbalanced Data Sets*, Springer International Publishing.

Fernández-Niño, J., Guerra-Gómez, J. & Idrovo, A. (2020), 'Multimorbidity patterns among covid-19 deaths: Proposal for the construction of etiological models', *Revista Panamericana de Salud Pública* **44**, 1.

Hanifah, F., Wijayanto, H. & Kurnia, A. (2015), 'Smote bagging algorithm for imbalanced dataset in logistic regression analysis (case: Credit of bank x)', *Applied Mathematical Sciences* **9**(13), 6857–6865.

He, H. & Garcia, E. A. (2009), 'Learning from imbalanced data', *IEEE Transactions on Knowledge and Data Engineering* **21**(9), 1263–1284.

Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. (2013), *Applied Logistic Regression*, 3rd edn, Wiley.

Laajaj, R., De Los Rios, C., Sarmiento-Barbieri, I., Aristizabal, D., Behrentz, E., Bernal, R., Buitrago, G., Cucunubá, Z., de la Hoz, F., Gaviria, A., Hernández, L. J., León, L., Moyano, D., Osorio, E., Varela, A. R., Restrepo, S., Rodriguez, R., Schady, N., Vives, M. & Webb, D. (2021), 'Covid-19 spread, detection, and dynamics in bogota, colombia', *Nature Communications* **12**(1), 4726.

Le Thi, H. A. & Nguyen, M. C. (2023), 'Dca-based weighted bagging: A new ensemble learning approach', *Advances in Data Analysis and Classification* . Disponible en: https://link.springer.com/article/10.1007/s00477-022-02185-6.

Li, J., Huang, D., Zou, B., Yang, H., Hui, W., Rui, F., Yee, N., Liu, C., Nerurkar, S., Kai, J., Teng, M., Li, X., Zeng, H., Borghi, J., Henry, L., Cheung, R. & Nguyen, M. (2020), 'Epidemiology of covid-19: A systematic review and meta-analysis of clinical characteristics, risk factors and outcomes', *Journal of Medical Virology* **93**. Disponible en: https://doi.org/10.1002/jmv.26424.

Lupei, M. I., Li, D., Ingraham, N. E., Baum, K. D., Benson, B., Puskarich, M., Milbrandt, D., Melton, G. B., Scheppmann, D., Usher, M. G. & Tignanelli, C. J. (2022), 'A 12-hospital prospective evaluation of a clinical decision support prognostic algorithm based on logistic regression as a form of machine learning to facilitate decision making for patients with suspected covid-19', *PLOS ONE* **17**(1), e0262193.

Morgenthaler, S. (2023), 'Robust regression against heavy heterogeneous contamination', *Metrika* . Disponible en: https://link.springer.com/article/10.1007/s00184-022-00832-6.

Neptune.ai (2023), 'How to deal with imbalanced classification and regression data'. Disponible en: https://neptune.ai/blog/imbalanced-data.

Roscino, A. & Pollice, A. (2006), A generalization of the polychoric correlation coefficient, *in* S. Zani, A. Cerioli, M. Riani & M. Vichi, eds, 'Data Analysis, Classification and the Forward Search', Springer Berlin Heidelberg, pp. 135–142.

Toya, H. & Skidmore, M. (2021), 'A cross-country analysis of the determinants of covid-19 fatalities', *SSRN Electronic Journal* . Disponible en: https://doi.org/10.2139/ssrn.3832483.

Upshaw, T. L., Brown, C., Smith, R., Perri, M., Ziegler, C. & Pinto, A. D. (2021), 'Social determinants of covid-19 incidence and outcomes: A rapid review', *PLoS ONE* **16**(3), e0248336.

Yalaman, A., Basbug, G., Elgin, C. & Galvani, A. (2021), 'Cross-country evidence on the association between contact tracing and covid-19 case fatality rates', *Scientific Reports* **11**(2145).

Clinical History Review Algorithm

This section describes the algorithm used for extracting relevant patient information from clinical histories which were in pdf format, the variable selection method, and the methodology used for balancing a highly unbalanced sample.

All clinical reports contain the information of a given patient which makes it complex and time consuming to extract information on the variables of interest. Therefore, an algorithm in Python Jupyter Notebook was settled to improve the efficiency of this task. Since medical records commonly contain misspelling, a deterministic algorithm approach was used to extract the text from the clinical histories.

Four different dictionaries were done, one for each category of variables entry data, comorbidities evaluation, complications during hospitalization, and intervention strategies in the ICU. These dictionaries contain the words describing each medical condition and several possible ways in which the condition can be written by the medical doctor. Consequently, all variables are dummies, and they take on the value 1 if the symptom, intervention strategy, comorbidity or complication is found in the clinical record. The four dictionaries are shown below in table 4.

Table A1: Dictionaries

Dictionary #1	"fever", "febrile", "with fever", "presenting fever", "afebrile", "without fever", "fever: no", "fever: yes", "deny fever", "persistent fever", " temperature 37.5 "," temperature 38 "," diarrhoea "," vomit "," headache "," refer headache "," deny headache "," without being headache itself "," no headache "," anosmia "," deny anosmia "," dysgeusia "," cva "," brain oedema "," without oedema "," exanthema "," cutaneous lesions "," risk cutaneous lesions "," skin lesions "," dermal lesion "," dermal "," dyspnoea "," dyspnoea: no "," throat "," odinophagea "," cough "," cough with "," tachypnea "," unquantified fever "," antecedent cva "," chronic obstructive disease "," respiratory distress ", "anosmia persists"
Dictionary #2	"aht", "arterial hypertension", "dm", "diabetes mellitus", "type diabetes mellitus", "noninsulin required diabetes", "aneurysm", "no aneurysm", "without aneurysm", "copd" " pulmonary chronic obstructive "," dild "" fibrosis ", pulmonary fibrosis", "asthma", "history of asthma", "kidney insufficiency "," kidney failure "," obesity "," thyroid "," hypothyroidism "," no hypothyroidism symptoms "," hyperthyroidism "," no symptoms of hypothyroidism "," cholesterol "," triglycerides "," arthritis "," leslupus "," lupus "," sclerosis "," cancer "," hiv "," dementia "," alzheimer "," parkinson "," senile "," smoking "," no smoking "," ex-smoking "
Dictionary #3	"coinfection", "respiratory failure", "ards", "renal failure", "pulmonary thromboembolism", "deep vein thrombosis", "angio tac", pte "," tevp "," bacterial superinfection "," without evidence of bacterial superinfection "," no bacterial superinfection "," suspected bacterial superinfection "," fungal superinfection "," no fungal structures observed "," fungal "," fungi "," blood cultures for fungi "," fungal tracing "," KOH "," Fungal tracing: positive "," stroke "," anosmia "," denies anosmia "," history of stroke "," no evidence of bacterial superinfection "," bacterial superinfection is ruled out "
Dictionary #4	"Ventilatory support", "eti", "eti patient", "ventilatory support is established", "dialysis", "haemodiafiltration", "enoxaparin", "thromboprophylaxis", "methylprednisolone", "prednisolone", "anticoagulation", " antibiotic "," antibiotic: no "," antibiotic is not started "," steroid "," corticoid "," steroid: no "," steroid: no "," dexamethasone "," methylpremisolone "," thromboprophylaxis is ruled out ", "anticoagulation persists"

Algorithm Implementation and Desk Check

The algorithm takes the pdf file of clinical records and transforms it into a docx file. Then, the algorithm places the file into a python string to search all matching coincidences between the medical history and the abovementioned dictionaries. Once the text is extracted and the dictionaries are loaded into the program, the content of the clinical history with the different set of words is compared. The results are presented in frequency tables, and a function verifies the context of the words to identify whether the symptom was active or not for a given patient. The logic underlying the algorithm was desk-checked before it was implemented in the data by using the Hansel and Gretel story and by rambling through every line in a pseudo-code to identify logic bugs.